Available online at www.sciencedirect.com

ScienceDirect

Information Fusion 7 (2006) 346–360

INFORMATION FUSION

www.elsevier.com/locate/inffus

A description of competing fusion systems ☆

Steven N. Thorsen, Mark E. Oxley *

Department of Mathematics and Statistics, Graduate School of Engineering and Management, Air Force Institute of Technology,
2950 Hobson Way, Wright-Patterson Air Force Base, OH 45433-7765, USA

Received 11 January 2005; received in revised form 14 October 2005; accepted 17 October 2005
Available online 6 December 2005

Abstract

A mathematical description of fusion is presented using category theory. A category of fusion rules is developed. The category definition is derived for a model of a classification system beginning with an event set and leading to the final labeling of the event. Functionals on receiver operating characteristic (ROC) curves are developed to form a partial ordering of families of classification systems. The arguments of these functionals point to specific ROCs and, under various choices of input data, correspond to the Bayes optimal threshold (BOT) and the Neyman–Pearson threshold of the families of classification systems. The functionals are extended for use over ROC curves and ROC manifolds where the number of classes of interest in the fusion system exceeds two and the parameters used are multi-dimensional. Choosing a particular functional, therefore, provides the qualitative requirements to define a fusor and choose the best competing classification system.

Keywords: Fusion; Optimization; Fusor; Category theory; Information fusion; Bayes optimal; Calculus of variations; Functional

1. Introduction

Information fusion is a rapidly advancing science. Researchers are daily adding to the known repertoire of fusion techniques (that is, fusion rules). An organization that is building a fusion system to detect or identify objects will want to get the best possible result for the money expended. It is this goal which motivates the need to construct a way to compete various fusion rules for acquisition purposes. There are many different methods and strategies involved with developing classification systems. Some rely on likelihood ratios, some on randomized techniques, and still others on a myriad of schemes. To add to this, there exists the fusion of all these technologies which create even more classification systems. Since the receiver operating characteristics (ROCs) can be developed for such systems under test conditions, we propose a functional defined on ROC curves as a method of quantifying the performance of a classification system. A ROC curve is a set of ROCs which define a continuous, non-decreasing (or depending upon the axes chosen, non-increasing) function in ROC space (the concept of ROC curves is developed in Section 3).

This functional then allows for the development of a cogent definition of what is fusion (i.e., the difference between fusion rules, which do not have a reliance upon any qualitative difference between the 'new' fused result and the 'old' non-fused result) and what we term fusors (a subcategory of fusion rules, see Section 2), which do rely upon the qualitative differences. In other words, how does one know the new fused result is "better" than what previously existed? (see Wald [2]). While the development of some classification systems require knowledge of class conditional probability density functions, others do not. A testing organization would not reveal the exact test

☆ A short version of this study was presented at the Seventh International Conference on Information Fusion (Fusion 2004) in Stockholm, Sweden [1]. 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.

* Corresponding author. Tel.: +1 937 2553636x4515; fax: +1 937 656 4413.

E-mail addresses: steven.thorsen@afit.edu (S.N. Thorsen), mark.oxley@afit.edu (M.E. Oxley).

1566-2535/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2005.10.003

scenario to those proposing different classification systems a priori. Therefore, even those systems relying upon class conditional density knowledge a priori can, at best, estimate the test scenario (and by extension can only estimate the operational conditions the system will be used in later!).

The functional we propose allows a researcher (or tester) who is competing classification systems to evaluate their performance. Each system generates a ROC, or a ROC curve in the case of a family of classification systems, based on the test scenario. The desired scenario of the test organization may be examined with a perturbation in the assumptions (without actually retesting), and functional averages compared as well, so performance can be compared over a range of assumed cost functions and prior probabilities. The result is a sound mathematical approach to comparing classification systems. The functional is scalable to a finite number of classes (the classical detection problem being two classes), with the development of ROC manifolds of dimension $m \geqslant 3$ (making a ROC $(m-1)$-manifold, see Section 5.1). The functional will operate on discrete ROC points in the m-dimensional ROC space as well as over a continuum of ROCs. Ultimately, under certain assumptions and constraints, we will be able to compete classification systems, fusion rules, fusors (fusion rules with a constraint), and fusion systems in order to choose the best from among finitely many competitors.

The relationships between ROCs, ROC curves, and performance has been studied for some time, and some properties are well-known. The foundations for two-class label sets can be reviewed in [3–10]. The method of discovery of these properties are different from our own. Previously, the conditional class density functions were assumed to be known, and differential calculus was applied to demonstrate certain properties. For example, for likelihood-based classifiers, the fact that the slope of a ROC curve at a point is the likelihood ratio which produces this point, was discovered in this manner [3]. Using cost functions in relation to ROC curves to analyze best performance seems to have recently (2001) been recognized by Provost and Foster [11], based on work previously published by [4,10,12]. The main assumption in most of the cited work, with regard to ROC curve properties, is that the distribution functions of the conditional class densities are known and differentiable with respect to the likelihood ratio (as a parameter). We take the approach that, as a beginning for the theory, the ROC curve is continuous and differentiable, and we apply variational calculus to a functional which has the effect of identifying the point on the curve which minimizes Bayes Cost. Under any particular assumption, such a point exists for every family of classification systems. This is not to say the classification system is Bayes Optimal with respect to all possible classification schemes, but rather it is Bayes optimal with respect to the classification systems within the family producing the ROC curve. The solution to the optimization of the functional allows us to extend this property to n-class classification systems, so that we can define and measure performance of ROC manifolds (and the families of classification systems producing them).

We believe this functional (which is really a family of functionals) eliminates the need to discuss classification system performance in terms of area under the curve (AUC), which is so prevalently used in the medical community, or volume under the ROC surface (VUS) [13,14], since these performance 'metrics' do nothing to describe a classification system's value under a specific cost–prior assumption. Any family of classification systems will be set to one particular threshold (at any one time), and so its performance will be measured at only one point on the ROC curve. The question is "What threshold will the user choose?" We submit that this performance can be calculated very quickly under the test conditions desired (using ROC manifolds) by applying vector space methods to the information revealed by the calculus of variations approach.

Additionally, the novelty of this approach also relies on the fact that no class conditional densities are assumed (by the tester), so that the parameters of the functional can be chosen to reflect the desired operational assumptions of interest to the tester. For example, the tester could establish that Neyman–Pearson criteria will form the data of the functional, or that he wants to minimize Bayes cost. The tester may wish to examine performance under a range of hypotheses. Once the data are established, the functional will induce a partial ordering on the set of competing systems.

We have found category theory useful for the description of fusion and fusion systems [1,15–17]. We are not alone in this, since there has already been some groundwork published [18,19] applying category theory to the science of information fusion. Category theory has also been used to prove certain properties of learning and memory using neural nets [20–23]. Category theory is a branch of mathematics useful for demonstrating mathematical relationships and properties of mathematical constructs, such as groups, rings, modules, etc., as well as properties which are universal among like constructs. It is a very useful tool to describe the relationships involved in the systems of classification families. We are using the language of category theory in order to discover universal properties among fusors and to provide mathematical rigor to the definitions. It has been our goal to engage the data fusion community to think in terms of generalities when studying fusion processes in order to abstract the processes and perhaps gain some knowledge and insight to properties that may go undetected otherwise. We have drawn upon the work of various authors in category theory literature [24–27] to present the definitions, which can be found in Appendix A, of this paper.

2. Modelling fusion within the event to label model

Let Ω be a set of states (or outcomes) for a universal (or sure) event, and $T \subset \mathbb{R}$ be a bounded interval of time.

Interval T sorts Ω such that we call $E \subseteq \Omega \times T$ an *event-state*. An event-state is then comprised of event-state elements, $e = (\omega, t) \in E$, where $\omega \in \Omega$ and $t \in T$. Thus, e denotes a state ω at an instant of time t. Let $\Omega \times T$, be the set of all event-states for an event over time interval T. A sigma algebra (or σ-field) $\mathscr{E}$, over $\Omega \times T$ is a collection of subsets of $\Omega \times T$, such that $\Omega \times T \in \mathscr{E}$, and for any $A \in \mathscr{E}$, then its complement, A^c, is also in $\mathscr{E}$. Finally, countable unions of elements of $\mathscr{E}$ are also elements of $\mathscr{E}$. A commonly known σ-field for a set $\Omega \times T$ is its power set, for example. Let $\mathscr{E}$ be a σ-field on $\Omega \times T$, and Pr be a probability measure defined on the measurable space $(\Omega \times T, \mathscr{E})$, then the triple $(\Omega \times T, \mathscr{E}, Pr)$ forms a probability space [28].

