18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Optimal Fusion Rules for Label Fusion of Dependent Classification Systems

James A. Fitch, Mark E. Oxley, Christine M. Schubert Kabban
Department of Mathematics and Statistics
Air Force Institute of Technology
Graduate School of Engineering and Management
Wright-Patterson AFB, OH 45433-7765
Email: james.fitch@linquest.com, Mark.Oxley@afit.edu, Christine.SchubertKabban@afit.edu

***Abstract*—A classification system with M possible output labels (or decisions) will have M(M-1) possible errors. The Receiver Operating Characteristic (ROC) manifold was created to quantify all of these errors. When multiple classification systems are fused, the assumption of independence is usually made in order to mathematically combine the individual ROC manifolds for each system into one ROC manifold. In this paper we will start with the independence assumption and then investigate fused statistically-dependent classification systems. Specifically, we will use label fusion (also called decision fusion) of multiple classification systems to combine these dependent systems and demonstrate the benefit in performance of incorporating the dependence effects into the fused classification system. We will derive the formula for the generalized AND rule for the resultant ROC manifold of the fused classification system which incorporates the individual dependent classification systems. We will also develop a method utilizing permutation matrices to generate formulas for other label-fusion rules. Examples will be given that demonstrate how the formulas are used.**

I. Introduction

This paper considers the fusion of two multi-class classification systems that are dependent but for which the dependency is unknown. Both systems have the same output set and are designed to classify the same events. We will start with the independence assumption and then investigate fused statistically-dependent classification systems. Specifically, we will use label fusion (also called decision fusion) of multiple classification systems to combine these dependent systems and demonstrate the benefit in performance of incorporating the dependence effects into the fused classification system. We will derive the formula for the generalized AND rule for the resultant ROC manifold of the fused classification systems which incorporates the individual dependent classification systems. We will also develop a method which utilizes permutation matrices to generate formulas for other label-fusion rules. Examples will be given that demonstrate how the formula is used.

Similar work on the topic of label fusion of classification systems has been accomplished but which relied on assumptions of the classification systems being statistically independent and / or the use of only two-label classifiers. See [1], [2], [3] for example. Still other types of work such as that discussed by Kittler [4] involves algorithms which combine several multi-class classifiers at once. In contrast, the work presented here pertains to label fusion of two statistically dependent multi-classification systems.

II. Mathematical Background

A. Classification System Theory

Let $\mathcal{E}$ be a population set of outcomes, and let $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_M\}$ be a label set. We assume there is a truth mapping $\mathbf{T} : \mathcal{E} \to \mathcal{L}$ such that $\mathbf{T}$ partitions the population set with $\mathscr{P} = \{\mathcal{E}_{\ell_1}, \mathcal{E}_{\ell_2}, \ldots, \mathcal{E}_{\ell_M}\}$, where $\mathcal{E}_{\ell_k} = \{e \in \mathcal{E} : \mathbf{T}(e) = \ell_k\}$. That is, $\mathcal{E}_{\ell_1} \cup \mathcal{E}_{\ell_2} \cup \cdots \cup \mathcal{E}_{\ell_M} = \mathcal{E}$ and $\mathcal{E}_{\ell_i} \cap \mathcal{E}_{\ell_j} = \varnothing \quad \forall i \neq j$. Let $\mathfrak{E}$ be a σ-algebra of subsets of $\mathcal{E}$ that contains the partition $\mathscr{P}$, then $(\mathcal{E}, \mathfrak{E})$ is a measurable space. We seek the truth mapping, but the best one can do is to approximate it using a classification system. Let $\Pr$ be a probability measure defined on $\mathfrak{E}$, then $(\mathcal{E}, \mathfrak{E}, \Pr)$ is a probability measure space. Let $\mathbf{s}$ be a sensor that produces data as its output, i.e., $\mathbf{s} : \mathcal{E} \to \mathcal{D}$, where $\mathcal{D}$ is the sensor data set. The outcome from this data set may be too difficult to quantify on its own, so a feature extractor map $\mathbf{p}$, defined on $\mathcal{D}$ produces a refined data object called a feature. The map $\mathbf{p}$ is a processor that takes a datum and produces a feature, i.e., $\mathbf{p} : \mathcal{D} \to \mathcal{F}$. (Typically, $\mathcal{F} = \mathbb{R}^N$ for some positive integer N.) Let Θ be a threshold set or parameter. So, for each $\theta \in \Theta$ let $\mathbf{a}_\theta$ be a classifier mapping $\mathcal{F}$ into a label set $\mathcal{L}$. That is, $\mathbf{a}_\theta : \mathcal{F} \to \mathcal{L}$ for each $\theta \in \Theta$. A graphical representation of the composition of these mappings, $\mathbf{A}_\theta \equiv \mathbf{a}_\theta \circ \mathbf{p} \circ \mathbf{s}$, is given in the following diagram.

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

We design the system $\mathbf{A}_\theta$ to map outcomes in $\mathcal{E}_{\ell_i}$ to ℓ_i , that is, if $e \in \mathcal{E}_{\ell_i}$ we designed for $\mathbf{A}_\theta(e) = \ell_i$. We use the probability measure to quantify the approximation $\mathbf{A}_\theta$ of $\mathbf{T}$ for some choice of $\theta \in \Theta$.

B. Receiver Operating Characteristic (ROC) Manifold

Each mapping in the classification system, as well as the composition of mappings, has a *pre-image* defined as follows.

Definition (Pre-image) Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty sets. Let the mapping $\mathbf{f}$ take an element from $\mathcal{X}$ and map it into $\mathcal{Y}$, that is, $\mathbf{f} : \mathcal{X} \to \mathcal{Y}$ or $\mathcal{X} \xrightarrow{\mathbf{f}} \mathcal{Y}$. Given a subset $Y \subset \mathcal{Y}$ we define the *pre-image* of $\mathbf{f}$ to be the subset in $\mathcal{X}$ by

$$\mathbf{f}^{\natural}(Y) = \{x \in \mathcal{X} : \mathbf{f}(x) \in Y\}. \tag{1}$$

Thus, the pre-image of a subset Y in $\mathcal{Y}$ is all the elements in $\mathcal{X}$ that are mapped by $\mathbf{f}$ into Y.

The use of pre-images allows us to take labels and express them in terms of the underlying events. A classification system with M labels has M possible correct classifications (one for each of the M labels) and $M^2 - M$ possible misclassifications (each label can misclassify an object in $M-1$ ways). The ROC manifold is defined in terms of these misclassifications. Then, since we are dealing with misclassifications, i.e. errors, we wish to minimize these errors. We denote the probability that system $\mathbf{A}_\theta$ classifies an outcome e as label ℓ_i when the outcome is really in $\mathcal{E}_{\ell_j}$ as

$$\begin{aligned} p_{i|j}(\mathbf{A}_\theta) &= \Pr\{\mathbf{A}_\theta(e) = \ell_i \mid e \in \mathcal{E}_{\ell_j}\} \\ &= \frac{\Pr\left(\mathbf{A}_\theta^\natural\left(\{\ell_i\}\right) \cap \mathcal{E}_{\ell_j}\right)}{\Pr\left(\mathcal{E}_{\ell_j}\right)}. \end{aligned} \tag{2}$$