The design of a classification system involves the ability to detect (or sense) an event in Ω, and process the detection into a label within a given label set L. For example, design a system that detects airborne objects and classifies them as 'friendly' or 'unfriendly'. To do this, we rely on several mappings, which are composed, to provide the user a classification system (from the event, to the label). Let $E \in \mathscr{E}$ be any member of $\mathscr{E}$, then a sensor is defined as a mapping from E into a (raw) data set D. We denote this with the diagram

$$E \xrightarrow{s} D,$$

so $s(e) = d \in D$ for all $e \in E$. The sensor is defined to produce a specific data type, so the codomain of s, $\mathrm{cod}(s) = D$, where D is the set describing the data output of mapping s. A processor p of this data must have domain, $\mathrm{dom}(p) = D$, and maps to a codomain of features F (a refined data set), $\mathrm{cod}(p) = F$. This is denoted by the diagram

$$D \xrightarrow{p} F.$$

Further, a classifier, c, of this system is a mapping such that $\mathrm{dom}(c) = F$ and $\mathrm{cod}(c) = L$, where L is a set of labels the user of the system finds useful. This is denoted by the diagram

$$F \xrightarrow{c} L.$$

Therefore, we can denote the entire system, which is diagrammed

$$E \xrightarrow{s} D \xrightarrow{p} F \xrightarrow{c} L,$$

as A, the classification system over an event E, where A is the composition of mappings

$$A = c \circ p \circ s.$$

For brevity, we will refer to this as a classification system. Thus, the system A is a (discrete) random variable which maps elements in $E \in \mathscr{E}$ into the labels in L and is diagrammed by

$$E \xrightarrow{A} L.$$

The following discussion can be expanded to a finite number of sensors, but for now consider the simple model of a multi-sensor system using two sensors in Fig. 1. The sets E_i, for $i \in \{1, 2\}$, are sets of event-states. It is useful to think of E_i as the set of possible states of an event (such as an aircraft flying) occurring within a sensor's (or several sensors') field(s) of view. Given E_i thus defined, now define a sensor s_i as a mapping from an event-state set to a data set, D_i. A data set could be a radar signature return of an object, multiple radar signature returns, a two-dimensional image, or even a video stream over the time period of the event-state set. In any case we would like to extract features from the data set. Hence, mapping p_i represents a processor which does just that. Processors are mappings from data sets into feature sets, F_i. One may also think of them as feature extractors. Finally, from the feature sets we want to determine a label or decision based upon the sensed event-state. This is achieved through use of the classifiers c_i which map the feature set into a label set. The label set L_i can be as simple as the two-class set {target, non-target} or could have a more complex nature to it, such as the *types* of targets and non-targets in order to define the battlefield more clearly for the warfighter [29]. Now the diagram in Fig. 1 represents a pair of classification systems having two sensors, two processors, and two classifiers, but can easily be extended to any finite number. Now consider two sensors not necessarily co-located. Hence they may sense different event-state sets. Fig. 1 models two sensors with differing fields of view. Performing fusion along any node or edge in this graph will result in an elevated level of fusion [30]—that of situation refinement or threat refinement, since we are not fusing common information about a particular event or events.

$$\begin{tikzcd}
\mathrm{E}_1 \arrow[r, "s_1"] & \mathrm{D}_1 \arrow[r, "p_1"] & \mathrm{F}_1 \arrow[r, "c_1"] & \mathrm{L}_1 \\
\mathrm{E}_2 \arrow[r, "s_2"] & \mathrm{D}_2 \arrow[r, "p_2"] & \mathrm{F}_2 \arrow[r, "c_2"] & \mathrm{L}_2
\end{tikzcd}$$

Fig. 1. Simple model of a two classification systems.

There are two other possible scenarios that Fig. 1 could depict. The sensors can overlap in their field of view, either partially or fully, in which case fusing the information regarding event-states within the intersection may be useful. Thus, a fusion process may be used to increase the reliability and accuracy of the system, above that which is possessed by either of the sensors on its own. Let E represent that event-state set that is common to both sensors, that is, $E = E_1 \cap E_2$. Hence, there are two basic challenges regarding fusion. The first is how to fuse information from multiple sources regarding common event-states (or target-states, if preferred) for the purpose of knowing the event-state (presumably for the purposes of tracking, identifying, and estimating future event-states). The second and much more challenging problem is to fuse information from multiple sources regarding event-states not common to all sensors, for the purpose of knowing the state of a situation (the situation-state), such as an enemy situation or threat assessment. We distinguish between the two types of fusion scenarios discussed by calling them *event-state fusion* and

$$\begin{tikzcd}
& D_1 \arrow[r, "p_1"] & F_1 \arrow[r, "c_1"] & L_1 \\
E \arrow[ur, "s_1"] \arrow[dr, "s_2"'] & & & \\
& D_2 \arrow[r, "p_2"] & F_2 \arrow[r, "c_2"] & L_2
\end{tikzcd}$$

Fig. 2. Two classification systems with overlapping field of view.

situation-state fusion, respectively. Therefore, Fig. 2 represents the Event-State-to-Label model of two classification systems. The only restriction necessary for the usefulness of this model is that a common field of view be used. Consequently, D_1 and D_2 can actually be the same data set under the model, while s_1 and s_2 could be different sensors.

At this point we begin to consider categories generated by these data sets. Let $\mathscr{D} = (D, \mathbf{Id}_D, \mathbf{Id}_D, \circ)$ be the discrete category generated by data set D. We use these categories to define fusion rules of classification systems.

Definition 1 (*Fusion rule of n* classification systems). Suppose we have n classification systems to be fused. For each $i = 1,\ldots,n$, let $\mathscr{C}_i$ be a category of data generated (if necessary) from the ith source of data (this could be raw data, features, or labels). Then the product

$$\pi(n) = \prod_{i=1}^{n} \mathscr{C}_i$$

is a product category. For a category of data, $\mathscr{C}_0$, the exponential, $\mathscr{C}_0^{\pi(n)}$, is a category of fusion rules, each rule of which maps the products of data objects $\mathbf{Ob}(\pi(n))$ to a data object in $\mathbf{Ob}(\mathscr{C}_0)$, and maps data arrows in $\mathbf{Ar}(\pi(n))$ to arrows in $\mathbf{Ar}(\mathscr{C}_0)$. These fusion rules are functors, which make up the objects of the category. The arrows of the category are natural transformations between them.

If the $\mathscr{C}_i$ are categories generated from sensor sources (i.e., outputs), then we call $\mathscr{C}_0^{\pi(n)}$ a category of data-fusion rules and use the symbols $\mathscr{D}_0^{\pi(n)}$. If they are generated by processor sources, then call $\mathscr{C}_0^{\pi(n)}$ a category of feature-fusion rules and use the symbols $\mathscr{F}_0^{\pi(n)}$. Finally, if they have classifiers as sources, then call them label-fusion rules (or, alternatively, decision-fusion rules) and use the symbols $\mathscr{L}_0^{\pi(n)}$. If we let $\mathscr{C}$ be the category which has as objects the $n+1$ data categories, $\mathscr{C}_i$ for $i = 0,1,\ldots,n$, and with arrows the functors between them, and include the products within this category, then we see that, in particular, the fusion rules are a category of functors (with arrows the natural transformations that may exist between the fusion rules).

A fusion rule could be a Boolean rule, a filter, an estimator, or an algorithm. There is no restriction on the output, with regard to being a "better" output than a system designed without a fusion rule, since that requires a new definition. We now desire to show how defining a fusor (see Definition 5) as a fusion rule with a constraint changes the classification system model into an event-state fusion model. Continuing to consider the two classification families in Fig. 2, a fusion rule can be applied to either the data sets or the feature sets. Given a fusion rule $\mathfrak{R}$ for the two data sets as in Fig. 3, our model becomes that of Fig. 4. Notice we use a different arrow to denote a fusion rule. A new data set, processor, feature set, and classifier may become necessary as a result of the fusion rule having a different codomain than the previous systems. The label set may change also, but for now, consider a two-class label set, that of

$$\begin{tikzcd}
D_1 \times D_2 \arrow[r, dotted, two heads, "\mathfrak{R}"] & D_3
\end{tikzcd}$$

Fig. 3. Fusion rule applied to data sets.

$$\begin{tikzcd}
& D_1 & & & \\
E \arrow[ur, "s_1"] \arrow[dr, "s_2"'] \arrow[r, dotted, two heads, "\mathfrak{R}"] & D_3 \arrow[r, "p"] & F \arrow[r, "c"] & L \\
& D_2 & & &
\end{tikzcd}$$

Fig. 4. Fusion rule applied within a dual sensor process.

$$L = L_1 = L_2 = \{\text{Target},\ \text{Non-target}\}.$$

In a within-fusion scenario (see [31]), the data sets are identical, $D_1 = D_2 = D_3$. This is easily seen in the case where two sensors are the same type (that is, they collect the same measurements, but from possibly different locations relative to the overlapping field of view). In the case where the data sets are truly different, a composite data set which is different from the first two (possibly even the product of the first two) is created as the codomain of the fusion rule.

At this point we may consider, in what way is the system in Fig. 4 **superior** to the original systems shown in Fig. 2 with $L = L_1 = L_2$? One way of comparing performance in such systems is to compare the systems' receiver operating characteristics (ROC) curves.

3. Developing a ROC curve

Setting aside the fusion of classification systems for a moment, we focus on a generic classification of an event-state itself. Let $(E, \mathscr{E}, Pr)$ be a probability space. Let the label set $L = \{T, N\}$, where $T =$ target and $N =$ non-target. Let $E = E_{\mathrm{T}} \cup E_{\mathrm{N}}$, where $E_{\mathrm{T}} \cap E_{\mathrm{N}} = \emptyset$, so that $\{E_{\mathrm{T}}, E_{\mathrm{N}}\}$ is a partition of E into the two classes. Let c be a classifier such that

$$E \xrightarrow{c} L$$

is a classification system. Recall that the mapping c induces a "natural" mapping, which we denote $c^{\natural}$, the pre-image of c ($\natural$ is the natural symbol in music, a becuadro in Spanish, and we use it to distinguish from the inverse symbol $^{-1}$. It is possible that the inverse exists, but not guaranteed). Hence, if $\mathscr{L}$ is the power set of L, then

$$c^{\natural} : \mathscr{L} \to E.$$

Thus, we can calculate the probability of true positive,

$$P_{\text{tp}} = \frac{Pr(c^{\natural}(T) \cap E_{\text{T}})}{Pr(E_{\text{T}})} = 1 - \frac{Pr(c^{\natural}(N) \cap E_{\text{T}})}{Pr(E_{\text{T}})} \tag{1}$$

which is estimated by the true positive rate (TPR), and the probability of false positive,

$$P_{\text{fp}} = \frac{Pr(c^{\natural}(T) \cap E_{\text{N}})}{Pr(E_{\text{N}})} \tag{2}$$

which is estimated by the false positive rate (FPR), or false alarm rate. The ordered pair $(P_{\text{fp}}, P_{\text{tp}}) \in [0,1] \times [0,1]$ is the ROC for the system. Now it is desirable for a classification system to have a parameter associated with the classifier, such that changing the parameter (which is possibly multi-dimensional) changes the ROC. In such a case, a parameter set Θ would be chosen such that the family of classification systems, $\mathbb{C} = \{c_\theta | \theta \in \Theta\}$, maps the event set into the label set, and such that the curve $f = \{(P_{\text{fp}}(c_\theta), P_{\text{tp}}(c_\theta)) : \theta \in \Theta\}$ is the projection of the trajectory $\tau = \{(\theta, P_{\text{fp}}(c_\theta), P_{\text{tp}}(c_\theta)) : \theta \in \Theta\}$ into the P_{fp}–P_{tp} plane. In this case, we have that

$$P_{\text{tp}}(c_\theta) = \frac{Pr(c_\theta^{\natural}(T) \cap E_{\text{T}})}{Pr(E_{\text{T}})}, \tag{3}$$

and

$$P_{\text{fp}}(c_\theta) = \frac{Pr(c_\theta^{\natural}(T) \cap E_{\text{N}})}{Pr(E_{\text{N}})}. \tag{4}$$

We call such a parameter set an *admissible* parameter set if the image of $P_{\text{fp}}(c_\theta)$ is onto $[0,1]$. Note the parameter need not necessarily be associated with the classifier of the system, but could be associated instead with the sensor(s), processor(s), or any combination of the three. Consider, for example, the three classification systems in Fig. 5. Each system will generate a ROC curve when Θ is an *admissible* parameter set. What is key is that the final parameter set must produce a corresponding ROC curve as a continuous curve from $(0,0)$ through $(1,1)$ in the $P_{\text{fp}} - P_{\text{tp}}$ plane as the example in Fig. 6 shows. The parameter θ is the threshold of the ROC. We can, at this point, advocate that the ROC curve be designated by the classification system that generates it, and not just the classifier. Therefore, the systems in Fig. 5 will be designated as in Fig. 7, a shorthand method for describing the composition of sensor(s), processor(s), and classifier(s), that is, $A_{1,\theta} = c \circ p \circ s_\theta$, $A_{2,\theta} = c \circ p_\theta \circ s$, and $A_{3,\theta} = c_\theta \circ p \circ s$.

$$\begin{tikzcd} E \arrow[r, "s_\theta"] & D \arrow[r, "p"] & F \arrow[r, "c"] & L \end{tikzcd}$$

$$\begin{tikzcd} E \arrow[r, "s"] & D \arrow[r, "p_\theta"] & F \arrow[r, "c"] & L \end{tikzcd}$$

$$\begin{tikzcd} E \arrow[r, "s"] & D \arrow[r, "p"] & F \arrow[r, "c_\theta"] & L \end{tikzcd}$$

Fig. 5. Three classification systems with admissible parameters each produce a ROC curve.

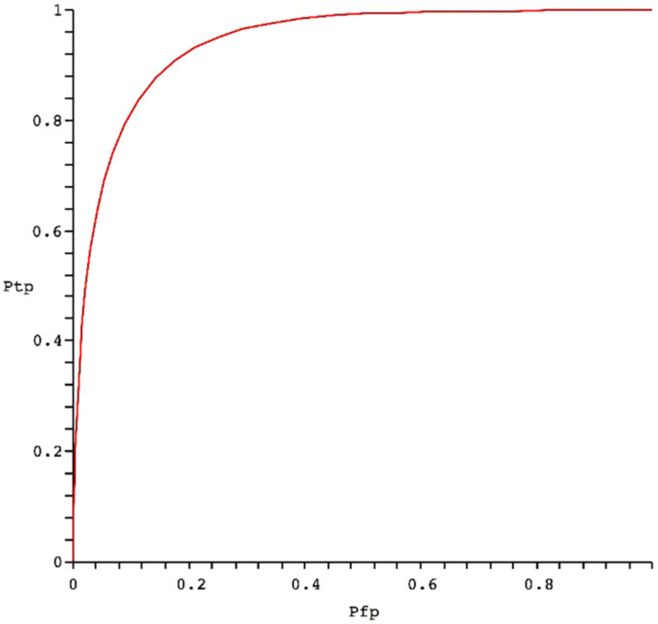

Fig. 6. A typical ROC curve.

Assume that $(E, \mathscr{E}, Pr)$ is a probability space, and $E = E_{\text{T}} \cup E_{\text{N}}$, with $E_{\text{T}} \cap E_{\text{N}} = \emptyset$, so that $\{E_{\text{T}}, E_{\text{N}}\} \subset \mathscr{E}$ is a partition of E into the two classes of events. Let $\mathbb{A} = \{A_\theta | \theta \in \Theta\}$, where $E \xrightarrow{A_\theta} L$. Is there a threshold, $\theta^* \in \Theta$, such that A_{θ^*} performs best in the family of classification systems, $\mathbb{A}$? It is well-known and accepted that the threshold for which the probability of a misclassification (or Bayes error) is minimized is considered best and denoted the Bayes optimal threshold (BOT). That is, does there exist $\theta^* \in \Theta$ which minimizes the quantity

$$\begin{aligned} &Pr(A_\theta^{\natural}(E_{\text{T}}) \cap E_{\text{N}}) \cup (A_\theta^{\natural}(E_{\text{N}}) \cap E_{\text{T}}) \\ &\quad = Pr(A_\theta^{\natural}(E_{\text{T}}) \cap E_{\text{N}}) + Pr(A_\theta^{\natural}(E_{\text{N}}) \cap E_{\text{T}}) \\ &\quad = P_{\text{fp}}(A_\theta) Pr(E_{\text{N}}) + (1 - P_{\text{tp}}(A_\theta)) Pr(E_{\text{T}}), \end{aligned} \tag{5}$$

where $Pr(E_{\text{T}})$ and $Pr(E_{\text{N}})$ are the prior probabilities of the target class and non-target class, respectively? If yes, then θ^* is the BOT for the family of classification systems $\mathbb{A}$.

$$\begin{tikzcd} E \arrow[r, "A_{1,\theta}"] & L \end{tikzcd}$$

$$\begin{tikzcd} E \arrow[r, "A_{2,\theta}"] & L \end{tikzcd}$$

$$\begin{tikzcd} E \arrow[r, "A_{3,\theta}"] & L \end{tikzcd}$$

Fig. 7. Three classification systems.