Properties of these quantities imply that for every $\theta \in \Theta$

$$\sum_{i=1}^{M} p_{i|j}\left(\mathbf{A}_\theta\right) = 1 \tag{3}$$

for each $j = 1, 2, \ldots, M$. Only the $p_{j|j}$ term is a correct classification; the other $M-1$ terms denote the misclassifications (i.e., the errors) of system $\mathbf{A}_\theta$ so,

$$\sum_{i=1, i\neq j}^{M} p_{i|j}(\mathbf{A}_\theta) = 1 - p_{j|j}(\mathbf{A}_\theta) \tag{4}$$

for each $j = 1, 2, \ldots, M$ and for every $\theta \in \Theta$. Define the $M \times M$ matrix $\mathtt{P}\left(\mathbf{A}_\theta\right)$ to be the matrix whose i, j entry is the conditional probability $p_{i|j}(\mathbf{A}_\theta)$ for $i = 1, ..., M$ and $j = 1, \ldots, M$. That is,

$$\left[\mathtt{P}\left(\mathbf{A}_\theta\right)\right]_{i,j} = p_{i|j}(\mathbf{A}_\theta) = \frac{\Pr(\mathbf{A}_\theta^\natural(\{\ell_i\}) \cap \mathcal{E}_{\ell_j})}{\Pr(\mathcal{E}_{\ell_j})}. \tag{5}$$

Equation (3) implies $\mathtt{P}\left(\mathbf{A}_\theta\right)$ is a (column) stochastic matrix for every $\theta \in \Theta$. Notice that the off-diagonal entries of $\mathtt{P}\left(\mathbf{A}_\theta\right)$ are the errors associated with misclassification. Next, we define the $M \times M$ matrix of misclassifications (dropping the θ subscript for brevity) as

$$\mathtt{R}(\mathbf{A}) = \begin{bmatrix} 0 & p_{1|2}(\mathbf{A}) & p_{1|3}(\mathbf{A}) & \cdots & p_{1|\mathrm{M}}(\mathbf{A}) \\ p_{2|1}(\mathbf{A}) & 0 & P_{2|3}(\mathbf{A}) & \cdots & p_{2|\mathrm{M}}(\mathbf{A}) \\ p_{3|1}(\mathbf{A}) & p_{3|2}(\mathbf{A}) & 0 & & p_{3|\mathrm{M}}(\mathbf{A}) \\ \vdots & \vdots & & \ddots & \vdots \\ p_{\mathrm{M}|1}(\mathbf{A}) & p_{\mathrm{M}|2}(\mathbf{A}) & p_{\mathrm{M}|3}(\mathbf{A}) & \cdots & 0 \end{bmatrix}. \tag{6}$$

C. 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 maps datums 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 $\phi \in \Phi$. 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 (CSFs) $\mathbb{A} \equiv \{\mathbf{A}_\theta : \theta \in \Theta\}$ and $\mathbb{B} \equiv \{\mathbf{B}_\phi : \phi \in \Phi\}$.

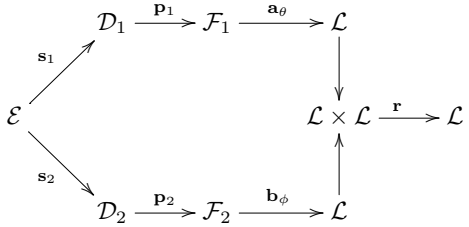
The two classification systems developed above map outcomes from the population set into different data, feature, and label sets, which are then used to fuse the classification systems together. There are, however, other ways to label the outcomes from the event set. In this discussion, classification 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 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.

D. Fusion Rules

Let the systems $\mathbb{A}$ and $\mathbb{B}$ be defined as above. We seek to combine these two families in some manner. The easiest way is to combine their outputs using a function, or rule, defined on $\mathcal{L}$, that is, $\mathbf{r} : \mathcal{L} \times \mathcal{L} \to \mathcal{L}$ with $\mathcal{D}om(\boldsymbol{r}) = \mathcal{L} \times \mathcal{L}$. Then a new CSF is created by, for every outcome $e \in \mathcal{E}$

$$\mathbf{C}_{\theta,\phi}(e) = \mathbf{r}(\mathbf{A}_\theta(e), \mathbf{B}_\phi(e))$$

for every $\theta \in \Theta$ and $\phi \in \Phi$. Hence, $\mathbb{C} = \{\mathbf{C}_{\theta,\phi} : \theta \in \Theta, \phi \in \Phi\}$ and we write $\mathbb{C} = \mathbf{r}\left(\mathbb{A}, \mathbb{B}\right)$ (a slight abuse of notation). The diagram that corresponds to this new CSF is



There are several rules $\mathbf{r}$ that exist for a label set with M labels. In fact, there are $M^{(M^2)}$ rules for 2 systems. If we wish to combine N systems then the number of rules $R = M^{(M^N)}$ [3]. The table below gives examples to show how the number of rules grows for a small number of labels.

Labels	Systems	Rules	Consistent*
M	N	R	C
2	2	16	2
2	3	256	
3	2	19,683	6
3	3	7,625,597,484,987	
4	2	4,294,967,296	24
5	2	298,023,223,876,953,125	120

*We define Consistent rules in section II-F.

E. Generalized AND Rule Assuming Independence

In this section we develop a method of combining classification systems via label fusion assuming that the two systems are statistically independent. In this discussion we focus on the Boolean-like generalized AND rule as it forms the foundation for other rules which follow. This will be expanded further in the next section.

Let the ROC manifold associated with the classification system family $\mathbb{A} \equiv \{\mathbf{A}_\theta : \theta \in \Theta\}$ be denoted by $\mathscr{M}_{\mathbb{A}}$ and the ROC manifold associated with the classification system family $\mathbb{B} \equiv \{\mathbf{B}_\phi : \phi \in \Phi\}$ be denoted by $\mathscr{M}_{\mathbb{B}}$. In the definitions that follow, we continue to use the label set $\mathcal{L} = \{\ell_1, \ell_2, \ldots, , \ell_M\}$. Also in the definitions that follow, we assume that the label set is ordered in terms of the labels' importance. That is ℓ_{i+1} is more important than label ℓ_i.

The AND rule is a binary operation defined on $\mathcal{L} = \{\ell_1, \ell_2, \ldots, , \ell_M\}$ and is defined as follows:

$$r(\ell_i, \ell_j) = \min_{\substack{\mathcal{L} \\ \leq}} \{\ell_i, \ell_j\} = \ell_{\min_{\substack{\mathbb{R} \\ \leq}} \{i,j\}}. \tag{7}$$

We denote this operation by $\wedge$, just as in the traditional Boolean AND operation, but here we call this rule the *generalized AND rule*. The new classification system $\mathbf{C}^{\text{AND}}_{\theta,\phi}$ is defined by the point-wise AND operation on its output, that is,

$$\begin{aligned} \mathbf{C}^{\text{AND}}_{\theta,\phi}(e) &= \mathbf{A}_\theta(e) \wedge \mathbf{B}_\phi(e) \\ &= \min\{\mathbf{A}_\theta(e), \mathbf{B}_\phi(e)\} \quad \text{for all } e \in \mathcal{E}. \end{aligned} \tag{8}$$