An obvious question at this point is, given two families of classification systems, $\mathbb{A} = \{A_\theta | \theta \in \Theta\}$ and $\mathbb{B} = \{B_\phi | \phi \in \Phi\}$, which family is best? This is not an easy question to answer, as demonstrated in [32]. It is tempting to use some measure of the BOT, but notice that the BOT is dependent upon the selection of prior probabilities. The priors are generally not known, so selection of a better classification system based on ROC curves may not be possible, since ROC curves for different families can overlap. Rather, we should ask the question, given an operating assumption of our prior probabilities, such as $Pr(E_{\mathrm{T}}) = \frac{1}{4}$, can we choose among competing families of classification systems one that is superior to the others? One way to answer the question is derived in an unexpected way.

4. A variational calculus solution to determining the Bayes optimal threshold of a family of classification systems

Suppose the ROC curves are smooth (differentiable) over the entire range, i.e., we consider the set $X = \{f : [0,1] \to \mathbb{R} | f$ is differentiable at each $x \in (0,1)$ and its derivative f' is continuous at each $x \in (0,1)\}$. This is consistent with Alsing's proof [32] in that, given enough data all ROC curve estimates converge to the ROC curve. Given a diagram describing the family of classification systems $\mathbb{A} = \{A_\theta : \theta \in \Theta\}$, with Θ an admissible parameter set (assumed to be one-dimensional), and $(E, \mathscr{E}, Pr)$ a probability space of features, there is a set $\tau_{\mathbb{A}} = \{(\theta, P_{\mathrm{fp}}(A_\theta), P_{\mathrm{tp}}(A_\theta)) : \theta \in \Theta\}$ which is called the *ROC trajectory* for the classification system family $\mathbb{A}$. The projection of the ROC trajectory onto the $(P_{\mathrm{fp}}, P_{\mathrm{tp}})$-plane is the set $f_{\mathbb{A}} = \{(P_{\mathrm{fp}}(A_\theta), P_{\mathrm{tp}}(A_\theta)) : \theta \in \Theta\}$ which is the ROC curve of the classification system family $\mathbb{A}$. Hence, for $h \in [0,1]$ such that $h = P_{\mathrm{fp}}(A_\theta)$ for some $\theta \in \Theta$, we have that

$$[P_{\mathrm{fp}}]^{\natural}(\{h\}) = \{A_\theta\} \iff \theta,$$

that is, the pre-image of h under $P_{\mathrm{fp}}(\,\cdot\,)$ is the classification system A_θ, which we assume has a one-to-one and onto correspondence to θ. Therefore, the BOT of the family of classification systems $\mathbb{A}$, denoted by θ^*, corresponds to some $h^* = P_{\mathrm{fp}}(A_{\theta^*}) \in [0,1]$, which may not be unique, unless the function $P_{\mathrm{fp}}(\,\cdot\,)$ is one-to-one. So, there is at least one such h^*, now what can we learn about it? Consider the problem stated as follows:

> Let $\alpha, \beta \geqslant 0$. Among all smooth curves which originate on the point $(0,1)$ and terminate on the ROC curve f, find the curve, defined by the function Y, which minimizes the functional
>
> $$J[Y] = \int_0^h [\alpha + \beta|\dot{Y}(x)|]\,\mathrm{d}x. \tag{6}$$
>
> The value of $h \in [0,1]$ depends on the function Y which satisfies the constraints
>
> $$\begin{aligned} Y(0) &= 1, \\ Y(h) &= f(h). \end{aligned} \tag{7}$$
>
> Observe that $h = P_{\mathrm{fp}}(A_\theta), f(h) = P_{\mathrm{tp}}(A_\theta)$ for some $\theta \in \Theta$, and $\beta = 1 - \alpha$ with $\alpha = Pr(N)$, the prior probability of a non-target.

The functional J identifies the curve with the smallest arclength (measured with respect to the weighted 1-norm) from the point $(0,1)$ to the ROC curve. The constraints of Eq. (7) imply that the curve must begin at $(0,1)$ and terminate on the ROC curve. Any differentiable function Y that minimizes J, subject to the constraints (7), necessarily must be a solution to Euler's equation [33]

$$\frac{\partial}{\partial y} G(x, Y(x), \dot{Y}(x)) - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial}{\partial z} G(x, Y(x), \dot{Y}(x)) = 0$$
$$\text{for all } x \in (0,h). \tag{8}$$

From Eq. (6) we define $G(x,y,z) = \alpha + \beta|z|$, so that $\frac{\partial}{\partial y} G = 0$ and $\frac{\partial}{\partial z} G = \beta\mathbf{sgn}(z)$. Hence, we have that Y solves the Euler equation

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{sgn}(\dot{Y}(x)) = 0 \quad \text{for all } x \in (0,h). \tag{9}$$

Integrating this equation yields $\mathbf{sgn}(\dot{Y}(x))$ is constant for all $x \in [0,h]$. Since $Y(x) \leqslant 1$ for all $x \in (0,h)$, and $Y(0) = 1$, from constraints (7), then $\mathbf{sgn}(\dot{Y}(x))$ must be 0 or -1. Now, if $\mathbf{sgn}(\dot{Y}(x)) = 0$ for all x, then $1 = Y(0) = Y(h) = Y(1)$ due to the smoothness of the ROC curve. Substituting this solution into the functional J in Eq. (6) yields

$$J[Y] = \alpha h = Pr(N)P_{\mathrm{fp}}(A_\theta) \tag{10}$$

with $P_{\mathrm{fp}}(A_\theta) = 1$. Thus, $J[Y] = Pr(N)$ and the weighted (1-norm) arclength of curve Y is therefore $Pr(N)$. On the other hand, if $\mathbf{sgn}(\dot{Y}(x)) = -1$ for all $x \in (0,h)$, then $|\dot{Y}(x)| = -\dot{Y}(x)$ and substituting this into J directly in Eq. (6) yields

$$J[Y] = \int_0^h [\alpha - \beta\dot{Y}(x)]\,\mathrm{d}x \tag{11}$$
$$\begin{aligned} &= \alpha h + \beta\mathbf{sgn}(\dot{Y}(h))Y(h) - \beta\mathbf{sgn}(\dot{Y}(0))Y(0) \\ &= \alpha h - \beta[Y(h) - Y(0)] \\ &= \alpha h + [1 - Y(h)]\beta \\ &= P_{\mathrm{fp}}(A_\theta)Pr(N) + (1 - P_{\mathrm{tp}}(A_\theta))Pr(T). \end{aligned} \tag{12}$$

Notice that Eq. (12) is identical to the unminimized Eq. (5). Therefore, $h = h^*$ which minimizes Eq. (12) corresponds to the BOT, θ^*, of the classification system family $\mathbb{A}$. The transversality condition [33] of this problem is

$$\alpha + \beta|\dot{Y}(x)||_{x=h^*} + \beta(f'(x) - \dot{Y}(x))\mathbf{sgn}(\dot{Y}(x))|_{x=h^*} = 0, \tag{13}$$