This produces a new classification system family $\mathbb{C}^{\text{AND}} = \{\mathbf{C}_{\theta,\phi} : \theta \in \Theta, \phi \in \Phi\}$. The conditional probability that system $\mathbf{C}_{\theta,\phi}$ classifies an object as label ℓ_i when the object is actually ℓ_j is given by an expression similar to (2) as

$$\begin{aligned} p_{i|j}(\mathbf{C}_{\theta,\phi}) &= \frac{\Pr\left(\mathbf{C}^{\natural}_{\theta,\phi}(\{\ell_i\}) \cap \mathcal{E}_{\ell_j}\right)}{\Pr\left(\mathcal{E}_{\ell_j}\right)} \\ &= \frac{\Pr\left((\mathbf{A}_\theta \wedge \mathbf{B}_\phi)^{\natural}(\{\ell_i\}) \cap \mathcal{E}_{\ell_j}\right)}{\Pr\left(\mathcal{E}_{\ell_j}\right)}. \end{aligned} \tag{9}$$

Assuming statistical independence between CFSs $\mathbb{A}$ and $\mathbb{B}$, Oxley [1] shows the conditional probability in terms of the individual probabilities as:

$$\begin{aligned} p_{i|j}(\mathbf{C}^{\text{AND}}_{\theta,\phi}) = p_{i|j}(\mathbf{A}_\theta)p_{i|j}(\mathbf{B}_\phi) &+ \sum_{k=i+1}^{M} p_{k|j}(\mathbf{A}_\theta)p_{i|j}(\mathbf{B}_\phi) \\ &+ \sum_{k=i+1}^{M} p_{i|j}(\mathbf{A}_\theta)p_{k|j}(\mathbf{B}_\phi) \end{aligned} \tag{10}$$

A more convenient means by which to use and analyze (10) is to form an equivalent representation. That is, the matrix representation of (10) is

$$\begin{aligned} \mathtt{P}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right) = \mathtt{P}(\mathbf{A}_\theta) \odot \mathtt{P}(\mathbf{B}_\phi) &+ \mathbf{U}\mathtt{P}(\mathbf{A}_\theta) \odot \mathtt{P}(\mathbf{B}_\phi) \\ &+ \mathbf{U}\mathtt{P}(\mathbf{B}_\phi) \odot \mathtt{P}(\mathbf{A}_\theta) \end{aligned} \tag{11}$$

where $\odot$ denotes the Hadamard matrix product and $\mathbf{U}$ is an $M \times M$ upper triangular matrix where $\mathbf{U}_{i,j} = 1$ for $i \leq j$ and 0 for $i > j$. Note that the matrix multiplications with $\mathbf{U}$ represent the summations in (10). To simplify notation in (11), let $\mathtt{C} = \mathtt{P}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right)$, $\mathtt{A} = \mathtt{P}(\mathbf{A}_\theta)$, $\mathtt{B} = \mathtt{P}(\mathbf{B}_\phi)$, and $\mathbf{V} = \frac{1}{2}\mathbf{I} + \mathbf{U}$, then we can write (11) as

$$\mathtt{C} = \mathbf{V}\mathtt{A} \odot \mathtt{B} + \mathbf{V}\mathtt{B} \odot \mathtt{A}. \tag{12}$$

We wish to transform this matrix equation once more into a matrix-vector form using properties of the Kronecker product and the vectorization operation. We denote the Kronecker product of an $m \times n$ matrix $\mathbf{A}$ and a $p \times q$ matrix $\mathbf{B}$ as $\mathbf{A} \otimes \mathbf{B}$. The operation $\mathtt{vec}(\mathbf{A})$ is the vertical concatenation of the columns of $\mathbf{A}$. That is, $\mathtt{vec}(\mathbf{A}) = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_n \end{bmatrix}^T$, where each $\mathbf{A}_i$ is the ith column of $\mathbf{A}$. The Kronecker product $\otimes$ and $\mathtt{vec}()$ operation are related according to the following theorem.

Theorem 2.1: For any three matrices $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{C}$ for which the matrix product $\mathbf{AXB}$ is defined,

$$\mathtt{vec}(\mathbf{AXC}) = (\mathbf{C}^T \otimes \mathbf{A}) \cdot \mathtt{vec}(\mathbf{X}).$$

In the case when $\mathbf{C}$ is the identity matrix $\mathbf{I}$, then

$$\mathtt{vec}(\mathbf{AX}) = \mathtt{vec}(\mathbf{AXI}) = (\mathbf{I} \otimes \mathbf{A}) \cdot \mathtt{vec}(\mathbf{X}).$$

The Hadamard product and $\mathtt{vec}()$ operation are also related by the following theorem.

Theorem 2.2: For any two matrices $\mathbf{A}$ and $\mathbf{X}$ for which the Hadamard product $\mathbf{A} \odot \mathbf{X}$ is defined,

$$\mathtt{vec}(\mathbf{A} \odot \mathbf{X}) = \mathtt{diag}(\mathtt{vec}(\mathbf{A})) \cdot \mathtt{vec}(\mathbf{X})$$

where $\mathtt{diag}(\mathbf{A})$ is the diagonal matrix formed by the elements of the vector $\mathtt{vec}(\mathbf{A})$.

We will also make use of the following two notation simplifications. First, we will denote the Kronecker product of the matrix $\mathbf{X}$ with the identity matrix, i.e., $\mathbf{I} \otimes \mathbf{X}$ as the block diagonal matrix $\mathbf{X}_D$ using the subscript D. Second, we will denote $\mathtt{vec}(\mathbf{X})$ as $\bar{\mathbf{X}}$ using the over bar.

We can now rewrite (12) as

$$\begin{aligned} \mathtt{vec}(\mathtt{C}) &= \mathtt{vec}(\mathbf{V}\mathtt{A}) \odot \mathtt{vec}(\mathtt{B}) + \mathtt{vec}(\mathbf{V}\mathtt{B}) \odot \mathtt{vec}(\mathtt{A}) \\ \bar{\mathtt{C}} &= \mathbf{V}_D\bar{\mathtt{A}} \odot \bar{\mathtt{B}} + \mathbf{V}_D\bar{\mathtt{B}} \odot \bar{\mathtt{A}}. \end{aligned} \tag{13}$$

F. Consistent Fusion Rules

Fitch [3], [5] shows that the set of consistent rules (defined below) are a meaningful and desirable set of fusion rules to be used in label-based classifier fusion.

Definition (Agreement Rule) Let $f : \mathcal{L}^N \to \mathcal{L}$ be a function. If for every $\mathbf{w} \in \mathcal{L}^N$, $f(\mathbf{w}) \in \mathbf{w}$ then f is an agreement function or agreement rule.

Definition (Symmetric Rule) Let $\phi : \mathcal{L} \to \mathcal{L}$ be a permutation on $\mathcal{L}$. Define the function $\Phi : \mathcal{L}^N \to \mathcal{L}^N$ by using ϕ as follows: for $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N) \in \mathcal{L}^N$, $\Phi(\mathbf{w}) = (\mathbf{w}_{\phi(1)}, \mathbf{w}_{\phi(2)}, \ldots, \mathbf{w}_{\phi(N)})$. Let $f : \mathcal{L}^N \to \mathcal{L}$ be a function. If for every $\mathbf{w} \in \mathcal{L}^N$ and permutation ϕ, $f(\Phi(\mathbf{w})) = f(\mathbf{w})$ then f is a symmetric function or symmetric rule.

Consistent rules are then defined as follows.