so that

$$f'(h^*) = \frac{\alpha}{\beta} \tag{14}$$

which is

$$f'(h^*) = \frac{Pr(N)}{Pr(T)}. \tag{15}$$

This is a global minimum, since, it is clear the J is a convex functional. So the transversality condition tells us that the BOT of the family of classification systems corresponds to a point on the ROC curve which has a derivative equal to the ratio of prior probabilities, $\frac{Pr(N)}{Pr(T)}$. Therefore, if one presumes a ratio of prior probabilities equal to 1, then the point on the curve corresponding to the BOT will have a tangent to the ROC curve with slope 1. We could substitute $\alpha = C_{\text{fp}}P(N)$ and $\beta = C_{\text{fn}}P(T)$ where C_{fp} and C_{fn} are the costs of making each error, or we could specify a cost–prior ratio $\frac{C_{\text{fp}}P(N)}{C_{\text{fn}}P(T)}$, if we wish to consider costs in addition to the prior probabilities. This gives us an idea of what would make a good functional for determining which families of classification systems are more desirable than others. An immediate approach would be to choose a preferred prior ratio and construct a linear variety through the optimal ROC point $(0,1)$. Then take the 2-norm of the vector which minimizes the distance from the ROC curve to the linear variety. However, it is still possible that many ROC curves could be constructed so that the BOT for each one has the same distance to the linear variety. This would set up an equivalence class of families of classification system, so that this distance induces a partial order of these families. This is similar to the problem faced when using area under the curve (AUC) of a ROC curve as a functional. In both cases the underlying posterior conditional probabilities are unknown and there are just too many possible combinations of posterior distributions that can produce ROC curves with the same AUC (or equal BOT functional values). The point, however, is that under a functional based on the BOT, we would have a "leveled playing field" since we are debating which ROC (and therefore the classification system it represents) is better based on the *same* prior probabilities. Families of classification systems with equal AUC are considered equal over the entire range of possible priors and therefore, AUC is of less value. Furthermore, the AUC functional does not relate its values to the unknown priors at all. Rather, it is related to the value of the class conditional probabilities associated with a family of classification systems over **all** possible false positive values. It is therefore essentially useless as a functional in trying to discover an appropriate operating threshold for a classification system.

5. ROC manifolds

5.1. Constructing the three-dimensional ROC manifold

So far we have considered the fusion of only those classification systems which produce a two-class output. What if there were a choice of three classes with the corresponding label set $L = \{\ell_1, \ell_2, \ell_3\}$?

Let $(E, \mathscr{E}, Pr)$ be the probability space over which we will apply a classification system. Let $\mathbf{\Theta} = \Theta_1 \times \Theta_2$ be an admissible parameter set for a family of classification systems. We will use the generalized approach and build a ROC manifold from the family of classification systems $E \overset{A_{\mathbf{\Theta}}}{\to} L$, where $L = \{\ell_1, \ell_2, \ell_3\}$ is a three-class label set. Let $\{E_1 E_2, E_3\}$ be a partition of E, so that E_i corresponds to class ℓ_i. We say $P_{i|j}(A_{\boldsymbol{\theta}})$ is the conditional probability of the classification system $A_{\boldsymbol{\theta}}$ labeling an elementary event $e \in E$ as class i when event E_j has occurred, that is

$$P_{i|j}(A_{\boldsymbol{\theta}}) = \frac{Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_i) \cap E_j)}{Pr(E_j)},$$

where $Pr(E_j)$ is the prior probability of class ℓ_j.

Consider now the Bayes error of the classification system. First, we define the Bayes error function as

$$\begin{aligned}\mathrm{BE}(A_{\boldsymbol{\theta}}) &= Pr[(A_{\boldsymbol{\theta}}^{\natural}(\ell_1) \cap E_3) \cup (A_{\boldsymbol{\theta}}^{\natural}(\ell_2) \cap E_3) \cup (A_{\boldsymbol{\theta}}^{\natural}(\ell_1) \cap E_2) \\ &\quad \cup (A_{\boldsymbol{\theta}}^{\natural}(\ell_3) \cap E_2) \cup (A_{\boldsymbol{\theta}}^{\natural}(\ell_2) \cap E_1) \cup (A_{\boldsymbol{\theta}}^{\natural}(\ell_3) \cap E_1)] \\ &= Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_1) \cap E_3) + \mathrm{Pr}(A_{\boldsymbol{\theta}}^{\natural}(\ell_2) \cap E_3) \\ &\quad + Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_1) \cap E_2) + Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_3) \cap E_2) \\ &\quad + Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_2) \cap E_1) + Pr(A_{\boldsymbol{\theta}}^{\natural}(\ell_3) \cap E_1) \\ &= \left[1 - P_{3|3}(A_{\boldsymbol{\theta}}^{\natural}(\ell_3))\right] Pr(E_3) + \left[1 - P_{2|2}(A_{\boldsymbol{\theta}}^{\natural}(\ell_3))\right] Pr(E_2) \\ &\quad + \left[1 - P_{1|1}(A_{\boldsymbol{\theta}}(\ell_3))\right] Pr(E_1).\end{aligned} \tag{16}$$

We wish to minimize the Bayes error function over the admissible set of parameters. If $\boldsymbol{\theta}^* \in \mathbf{\Theta}$ minimizes $\mathrm{BE}(A_{\boldsymbol{\theta}})$, then $\boldsymbol{\theta}^*$ is called the BOT for the family of classification systems $\{A_{\boldsymbol{\theta}} | \boldsymbol{\theta} \in \mathbf{\Theta}\}$. If the errors within class have the same cost (as is the case with this construction) then we can have three error axes, $1 - P_{3|3}(A_{\boldsymbol{\theta}})$, $1 - P_{2|2}(A_{\boldsymbol{\theta}})$, and $1 - P_{1|1}(A_{\boldsymbol{\theta}})$. Assume the first parameter θ_1 corresponds to axis $1 - P_{1|1}(A_{\boldsymbol{\theta}})$, and the second parameter θ_2 corresponds to the axis $1 - P_{2|2}(A_{\boldsymbol{\theta}})$. If error costs are not equal, then we need $3^2 - 3 = 6$ axes for our ROC space and five free parameters to create a ROC 5-manifold. This leads us to extend the calculus of variations approach identifying the corresponding BOT for a given set of prior probabilities.

5.2. Extending the calculus of variations approach to ROC manifolds

For n classes, we specify $m = n^2 - n$ dimensions assuming differing costs for each error. Each axis then is a type of false positive for a given class, so that the optimal ROC point is the origin, $(0,0,\ldots,0)$. The method used in this section follows and extends that of [34]. Let $x_m = f(x_1, x_2, \ldots, x_{m-1})$ be the equation of the ROC manifold residing in m-dimensional space. Define the function

$$\Psi(x_1, x_2, \ldots, x_m) \doteq f(x_1, x_2, \ldots, x_{m-1}) - x_m,$$

then $\mathfrak{M} = \{(x_1, \ldots, x_m) | \Psi(x_1, \ldots, x_m) = 0\}$ is the ROC manifold. We assume all first-order partial derivatives exist and are continuous for Ψ. For each $t \in [0,1]$ let $\mathbf{R}(t) = (X_1(t), \ldots, X_m(t))$ be the position vector which points to a point on a smooth trajectory, beginning at the initial point $(0, \ldots, 0)$ and terminating on the manifold $\mathfrak{M}$. Thus, $\mathbf{R}(0) = (0, \ldots, 0)$ and there is some $t_{\text{f}} \in (0,1]$ such that $\mathbf{R}(t_{\text{f}}) \in \mathfrak{M}$, with t_{f} dependent upon the particular $\mathbf{R}$.

Choose weights $a_i > 0$ for $i = 1, \ldots, m$ such that $\sum_{i=1}^{m} a_i \leqslant 1$, and let $\|\cdot\|_1$ denote the weighted 1-norm defined on a vector $\mathbf{v} = (v_1, \ldots, v_m)$ by

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{m} a_i |v_i|. \tag{17}$$

Define the functional J_1

$$J_1[\mathbf{R}] = \int_0^{t_f} \|\dot{\mathbf{R}}(t)\|_1 \,\mathrm{d}t. \tag{18}$$

For ease of notation, define

$$G(t, \mathbf{x}, \mathbf{y}) = \|\mathbf{y}\|_1.$$

Hence, we write Eq. (18) as

$$J_1[\mathbf{R}] = \int_0^{t_f} G \,\mathrm{d}t, \tag{19}$$

and suppress the integrand variables as is customary in the calculus of variables. We wish to minimize J_1, and seek to find the optimal function $\mathbf{R}^*$ with initial and terminal points as discussed subject to the constraints. The mathematics describing this process are shown in Appendix B. Once the Euler equations with transversality conditions are solved, we have the result that

$$\nabla \Psi(\mathbf{R}^*(t_f^*)) = \frac{-1}{a_m}(a_1, \ldots, a_m) = \mathbf{n} \tag{20}$$

is the normal to the ROC manifold $\mathfrak{M}$ at the terminal point of $\mathbf{R}^*(t_f^*)$ on the smooth trajectory minimizing J_1. The weights, a_i, are the product of a prior probability of a particular class and a cost of the particular error.

6. A functional for comparing classifier families

Having shown using calculus of variations that the optimal points of the ROC manifold can be found corresponding to a simple normal vector, based on the initial data of prior probabilities and costs, we turn to developing a functional, which will calculate a value in $\mathbb{R}$ corresponding to the optimal value of the cost. The functional requires input of prior probabilities, costs, and constraints. The ROC manifold point which yields a minimum norm is the point on the ROC manifold which is optimal under the assumed data. Thus, families of classification systems can be compared using this functional, and the best classification system (and perhaps the best operating parameter) can be chosen. When the best classification system is chosen from competing families of classification systems which use fusion rules, these rules are in essence competed against each other and against the original systems without fusion. This enables a definition of fusors which is given in Section 7. Recall that the main point of this paper is to generalize a method to compete families of classification systems with the specific intent to define and compete fusion rules. The functional we propose will do this, however, other functionals can be developed as well. Ultimately, once the functional, along with its associated data is chosen, one has a way of defining fusion (and what we call fusors) for the given problem.

Let $n \in \mathbb{N}$ be the number of classes of interest, and $m = n^2 - n$. We construct the functional over the space $X = C([0,1]^{m-1}, \mathbb{R}) \cup C^1((0,1)^{m-1}, \mathbb{R})$, recognizing that we are competing ROC curves, which are by definition a subset of X. The functional

$$F_2 : X \to \mathbb{R},$$

where $n = 2$ is the number of classes, is denoted $F_2(\cdot; \gamma_1, \gamma_2, \alpha, \beta)$ for the ROC curves corresponding to a two-class family of classification systems, where $\gamma_1 = C_{2|1} Pr(\ell_1)$ is the cost of the error of declaring class E_2 when the class is truthfully E_1 times the prior probability of class E_1, $\gamma_2 = C_{1|2} Pr(\ell_2)$ is the cost of the error of declaring class E_1 when the class is truthfully E_2 times the prior probability of class E_2, while $\alpha = P_{1|2}$ and $\beta = P_{1|1}$ are the acceptable limits of false positive and true positive rates. Without loss of generality, we assume γ_1 to be the dependent constraint. The quadruple $(\gamma_1, \gamma_2, \alpha, \beta)$ comprises the *data* of the functional F_2.

Definition 2 (*ROC curve functional*). Let $(\gamma_1, \gamma_2, \alpha, \beta)$ be given data. Let

$$\mathbf{y}_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{\Gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix},$$

and

$$V_{\boldsymbol{\Gamma}} = \{\mathbf{v} | \mathbf{v} = k\boldsymbol{\Gamma} \ \forall k \in \mathbb{R}\}.$$

Then $V_{\boldsymbol{\Gamma}} + \mathbf{y}_0$ is a linear variety through the supremum ROC point, $(0,1)$, over all possible ROC curves, under the data. Let $f \in X$ and let f be non-decreasing. Let $\mathscr{R}(f)$ be the range of f, and let

$$\mathbf{T} = ([0, \alpha] \times [\beta, 1]) \cap \mathscr{R}(f).$$

Let $z_{\boldsymbol{\Gamma}} = \min_{\substack{\mathbf{v} \in V_{\boldsymbol{\Gamma}} \\ \mathbf{y} \in \mathbf{T}}} \|\mathbf{v} + \mathbf{y}_0 - \mathbf{y}\|_2$. Then define

$$F_2(\cdot; \gamma_1, \gamma_2, \alpha, \beta) : X \to \mathbb{R}$$

by

$$F_2(f; \gamma_1, \gamma_2, \alpha, \beta) = \sqrt{2} - z_{\boldsymbol{\Gamma}} \quad \forall f \in X. \tag{21}$$

It should be clear that the constant $\sqrt{2}$ is the largest theoretical distance from all linear varieties to a curve in ROC space.

So far, it is shown that F_n is minimal at the Bayes optimal point of the ROC curve under no constraints restricting the values possible for it to take on in ROC space (i.e., $\alpha = 1$ and $\beta = 0$ in the two-class case, and $\boldsymbol{\alpha} = (1, \ldots, 1)$ in the n-class case). We can now relate this functional to the Neyman–Pearson (N–P) criteria. Recall that the N–P criteria is also known as the most powerful test of size α_0, when α_0 is the a priori assigned maximum false positive rate [5]. Given a family of classification systems $\mathbb{A} = \{A_\Theta : \theta \in \Theta\}$, the N–P criteria could be written as

$$\max_{\theta \in \Theta} P_{1|1}(A_\theta) \text{ subject to } P_{1|2}(A_\theta) \leqslant \alpha_0.$$

Theorem 3 (ROC functional-Neyman–Pearson equivalence). *Let γ_1 be the dependent constraint, and $\sum_{i=1}^{2}\gamma_i \leqslant 1$. The ROC functional $F_2(\cdot;\gamma_1,\gamma_2,\alpha,\beta)$ under data $(1,0,\alpha_0,0)$ yields the same point on a ROC curve as the Neyman–Pearson criteria with $\alpha \leqslant \alpha_0$.*

Proof. Suppose $(\gamma_1,\gamma_2,\alpha,\beta)=(1,0,\alpha_0,0)$. Then $\Gamma=(1,0)$ and

$$V_\Gamma=\left\{\mathbf{v}\,|\,\mathbf{v}=\begin{pmatrix}k\\0\end{pmatrix}\ \forall k\in\mathbb{R}\right\},$$

and let

$$\mathbf{y}_0=\begin{pmatrix}0\\1\end{pmatrix}.$$

Thus, $V_\Gamma+\mathbf{y}_0$ is the appropriate linear variety. Let

$$\mathbf{T}=([0,\alpha_0]\times[0,1])\cap\mathscr{R}(f),$$

where f is a ROC curve and consider $\beta_{\mathrm{N}}\in f([0,\alpha_0])$ as the optimal point in the image of f under the N–P criteria. Then $z_{\mathrm{N}}=1-\beta_{\mathrm{N}}$ is the distance to $V_\Gamma+\mathbf{y}_0$. Now,

$$F_2(f)=\sqrt{2}-z_\Gamma,$$

where

$$z_\Gamma=\min_{\substack{\mathbf{v}\in V_\Gamma\\ \mathbf{y}\in\mathbf{T}}}\|\mathbf{v}+\mathbf{y}_0-\mathbf{y}\|_2.$$

Thus, we have that $\beta_{\mathrm{N}}\geqslant\beta$, $\forall\beta=f(\alpha)$ $\forall\alpha\leqslant\alpha_0$. Hence, $1-\beta_{\mathrm{N}}\leqslant 1-\beta$, $\forall\beta=f(\alpha)$, $\forall\alpha\leqslant\alpha_0$. Then for

$$\mathbf{y}_{\mathrm{N}}=\begin{pmatrix}\alpha_{\mathrm{N}}\\ \beta_{\mathrm{N}}\end{pmatrix},$$

we have that

$$\begin{aligned}[(1-\beta_{\mathrm{N}})^2]^{1/2}&=\left\|\begin{pmatrix}\alpha_{\mathrm{N}}\\1\end{pmatrix}-\begin{pmatrix}\alpha_{\mathrm{N}}\\ \beta_{\mathrm{N}}\end{pmatrix}\right\|\\&=\left\|\begin{pmatrix}\alpha_{\mathrm{N}}\\1\end{pmatrix}-\mathbf{y}_{\mathrm{N}}\right\|\leqslant\left\|\begin{pmatrix}\alpha_{\mathrm{N}}\\1\end{pmatrix}-\begin{pmatrix}\alpha\\ \beta\end{pmatrix}\right\|\end{aligned}$$

$\forall\beta=f(\alpha)$, $\forall\alpha\leqslant\alpha_0$. Thus, letting $\mathbf{y}=\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$ we have that

$$\min_{\alpha\leqslant\alpha_0}\left\|\begin{pmatrix}\alpha\\1\end{pmatrix}-\mathbf{y}_{\mathrm{N}}\right\|\leqslant\min_{\substack{\alpha\leqslant\alpha_0\\ \mathbf{y}\in[0,\alpha_0]\times f([0,\alpha_0])}}\left\|\begin{pmatrix}\alpha\\1\end{pmatrix}-\mathbf{y}\right\| \tag{22}$$

$$=\min_{\substack{\alpha\leqslant\alpha_0\\ \mathbf{y}\in[0,\alpha_0]\times f([0,\alpha_0])}}\left\|\begin{pmatrix}\alpha\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}-\mathbf{y}\right\| \tag{23}$$

$$=\min_{\substack{\mathbf{v}\in V_\Gamma\\ \mathbf{y}\in[0,\alpha_0]\times f([0,\alpha_0])}}\|\mathbf{v}+\mathbf{y}_0-\mathbf{y}\|. \tag{24}$$