Definition (Consistent Rules) The consistent rules are formed by the intersection of the sets of agreement rules and symmetric rules. That is,

$$Rules_{consistent} = Rules_{agreement} \cap Rules_{symmetric}$$

The conditional probability of the fused system after applying a consistent rule is computed via use of a permutation matrix in (12) as follows [3], [5]

$$\mathtt{C}_{(i)} = \mathbf{Q}_{(i)}^T \mathbf{V} \mathbf{Q}_{(i)} \mathtt{A} \odot \mathtt{B} + \mathbf{Q}_{(i)}^T \mathbf{V} \mathbf{Q}_{(i)} \mathtt{B} \odot \mathtt{A} \tag{14}$$

where, $\mathbf{Q}_{(i)}$ is a permutation matrix designed to permute the ordering of the label set $\mathcal{L}$. In vector form (14) is written as

$$\bar{\mathtt{C}}_{(i)} = \mathbf{Q}_{D(i)}^T \mathbf{V}_D \mathbf{Q}_{D(i)} \bar{\mathtt{A}} \odot \bar{\mathtt{B}} + \mathbf{Q}_{D(i)}^T \mathbf{V}_D \mathbf{Q}_{D(i)} \bar{\mathtt{B}} \odot \bar{\mathtt{A}} \tag{15}$$

where, $\mathbf{Q}_{D(i)} = \mathbf{I} \otimes \mathbf{Q}_{(i)}$ and $\mathbf{V}_D = \mathbf{I} \otimes \mathbf{V}$.

Fitch [5] shows that there are $M!$ consistent rules. Thus, for a pair of 3-label classification systems, there are $3 \cdot 2 \cdot 1 = 6$ consistent label fusion rules and thus 6 permutation matrices which represent those 6 rules.

G. Performance Functionals

Clearly, analysis is needed to determine the best rule prior to building a system. We define the notion of "best" by using a performance functional ϱ so that $\varrho(\mathbb{A})$ is a real number, which can be used to quantify "better". Then we use optimization to quantify "best". Assume a smaller value means better, that is, if $\varrho(\mathbb{A}) < \varrho(\mathbb{B})$ then $\mathbb{A}$ is better than $\mathbb{B}$. Our definition of Fusion is the following: if

$$\varrho(\mathbf{r}(\mathbb{A}, \mathbb{B})) < \min\{\varrho(\mathbb{A}), \varrho(\mathbb{B})\}$$

then we call the rule $\mathbf{r}$ a *fusor* for families $\mathbb{A}$ and $\mathbb{B}$ with respect to the performance functional ϱ. If the performance ϱ is fixed then we can seek the best fusor via the optimization problem

$$\min_{\substack{\mathbf{r} \in \mathtt{Rules} \\ \mathbf{A} \in \mathbb{A}, \mathbf{B} \in \mathbb{B}}} \varrho(\mathbf{r}(\mathbb{A}, \mathbb{B})).$$

The performance functional we will use is Bayes cost which allows us to incorporate prior probabilities and costs of misclassifications as in

$$\begin{aligned} \rho(\mathbf{A}_\theta) &= \varphi(\mathtt{R}(\mathbf{A}_\theta)) \\ &= \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} c_{i|j} P_j p_{i|j}(\mathbf{A}_\theta) = \langle \Gamma, \mathtt{R}(\mathbf{A}_\theta) \rangle \end{aligned} \tag{16}$$

where the matrix $\Gamma = \mathtt{C} \cdot \mathrm{diag}(P)$ for the cost matrix $\mathtt{C}$, and diagonal matrix, $\mathrm{diag}(P)$ of prior probabilities.

III. Main Results

A. Generalized AND Rule with Dependence

In this section we develop a method of combining two classification systems similar to that shown in section II-E. However, in this case we do not assume independence between the two systems. We begin by defining a total ordering of the label set $\mathcal{L} = \{\ell_1 \prec \ell_2 \prec \ldots \prec \ell_M\}$. Then we develop an expression for each $p_{i|j}(\mathbf{C}_{\theta,\phi})$ in terms of the individual classification systems $\mathbf{A}_\theta$ and $\mathbf{B}_\phi$.

Choose a label ℓ_i and fix the index i. By the total ordering of $\mathcal{L} = \{\ell_1, \ell_2, ..., \ell_M\}$ then $r_{\mathrm{AND}}(\ell_{i'}, \ell_{j'}) = \ell_{\min\{i',j'\}} = \ell_i$ implies that the possible choices of (i', j') integer pairs are

$$\begin{aligned} &(i,i), (i+1,i), (i+2,i), ..., (M,i), \\ &(i,i+1), (i,i+2), ..., (i,M). \end{aligned} \tag{17}$$

Therefore, $\mathbf{C}_{\theta,\phi}(e) = r_{\text{AND}}(\mathbf{A}_\theta(e), \mathbf{B}_\phi(e)) = \ell_i$ implies the set

$$\begin{aligned} \mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\}) &= \{e \in \mathcal{E} : \mathbf{C}_{\theta,\phi}(e) = \ell_i\} \\ &= \{e \in \mathcal{E} : r_{\text{AND}}(\mathbf{A}_\theta(e), \mathbf{B}_\phi(e)) = \ell_i\} \end{aligned} \tag{18}$$

can be partitioned into the collection of the sets

$$\begin{aligned} &\{e \in \mathcal{E} : [\mathbf{A}_\theta(e) = \ell_i] \cap [\mathbf{B}_\phi(e) = \ell_i]\}, \\ &\{e \in \mathcal{E} : [\mathbf{A}_\theta(e) = \ell_{i+1}] \cap [\mathbf{B}_\phi(e) = \ell_i]\}, \ldots \\ &\{e \in \mathcal{E} : [\mathbf{A}_\theta(e) = \ell_M] \cap [\mathbf{B}_\phi(e) = \ell_i]\}, \\ &\{e \in \mathcal{E} : [\mathbf{A}_\theta(e) = \ell_i] \cap [\mathbf{B}_\phi(e) = \ell_{i+1}]\}, \ldots \\ &\{e \in \mathcal{E} : [\mathbf{A}_\theta(e) = \ell_i] \cap [\mathbf{B}_\phi(e) = \ell_M]\}. \end{aligned} \tag{19}$$

Then the set $\mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\})$ can be written as

$$\begin{aligned} \mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\}) = &\left[\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\})\right] \bigcup \\ &\bigcup_{i'=i+1}^{M} \left[\mathbf{A}_\theta^{\natural}(\{\ell_{i'}\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\})\right] \bigcup \\ &\bigcup_{i''=i+1}^{M} \left[\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_{i''}\})\right]. \end{aligned} \tag{20}$$

Now we intersect $\mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\})$ with event $\mathcal{E}_j$, then take the probability of both sides, and use the fact that the set components are mutually disjoint to yield

$$\begin{aligned} \mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j = &\left[\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right] \bigcup \\ &\bigcup_{i'=i+1}^{M} \left[\mathbf{A}_\theta^{\natural}(\{\ell_{i'}\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right] \bigcup \\ &\bigcup_{i''=i+1}^{M} \left[\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_{i''}\}) \cap \mathcal{E}_j\right], \end{aligned} \tag{21}$$

$$\begin{aligned} &\Pr\left(\mathbf{C}_{\theta,\phi}^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right) \\ &= \Pr\left(\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right) \\ &+ \sum_{k=i+1}^{M} \Pr\left(\mathbf{A}_\theta^{\natural}(\{\ell_k\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_i\}) \cap \mathcal{E}_j\right) \\ &+ \sum_{k=i+1}^{M} \Pr\left(\mathbf{A}_\theta^{\natural}(\{\ell_i\}) \cap \mathbf{B}_\phi^{\natural}(\{\ell_k\}) \cap \mathcal{E}_j\right). \end{aligned} \tag{22}$$