On the other hand,

$$\min_{\substack{\mathbf{v}\in V_\Gamma\\ \mathbf{y}\in[0,\alpha_0]\times f([0,\alpha_0])}}\|\mathbf{v}+\mathbf{y}_0-\mathbf{y}\|\leqslant\min_{\mathbf{v}\in V_\Gamma}\|\mathbf{v}+\mathbf{y}_0-\mathbf{y}_{\mathrm{N}}\| \tag{25}$$

$$\leqslant\left\|\begin{pmatrix}\alpha_{\mathrm{N}}\\1\end{pmatrix}-\mathbf{y}_{\mathrm{N}}\right\| \tag{26}$$

$$=\left\|\begin{pmatrix}0\\1-\beta_{\mathrm{N}}\end{pmatrix}\right\| \tag{27}$$

$$=1-\beta_{\mathrm{N}}. \tag{28}$$

Therefore, we have that

$$z_\Gamma=\min_{\substack{\mathbf{v}\in V_\Gamma\\ \mathbf{y}\in[0,\alpha_0]\times f([0,\alpha_0])}}\|\mathbf{v}+\mathbf{y}_0-\mathbf{y}\|=1-\beta_{\mathrm{N}}.$$

But $z_\Gamma=1-\beta_{\mathrm{R}}$, so that $\beta_{\mathrm{R}}=\beta_{\mathrm{N}}$. So, we have that the ROC functional, under data $(1,0,\alpha_0,0)$, acting on a ROC curve corresponds to the power of the most powerful test of size α_0. □

The calculation and scalability of the functional is straightforward. Suppose we have n classes. In the two-class case, one axis may be chosen as $P_{1|1}$, but in the n-class case, each axis is an error axis. This is absolutely necessary in the case where costs of errors differ within a class. If we apply this methodology to the two-class case, the two axes would be $P_{1|2}$ and $P_{2|1}$ with the ROC curve starting at point $(0,1)$ and terminating at point $(1,0)$. A ROC at the origin would represent the perfect classification system under this scheme. For the n-class case, we have $m=n^2-n$ error axes. Without loss of generality, we choose the conditional class probability of class n given $n-1$ to be the dependent variable, so that γ_m is the cost–prior product associated with $p_{n|n-1}$. Let $m=k^2-k$. Let $\mathbf{d}=(\gamma_1,\ldots,\gamma_m,\alpha_1,\ldots,\alpha_m)$ be the data, and let each $r=1,2,\ldots,m$ be associated with one of the m pairs, (i,j), where for each $i=1,2,\ldots,k$ with $i\neq j$, we have a $j=1,2,\ldots,k$. Let α_m be associated with $p_{k|k-1}$. Then let $\mathbf{q}=(q_1,\ldots,q_m)$, so that

$$Q=\Big\{\mathbf{q}|q_r=p_{i|j},\ r=1,\ldots,m;\ p_{i|j}\leqslant\alpha_r,r\neq m;\\ i,j=1,\ldots,k;\ i\neq j\Big\}$$

be the set of points comprising the ROC curve within the constraints. Then we have that $\mathbf{y}_0=(0,0,\ldots,0)$ and

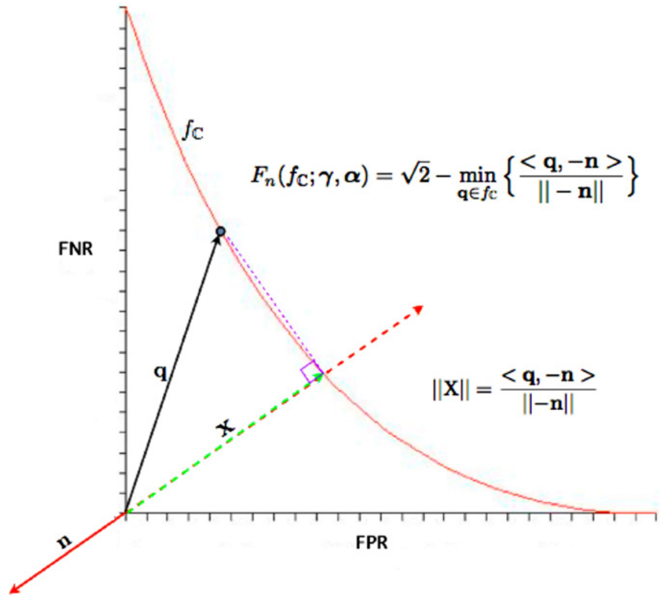
f_{C}
$F_n(f_{\mathrm{C}};\gamma,\alpha)=\sqrt{2}-\min_{\mathbf{q}\in f_{\mathrm{C}}}\left\{\frac{<\mathbf{q},-\mathbf{n}>}{\|-\mathbf{n}\|}\right\}$
FNR
q
X
$\|\mathbf{X}\|=\frac{<\mathbf{q},-\mathbf{n}>}{\|-\mathbf{n}\|}$
n
FPR

Fig. 8. Geometry of calculating the ROC functional, F_2 for a point (with vector $\mathbf{q}$) on ROC curve f_{C}.

$\mathbf{n} = \frac{-1}{\gamma_m}(\gamma_1, \ldots, \gamma_m)$, so that if we are given the ROC curve represented by Q, denoted by f_Q, we have that

$$\begin{aligned} F_n(f_Q;\mathbf{d}) &= \sqrt{2} - \min_{\mathbf{q}\in Q}\left\{\frac{\langle \mathbf{q}-\mathbf{y}, -\mathbf{n}\rangle}{\|-\mathbf{n}\|}\right\} \\ &= \sqrt{2} - \min_{\mathbf{q}\in Q}\left\{\frac{\langle \mathbf{q}, -\mathbf{n}\rangle}{\|-\mathbf{n}\|}\right\} \end{aligned} \tag{29}$$

if Q is not empty, and

$$F_n(f_Q;\mathbf{d}) = 0,$$

otherwise. Fig. 8 shows the geometry of the ROC functional calculation where the number of classes is $n = 2$, and the given data is $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$. In this case, $m = 2^2 - 2 = 2$.

7. Fusors

We are now in a position to define a way in which we can compete fusion rules. Suppose we have a system such as that in Fig. 2. Each branch has a ROC curve that can be associated with the family of classification systems, and we now have a viable means of competing each branch. If we can only choose among the two classification systems, choose the one whose associated ROC functional is greater. Therefore, we can also compete these two classification systems with a new system that fuses the two data sets (or the feature sets for that matter) by fixing a third family of classification systems, which is based on the fusion rule, and finding the ROC functional of the event-to-label system corresponding to the fused data (features). If the fused system's ROC functional is greater than either of the original two, then the fusion rule is a fusor. Repeating this process on a finite number of fusion rules, we discover a finite collection of fusors with associated ROC functional values. The fusor that is the best choice is the fusor corresponding to the largest ROC functional value.

Do you want to change your a priori probabilities? Simply adjust $\boldsymbol{\gamma}$ in the ROC functional's data and recalculate the ROC functional for each corresponding ROC and choose the largest value. The corresponding fusor is then the best fusor to select under your criteria. Therefore, given a finite collection of fusion rules, we have for fixed ROC functional data a partial ordering of fusors.

Definition 4 (*Similar families of classification systems*). Two families of classification systems $\mathbb{A}$ and $\mathbb{B}$ are called similar if and only if they operate on the same σ-field and their output is the same label set, where each set element is defined the same way for each classification system.

Definition 5 (*Fusor*). Let $\mathbb{I} \subset \mathbb{N}$ be a finite subset of the natural numbers, with $\sup \mathbb{I} = n$. Given $\{\mathbb{A}_i\}_{i\in\mathbb{I}}$ a finite collection of similar families of classification systems, let $\mathscr{O}_0^{\pi(n)}$ be the category of fusion rules associated with the product of n data sets. Assume a functional, ρ, on the associated ROC curves of the classification systems, both original and fused. Then given that $f_{\mathbb{A}_i}$ is the ROC curve of the ith family of classification systems, and $f_{\mathfrak{R}}$ the ROC curve of the classification family, $\mathbb{A}_{\mathfrak{R}}$, associated with fusion rule $\mathfrak{R} \in \mathbf{Ob}(\mathscr{O}_0^{\pi(n)})$, we say that

$$\mathbb{A}_i \succeq \mathbb{A}_j \Longleftrightarrow \rho(f_{\mathbb{A}_i}) \geqslant \rho(f_{\mathbb{A}_j}), \tag{30}$$

so that if $\mathbb{A}_{\mathfrak{R}} \succeq \mathbb{A}_i$ for all $i \in \mathbb{I}$, then $\mathfrak{R}$ is called a fusor.