We then divide by the probability of event $\mathcal{E}_j$ to form conditional probabilities so that (22) becomes

$$\begin{aligned}\frac{\Pr\left(\mathbf{C}^{\natural}_{\theta,\phi}\left(\{\ell_i\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)} &= \frac{\Pr\left(\mathbf{A}^{\natural}_{\theta}\left(\{\ell_i\}\right)\cap\mathbf{B}^{\natural}_{\phi}\left(\{\ell_i\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)} \\ &+ \sum_{k=i+1}^{M}\frac{\Pr\left(\mathbf{A}^{\natural}_{\theta}\left(\{\ell_k\}\right)\cap\mathbf{B}^{\natural}_{\phi}\left(\{\ell_i\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)} \\ &+ \sum_{k=i+1}^{M}\frac{\Pr\left(\mathbf{A}^{\natural}_{\theta}\left(\{\ell_i\}\right)\cap\mathbf{B}^{\natural}_{\phi}\left(\{\ell_k\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)}.\end{aligned} \tag{23}$$

To help simplify the notation, we define the left hand side conditional probability as

$$p_{i|j}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right) \equiv \frac{\Pr\left(\mathbf{C}^{\natural}_{\theta,\phi}\left(\{\ell_i\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)} \tag{24}$$

and define each conditional probability term on the right hand side as

$$p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) \equiv \frac{\Pr\left(\mathbf{A}^{\natural}_{\theta}\left(\{\ell_i\}\right)\cap\mathbf{B}^{\natural}_{\phi}\left(\{\ell_k\}\right)\cap\mathcal{E}_j\right)}{\Pr(\mathcal{E}_j)}. \tag{25}$$

That is, $p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi)$ is the conditional probability that system $\mathbf{A}_\theta$ classifies the event e as label ℓ_i while system $\mathbf{B}_\phi$ classifies the event as label ℓ_k, given that the event was $\mathcal{E}_j$. $p_{i|j}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right)$ is the conditional probability that the "ANDed" system classifies the event e as label ℓ_i, given that the event was $\mathcal{E}_j$. So now (23) can be written as

$$\begin{aligned}p_{i|j}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right) = p_{i,i|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) &+ \sum_{k=i+1}^{M} p_{k,i|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) \\ &+ \sum_{k=i+1}^{M} p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi).\end{aligned} \tag{26}$$

Next we define a conditional dependency ratio, μ, corresponding to each conditional probability as

$$\mu_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) \equiv \frac{p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi)}{p_{i|j}(\mathbf{A}_\theta)p_{k|j}(\mathbf{B}_\phi)}. \tag{27}$$

Thus, the conditional probability terms in the summation can be written as

$$p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) = \mu_{i,k|j}\, p_{i|j}(\mathbf{A}_\theta)p_{k|j}(\mathbf{B}_\phi). \tag{28}$$

The conditional probability for each i,j term in the resulting conditional probability matrix for $\mathbf{C}_{\theta,\phi}$ becomes the sum of the probabilities, so therefore

$$\begin{aligned}p_{i|j}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right) = \mu_{i,i|j}p_{i|j}(\mathbf{A}_\theta)p_{i|j}(\mathbf{B}_\phi) &+ \left(\sum_{k=i+1}^{M}\mu_{k,i|j}p_{k|j}\left(\mathbf{A}_\theta\right)\right)p_{i|j}\left(\mathbf{B}_\phi\right) \\ &+ \left(\sum_{k=i+1}^{M}\mu_{i,k|j}p_{k|j}\left(\mathbf{B}_\phi\right)\right)p_{i|j}\left(\mathbf{A}_\theta\right).\end{aligned} \tag{29}$$

Note that for notational brevity θ and ϕ are dropped from the μ terms.

The i,j th term of the conditional probability matrix for system $\mathbf{A}_\theta$ can be written as the sum of $p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi)$ terms as:

$$\begin{aligned}p_{i,j}\left(\mathbf{A}_\theta\right) &= \sum_{k=1}^{M} p_{i,k|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) \\ &= \sum_{k=1}^{M}\mu_{i,k|j}\, p_{i|j}(\mathbf{A}_\theta)p_{k|j}(\mathbf{B}_\phi) \\ &= p_{i|j}(\mathbf{A}_\theta)\sum_{k=1}^{M}\mu_{i,k|j}\, p_{k|j}(\mathbf{B}_\phi).\end{aligned} \tag{30}$$

Thus,

$$\sum_{k=1}^{M}\mu_{i,k|j}\, p_{k|j}(\mathbf{B}_\phi) = 1 \text{ for each } i,j \in 1\ldots M \tag{31}$$

Similarly, the i,j th term of the conditional probability matrix for system $\mathbf{B}_\phi$ can be written as the sum of $p_{k,i|j}(\mathbf{A}_\theta,\mathbf{B}_\phi)$ terms as:

$$\begin{aligned}p_{i,j}\left(\mathbf{B}_\phi\right) &= \sum_{k=1}^{M} p_{k,i|j}(\mathbf{A}_\theta,\mathbf{B}_\phi) \\ &= \sum_{k=1}^{M}\mu_{k,i|j}\, p_{k|j}(\mathbf{A}_\theta)p_{i|j}(\mathbf{B}_\phi) \\ &= p_{i|j}(\mathbf{B}_\phi)\sum_{k=1}^{M}\mu_{k,i|j}\, p_{k|j}(\mathbf{A}_\theta).\end{aligned} \tag{32}$$

Thus,

$$\sum_{k=1}^{M}\mu_{k,i|j}\, p_{k|j}(\mathbf{A}_\theta) = 1 \text{ for each } i,j \in 1\ldots M. \tag{33}$$

These two equations, $\sum_{k=1}^{M}\mu_{k,i|j}\, p_{k|j}(\mathbf{A}_\theta) = 1$ and $\sum_{k=1}^{M}\mu_{i,k|j}\, p_{k|j}(\mathbf{B}_\phi) = 1$, for each j, can be written as a linear system of equations in block diagonal form as:

$$K\bar{\mu} = \mathbf{1}. \tag{34}$$

Here is an $M=2$ example showing the j th block. As before using a simplified notation we let $\mathtt{C} = \mathtt{P}\left(\mathbf{C}^{\text{AND}}_{\theta,\phi}\right)$, $\mathtt{A} = \mathtt{P}\left(\mathbf{A}_\theta\right)$, and $\mathtt{B} = \mathtt{P}\left(\mathbf{B}_\phi\right)$. Also, here the terms $\mathtt{A}_{ij}$ and $\mathtt{B}_{ij}$ represent $p_{i|j}(\mathbf{A}_\theta)$ and $p_{i|j}(\mathbf{B}_\phi)$, respectively.

$$\begin{bmatrix} \mathtt{A}_{1j} & \mathtt{A}_{2j} & 0 & 0 \\ 0 & 0 & \mathtt{A}_{1j} & \mathtt{A}_{2j} \\ \mathtt{B}_{1j} & 0 & \mathtt{B}_{2j} & 0 \\ 0 & \mathtt{B}_{1j} & 0 & \mathtt{B}_{2j} \end{bmatrix}\begin{bmatrix} \mu_{11j} \\ \mu_{21j} \\ \mu_{12j} \\ \mu_{22j} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Each $[K]_j$ is size $2M \times M^2$ and thus K is size $2M^2 \times M^3$; $\bar{\mu}$ is size $M^3 \times 1$, and $\mathbf{1}$ is size $2M^2 \times 1$.

Now, using techniques similar to those shown in section II-E we will form the matrix-vector representation of (29). First, the $\mu_{i,k,j}$ arrays are written as a block diagonal matrix as

$$\mathcal{M} \equiv \begin{bmatrix} \boldsymbol{\mu}^{(1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mu}^{(2)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\mu}^{(M)} \end{bmatrix}$$

where

$$\boldsymbol{\mu}^{(j)} \equiv \begin{bmatrix} \mu_{11j} & \mu_{12j} & \cdots & \mu_{1Mj} \\ \mu_{21j} & \mu_{22j} & \cdots & \mu_{2Mj} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{M1j} & \mu_{M2j} & \cdots & \mu_{MMj} \end{bmatrix}.$$

Thus, the $M \times M \times M$ array of μ_{ikj} terms is arranged as an $(M \times M) \times M$ block diagonal matrix. We now can write the matrix-vector representation of (29) as:

$$\bar{\mathtt{C}} = \left(\mathcal{M}^T \odot \mathbf{V}_D\right) \bar{\mathtt{A}} \odot \bar{\mathtt{B}} + \left(\mathcal{M} \odot \mathbf{V}_D\right) \bar{\mathtt{B}} \odot \bar{\mathtt{A}}. \tag{35}$$

Then, similar to the independence case as with (14), Fitch [5] shows that we can write the consistent rule form of (35) as:

$$\begin{aligned} \bar{\mathtt{C}}_{(i)} = &\left(\mathcal{M}^T \odot \mathbf{Q}_{D(i)}^T \mathbf{V}_D \mathbf{Q}_{D(i)}\right) \bar{\mathtt{A}} \odot \bar{\mathtt{B}} \\ &+ \left(\mathcal{M} \odot \mathbf{Q}_{D(i)}^T \mathbf{V}_D \mathbf{Q}_{D(i)}\right) \bar{\mathtt{B}} \odot \bar{\mathtt{A}}. \end{aligned} \tag{36}$$

As implied by (36), Fitch also shows that the dependency matrix $\mathcal{M}$ is invariant with respect to the rule employed. Thus, if the dependency matrix $\mathcal{M}$ is known for one rule, those same dependency terms can be used to predict the ROC matrix (or equivalent vector) of the label-fused system using each rule depicted by permutation matrix $\mathbf{Q}_{(i)}$, for each $i = 1 \ldots M!$.

B. Compute the Dependency Terms

The dependency terms in (35), i.e., the terms in matrix $\mathcal{M}$, generally are unknown. However, here we will consider the case in which the conditional probability matrix for the "ANDed" system (represented in (35) by the vector $\bar{\mathbf{C}}$) is known via measurements performed on the actual fused system. Thus, in (35) with $\bar{\mathbf{C}}, \bar{\mathbf{A}}, \bar{\mathbf{B}}$, and $\mathbf{V}_D$ all known quantities, the expression is a linear equation in terms of unknown $\mathcal{M}$.

We now rewrite (35) as the linear transformation

$$L(\mathcal{M}) \equiv \left(\mathcal{M}^T \odot \mathbf{V}_D\right) \bar{\mathbf{A}} \odot \bar{\mathbf{B}} + \left(\mathcal{M} \odot \mathbf{V}_D\right) \bar{\mathbf{B}} \odot \bar{\mathbf{A}}. \tag{37}$$

Then, with the known $\bar{\mathbf{C}}$ outcome we can write (35) as

$$L(\mathcal{M}) = \bar{\mathbf{C}}. \tag{38}$$

In order to solve this linear system we must also include the constraints from (34). Thus (38) is augmented as follows. Let

$$\mathbb{L} = \begin{bmatrix} [L] \\ \cdots \\ K \end{bmatrix}$$

We form the matrix representation of L using the standard basis to get $[L]$ who's size is $M^2 \times M^3$, then, the augmented matrix equation becomes

$$\begin{bmatrix} [L] \\ \cdots \\ K \end{bmatrix} \bar{\mu} = \begin{bmatrix} \bar{\mathbf{C}} \\ \cdots \\ \mathbf{1} \end{bmatrix} = \mathbf{b}$$

or simply

$$\mathbb{L}\bar{\mu} = \mathbf{b} \tag{39}$$

where $\bar{\mu}$ is the concatenated vectorization of each $\boldsymbol{\mu}_{(j)}$, $j = 1 \ldots M$. We then pose the optimization problem as

$$\min\left\{\|\bar{\mu} - \mathbf{1}\|_2 : \mu_{i,k|j} \in \mathbb{R} \text{ subject to } \mathbb{L}\bar{\mu} = \mathbf{b}\right\}. \tag{40}$$

We are motivated by this form, i.e., minimizing $\|\bar{\mu} - 1\|$ because we want a $\bar{\mu}$ solution which is closest in some respect to the independent result which consists of all $\bar{\mu}$ terms being 1. We choose the 2-norm here since it provides a convenient method of computing a solution. We then compute the optimal solution for $\bar{\mu}$ as

$$\hat{\bar{\mu}} = \mathbb{L}^T \left(\mathbb{L}\mathbb{L}^T\right)^\dagger \mathbf{b}. \tag{41}$$

As shown in [5] a unique inverse for $\left(\mathbb{L}\mathbb{L}^T\right)$ does not exist, therefore we use a pseudoinverse in the numerical computations.

IV. Examples

In this section we pose a 3-label classification problem similar to one depicted in [6]. Let the label set $\mathcal{L} = \{\ell_1, \ell_2, \ell_3\}$ where the indices 1, 2, and 3 imply the natural ordering of the set. We will assume that the sensor $\mathbf{s}$ and processor $\mathbf{p}$ mapped disjoint events $\mathcal{E}_1, \mathcal{E}_2$, and $\mathcal{E}_3$ into the feature set $\mathcal{F} = \mathbb{R}^2$ and produce sets which are not disjoint but distributed via the following distributions.

$$f_1(x,y) = \frac{e^{-\left(\frac{x-2}{0.5}\right)} e^{-\left(\frac{y-3}{1}\right)}}{(0.5)(1)\left[1 + e^{-\left(\frac{x-2}{0.5}\right)}\right]^2 \left[1 + e^{-\left(\frac{y-3}{1}\right)}\right]^2}$$

$$\begin{aligned} f_2(x,y) = &\frac{0.9}{2\pi(1)(3)} e^{\left(-\left(\frac{x+1}{1}\right)^2 - \left(\frac{y-0}{3}\right)^2\right)} \\ &+ \frac{0.1}{2\pi(0.7)(0.6)} e^{\left(-\left(\frac{x-1.5}{0.7}\right)^2 - \left(\frac{y+0.5}{0.6}\right)^2\right)} \end{aligned}$$

$$f_3(x,y) = \frac{1}{\pi^2(1)(1)} \left[1 + \left(\frac{x-2}{1}\right)^2\right] \left[1 + \left(\frac{y+2}{1}\right)^2\right]$$

Note that f_1 is Logistic, f_2 is bi-modal Gaussian, and f_3 is Cauchy. The density distributions are graphed in Fig. 1. Contour plots of the density functions are shown in Fig. 2. In this example there are two CSFs.

Define classification system $\mathbb{A}$ as $\mathbf{A}_\Theta = \mathbf{a}_{\boldsymbol{\theta}} \circ \mathbf{p} \circ \mathbf{s}$ where

$$\mathbf{a}_{\theta_1,\theta_2,\theta_3}(x,y) = \begin{cases} \ell_1 & \text{if} \quad x \geq \theta_1, \quad y \geq \theta_2 + (x - \theta_1)\theta_3 \\ \ell_2 & \text{if} \quad x < \theta_1, \quad y \in \mathbb{R} \\ \ell_3 & \text{if} \quad x \geq \theta_1, \quad y < \theta_2 + (x - \theta_1)\theta_3 \end{cases}.$$

This classifier creates a vertical plane that shifts horizontally with θ_1. It contains a ray that begins at (θ_1, θ_2) with slope θ_3.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \Theta \equiv [-10, 10] \times [-10, 10] \times \mathbb{R}$ then define the pre-image

$$\mathbf{a}_{\boldsymbol{\theta}}^{\natural}(\ell_i) = \mathbf{a}_{\theta_1,\theta_2,\theta_3}^{\natural}(\ell_i) = \{(x,y) \in \mathbb{R}^2 : \mathbf{a}_{\theta_1,\theta_2,\theta_3}(x,y) = \ell_i\}$$

and generally, the conditional probability as

$$p_{i|j}(\mathbf{A}_{\boldsymbol{\theta}}) = \int_{\mathbf{a}_{\boldsymbol{\theta}}^{\natural}(\ell_i)} f_j(x,y) dy dx.$$

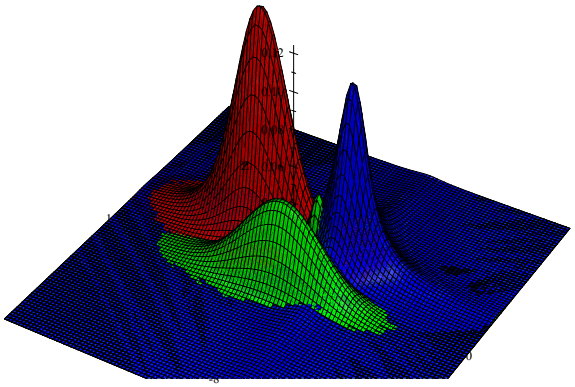


Fig. 1. The plots of two-dimensional density functions: f_1 is Logistic, f_2 is Gaussian (bi-modal), and f_3 is Cauchy.

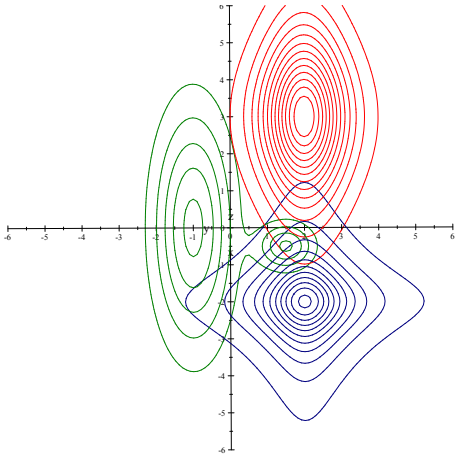


Fig. 2. The contour plots of two-dimensional density functions: f_1 is Logistic, f_2 is Gaussian (bi-modal), and f_3 is Cauchy.

This generates the following six expressions for the misclassification probabilities.

$$\begin{aligned} p_{1|j}(\mathbf{A}_{\boldsymbol{\theta}}) &= \int_{\theta_1}^{\infty}\int_{\theta_2+(x-\theta_1)\theta_3}^{\infty} f_j(x,y)dydx, \quad \text{for } j=2,3 \\ p_{2|j}(\mathbf{A}_{\boldsymbol{\theta}}) &= \int_{-\infty}^{\theta_1}\int_{-\infty}^{\infty} f_j(x,y)dydx, \quad \text{for } j=1,3 \\ p_{3|j}(\mathbf{A}_{\boldsymbol{\theta}}) &= \int_{\theta_1}^{\infty}\int_{-\infty}^{\theta_2+(x-\theta_1)\theta_3} f_j(x,y)dydx, \quad \text{for } j=1,2 \end{aligned}$$

Next we introduce a second classifier. This second classifier by design is not as effective as the first. However, the goal is to demonstrate that a fused system of two classifiers can perform better than the single original classifier. Define classification system $\mathbb{B}$ as $\mathbf{B}_{\Phi} = \mathbf{b}_{\boldsymbol{\phi}} \circ \mathbf{p} \circ \mathbf{s}$ where

$$\mathbf{b}_{\phi_1,\phi_2}(x,y) = \begin{cases} \ell_1 & \text{if} \quad x \in \mathbb{R}, \quad y \geq \phi_2 \\ \ell_2 & \text{if} \quad x \in \mathbb{R}, \quad \phi_1 \leq y < \phi_2 \\ \ell_3 & \text{if} \quad x \in \mathbb{R}, \quad y < \phi_1 \end{cases}.$$

This classifier creates two horizontal lines which shift vertically with ϕ_1 and ϕ_2.

Let $\boldsymbol{\phi} = (\phi_1, \phi_2) \in \Phi \equiv [-11, 11] \times [-11, 11]$ then define the pre-image

$$\mathbf{b}^{\natural}_{\boldsymbol{\phi}}(\ell_i) = \mathbf{b}^{\natural}_{\phi_1,\phi_2}(\ell_i) = \{(x,y) \in \mathbb{R}^2 : \mathbf{b}_{\phi_1,\phi_2}(x,y) = \ell_i\}$$

and generally, the conditional probability as

$$p_{i|j}(\mathbf{B}_{\boldsymbol{\phi}}) = \int_{\mathbf{b}^{\natural}_{\boldsymbol{\phi}}(\ell_i)} f_j(x,y)dydx.$$

This generates the following six expressions for the misclassification probabilities.

$$\begin{aligned} p_{1|j}(\mathbf{B}_{\boldsymbol{\phi}}) &= \int_{-\infty}^{\infty}\int_{\phi_2}^{\infty} f_j(x,y)dydx, \quad \text{for } j=2,3 \\ p_{2|j}(\mathbf{B}_{\boldsymbol{\phi}}) &= \int_{-\infty}^{\infty}\int_{\phi_1}^{\phi_2} f_j(x,y)dydx, \quad \text{for } j=1,3 \\ p_{3|j}(\mathbf{B}_{\boldsymbol{\phi}}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\phi_1} f_j(x,y)dydx, \quad \text{for } j=1,2 \end{aligned}$$

We will use the Bayes performance functional as presented in section II-G given by (16):

$$\begin{aligned} \Gamma &= \begin{bmatrix} 0 & c_{1|2} & c_{1|3} \\ c_{2|1} & 0 & c_{2|3} \\ c_{3|1} & c_{3|2} & 0 \end{bmatrix} \begin{bmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 3 \\ 2 & 0 & 2 \\ 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} \\ 1 & 0 & \frac{1}{3} \\ \frac{1}{2} & 1 & 0 \end{bmatrix}. \end{aligned} \tag{42}$$

A. Performance Results

To find Bayes cost for system $\mathbb{A}$ we minimize $\langle \Gamma, \mathtt{R}(\mathbf{A}_{\boldsymbol{\theta}}) \rangle$ over all $(\theta_1, \theta_2, \theta_3) \in \Theta$, that is, minimize

$$\begin{aligned} &\langle \Gamma, \mathtt{R}(\mathbf{A}_{\boldsymbol{\theta}}) \rangle \\ &= \text{trace} \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} \\ 1 & 0 & \frac{1}{3} \\ \frac{1}{2} & 1 & 0 \end{bmatrix}^T \begin{bmatrix} 0 & p_{1|2}(\mathbf{A}_{\boldsymbol{\theta}}) & p_{1|3}(\mathbf{A}_{\boldsymbol{\theta}}) \\ p_{2|1}(\mathbf{A}_{\boldsymbol{\theta}}) & 0 & p_{2|3}(\mathbf{A}_{\boldsymbol{\theta}}) \\ p_{3|1}(\mathbf{A}_{\boldsymbol{\theta}}) & p_{3|2}(\mathbf{A}_{\boldsymbol{\theta}}) & 0 \end{bmatrix} \\ &= \frac{1}{3}p_{1|2}(\mathbf{A}_{\boldsymbol{\theta}}) + \frac{1}{2}p_{1|3}(\mathbf{A}_{\boldsymbol{\theta}}) + p_{2|1}(\mathbf{A}_{\boldsymbol{\theta}}) \\ &\quad + \frac{1}{3}p_{2|3}(\mathbf{A}_{\boldsymbol{\theta}}) + \frac{1}{2}p_{3|1}(\mathbf{A}_{\boldsymbol{\theta}}) + p_{3|2}(\mathbf{A}_{\boldsymbol{\theta}}). \end{aligned}$$

The resulting Bayes cost is

$$\varrho(\mathbb{A}) = \min_{\boldsymbol{\theta}\in\Theta} \langle \Gamma, \mathtt{R}(\mathbf{A}_{\theta}) \rangle = 0.266$$

occurring at $(\theta_1^*, \theta_2^*, \theta_3^*) = (0.357, -1.678, 0.567)$. Similarly for system $\mathbb{B}$

$$\varrho(\mathbb{B}) = \min_{\boldsymbol{\phi}\in\Phi} \langle \Gamma, \mathtt{R}(\mathbf{B}_{\boldsymbol{\phi}}) \rangle = 0.374$$

occurring at $(\phi_1^*, \phi_2^*) = (-1.33, -0.07)$.

Table I shows Bayes cost for the fused system for each rule, depicted as $\mathbf{Q}_{(i)}$, using the independence assumption and using

TABLE I
BAYES COST FOR INDEP AND DEPEND CASES ACROSS ALL RULES

$\varrho(\mathbb{C})$	$\mathbf{Q}_{(1)}$	$\mathbf{Q}_{(2)}$	$\mathbf{Q}_{(3)}$	$\mathbf{Q}_{(4)}$	$\mathbf{Q}_{(5)}$	$\mathbf{Q}_{(6)}$
indep	0.621	0.487	0.416	0.437	0.365	0.232
depend	0.554	0.375	0.378	0.418	0.420	0.241

the dependency terms. Thus, for the fused system (assuming independence), the resulting Bayes cost for classifier $\mathbb{C}$ is

$$\varrho(\mathbb{C}) = \min_{\substack{\mathbf{r}\in\mathtt{Rules} \\ \boldsymbol{\theta}\in\Theta,\boldsymbol{\phi}\in\Phi}} \varrho(\mathbf{r}\left(\mathbf{A}_{\boldsymbol{\theta}}, \mathbf{B}_{\boldsymbol{\phi}}\right)) = 0.232$$

occurring at $(\theta_1^*, \theta_2^*, \theta_3^*) = (0.285, -1.966, 0.559)$, and $(\phi_1^*, \phi_2^*) = (-9.863, -1.256)$ using Rule $\mathbf{Q}_{(6)}$. For the fused system (including the dependency terms), the resulting Bayes cost for classifier $\mathbb{C}$ is

$$\varrho(\mathbb{C}) = \min_{\substack{\mathbf{r}\in\mathtt{Rules} \\ \boldsymbol{\theta}\in\Theta,\boldsymbol{\phi}\in\Phi}} \varrho(\mathbf{r}\left(\mathbf{A}_{\boldsymbol{\theta}}, \mathbf{B}_{\boldsymbol{\phi}}\right)) = 0.241$$

occurring at $(\theta_1^*, \theta_2^*, \theta_3^*) = (0.272, -2.206, 0.773)$, and $(\phi_1^*, \phi_2^*) = (-10.762, -0.189)$ using Rule $\mathbf{r}^*$ which corresponds to permutation matrix $\mathbf{Q}_{(6)}$.

This result shows that assuming independence in this case would result in the predicted performance (in terms of Bayes cost) of the fused system to be over-reported by almost 10 percent. That is, the predicted independent Bayes cost result of 0.232 is almost 10 percent lower than the realistic value of 0.241 computed using the dependency terms. Additionally, these label-fused results show an improvement of more than 9 percent (using the dependency terms) over that of using system $\mathbb{A}$ alone.

B. Dependency Results

The solution to (40) at the optimal $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}^*$ for the dependency case is

$$\hat{\boldsymbol{\mu}}^{(1)} = \begin{bmatrix} 1.0233 & 0.4066 & -0.0477 \\ 0.5773 & 11.7620 & 0.9662 \\ 0.6479 & 9.9639 & 39.6790 \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}}^{(2)} = \begin{bmatrix} 1.0476 & 0.9515 & 0.9874 \\ 0.9915 & 1.0087 & 0.9297 \\ 0.9994 & 1.0002 & 2.3980 \end{bmatrix}$$

$$\hat{\boldsymbol{\mu}}^{(3)} = \begin{bmatrix} 4.4547 & 0.4224 & 0.6317 \\ 0.8321 & 1.0629 & 0.7668 \\ 0.1703 & 1.1314 & 1.1412 \end{bmatrix}.$$

Notice that many of the dependency terms are close to 1, and several are much larger than 1, which is why the performances in Table I are different for the independent verses dependent cases.

V. CONCLUSION

We have proposed and derived expressions for computing the dependency relation between two classification systems with M labels using the ROC manifold of the two individual systems and a known measurement of a combined system using the AND rule. With the dependency terms computed we are able to iterate over all consistent label fusion rules and using a cost functional to evaluate the performance of the fused system, determine the optimal parameter settings for the fused system. This provides not only the expected performance of the fused system but also the best rule and parameter settings by which to fuse the two systems. We have demonstrated that using these techniques allows for significant improvement in the estimation of the fused ROC manifold over that of assuming independence of the two classification systems.

REFERENCES

[1] M. E. Oxley, S. N. Thorsen, and C. M. Schubert, "The ROC Manifold of Fused Independent Classification Systems," in *Information Fusion, 2009. FUSION '09. 12th International Conference on*, July 2009, pp. 466 –473.

[2] M. E. Oxley, C. M. Schubert, and S. N. Thorsen, "ROC manifolds of multiple fused independent ATR systems," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 7336, May 2009.

[3] J. A. Fitch, M. E. Oxley, and C. M. Schubert Kabban, "Label Fusion of Classification Systems via Their ROC Functions," pp. 839 214–839 214–12.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[5] J. A. Fitch, "Multi-label Classifier Performance of Dependent Classification Systems," Ph.D. dissertation, AFIT/DS/ENC/15-01 Air Force Institute of Technology, Wright-Patterson AFB, OH, 2015.

[6] C. M. Schubert, S. N. Thorsen, and M. E. Oxley, "The ROC Manifold for Classification Systems," *Pattern Recognition*, vol. 44, no. 2, pp. 350 – 362, 2011.

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.