There is then a category of fusors, which is a subcategory of $\mathscr{O}_0^{\pi(n)}$, and whose arrows are such that given $\mathfrak{R}, \mathfrak{S}$ objects of this subcategory, then there exists an arrow, $\mathfrak{R} \xrightarrow{\rho} \mathfrak{S}$, if and only if, $\mathbb{A}_{\mathfrak{R}} \succeq \mathbb{A}_{\mathfrak{S}}$.

By way of example, suppose we start with the system

$$\begin{tikzcd}
& \mathrm{D}_1 \arrow[r, "p_1"] & \mathrm{F}_1 \arrow[dr, "a_\theta"] & \\
\mathrm{E} \arrow[ur, "s_1"] \arrow[dr, "s_2"'] & & & \mathrm{L} \\
& \mathrm{D}_2 \arrow[r, "p_2"] & \mathrm{F}_2 \arrow[ur, "b_\phi"'] &
\end{tikzcd}$$

with L an n-class label set. Let $A_\theta = a_\theta \circ p_1 \circ s_1$ and $B_\phi = b_\phi \circ p_2 \circ s_2$, and consider a functional F_n on the ROC curves $f_{\mathbb{A}}$ and $f_{\mathbb{B}}$ where $\mathbb{A}$ and $\mathbb{B}$ are defined as families of the respective classification systems shown (F_n being created under the assumptions and data of the researcher's choice). Then, given fusion rules $\mathfrak{S}$, such as that in Fig. 10, and $\mathfrak{T}$ and a second fusion branch

$$\begin{tikzcd}
& \mathrm{D}_1 & & & & \\
\mathrm{E} \arrow[ur, "\mathrm{s}_1"] \arrow[rr, "<s_1,s_2>"] \arrow[dr, "\mathrm{s}_2"'] & & \mathrm{D}_1 \times \mathrm{D}_2 \arrow[ul, "q_1"'] \arrow[dl, "q_2"] \arrow[r, "\mathfrak{T}"] & \mathrm{D}_3 \arrow[r, "p_3"] & \mathrm{F}_3 \arrow[r, "d_\kappa"] & \mathrm{L} \\
& \mathrm{D}_2 & & & &
\end{tikzcd}$$

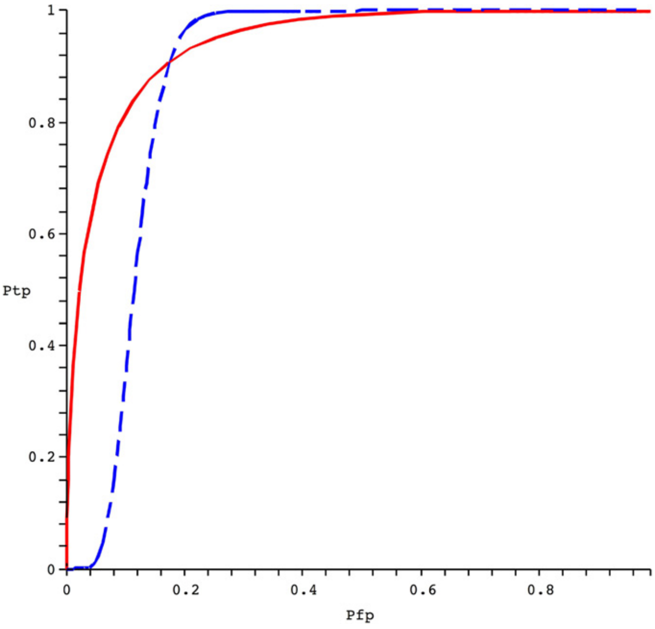

Fig. 9. ROC curves of two competing classification systems.

$$\begin{tikzcd}
& \mathrm{E} \arrow[dl, "s_1"'] \arrow[d, "<s_1{,}s_2>"] \arrow[dr, "s_2"] & \\
\mathrm{D}_1 \arrow[dd, "p_1"'] & \mathrm{D}_1 \times \mathrm{D}_2 \arrow[l, "q_1"'] \arrow[r, "q_2"] \arrow[d, "\mathfrak{S}"] & \mathrm{D}_2 \arrow[dd, "p_2"] \\
& \mathrm{D}_0 \arrow[d, "p"] & \\
\mathrm{F}_1 \arrow[dr, "a_\theta"'] & \mathrm{F}_0 \arrow[d, "c_\psi"] & \mathrm{F}_2 \arrow[dl, "b_\phi"] \\
& \mathrm{L} &
\end{tikzcd}$$

Fig. 10. Data fusion of two classification systems.

let $f_{\mathfrak{S}}$ and $f_{\mathfrak{T}}$ refer to the corresponding ROC curves to each of the fusion rule's systems (as a possible example of ROC curves of competing fusion rules see Fig. 9). Then we have that if $F_n(f_{\mathfrak{S}}) \geqslant F_n(f_{\mathbb{A}})$ and $F_n(f_{\mathfrak{S}}) \geqslant F_n(f_{\mathbb{B}})$ and similarly, if $F_n(f_{\mathfrak{T}}) \geqslant F(f_{\mathbb{A}})$ and $F_n(f_{\mathfrak{T}}) \geqslant F_n(f_{\mathbb{B}})$ then we say that $\mathfrak{S}, \mathfrak{T}$ are fusors. Furthermore, suppose $F_n(f_{\mathfrak{S}}) \geqslant F_n(f_{\mathfrak{T}})$. Then we have that $\mathfrak{S} \succeq \mathfrak{T}$. Thus, $\mathfrak{S}$ is the fusor a researcher would select under the given assumptions and data. Fig. 10 is a diagram showing all branches and products (along with the associated projectors q_1, q_2) in category theory notation.

8. Changing assumptions, robustness, and an illustrative example

While we have suggested a collection of functionals $(\{F_n : n \in \mathbb{N}\})$ to use as a way of competing classification systems, this collection is not the only choice a researcher has. There may be many others. Furthermore, one may desire to average functionals or transform them into new functionals. In many ways, the functional we have presented is general. We have shown its relationship to the Bayes optimal and Neyman–Pearson points on a ROC curve. It can also be shown to be related to Adam's and Hand's development of a loss comparison functional. In [35], the loss comparison of a classification system (LC) is denoted by

$$\mathrm{LC} = \int I(c_1) L(c_1)\,\mathrm{d}c_1, \tag{31}$$

where, although a slight abuse of notation, we have I as an indicator function of whether or not the classification system is still minimal under cost c_1, and c_1 is the cost of one type of error while c_0 is the cost of the other. $L(c_1)$ is a belief function which linearly weights how far c_1 is from the believed true cost of the error (or ratio $\frac{c_0}{c_1}$). This functional, LC, can be reformulated as follows:

Given competing classification systems $R = \{\mathbb{A}_i\}_{i=1}^{k}$ for $k \in \mathbb{N}$ fixed, fix $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. Let $\boldsymbol{\Gamma}$ be the set of all possible $\boldsymbol{\gamma}$. Define a set $H_{\boldsymbol{\gamma}}$ by

$$\mathrm{H}_{\boldsymbol{\gamma}} = \{\mathbb{A}_j \in R \mid F_2(f_{\mathbb{A}_j};\ \boldsymbol{\gamma}, \boldsymbol{\alpha}) \geqslant F_2(f_{\mathbb{A}_i};\ \boldsymbol{\gamma}, \boldsymbol{\alpha})\ \forall i \neq j,\\ i = 1, 2, \ldots, k\}.$$

Then, for $\mathbb{A}_i$ we have that

$$\mathrm{LC}(\mathbb{A}_i) = \int_{\boldsymbol{\Gamma}} I_{\mathrm{H}_{\boldsymbol{\gamma}}}(\mathbb{A}_i) W(\boldsymbol{\gamma})\,\mathrm{d}\boldsymbol{\gamma}, \tag{32}$$

where $W(\boldsymbol{\gamma})$ is the weight given to supposition $\boldsymbol{\gamma}$ (a belief function in this case). Thus LC scores the classification families, and induces an ordering on R.

One more suggested use of F_n would be to apply the belief function in a simpler way, and average F_n over the believed true $\boldsymbol{\gamma}$ and the believed extreme values of the set $\boldsymbol{\Gamma}$, so that

$$S_n(f_{\mathbb{A}}) = \frac{1}{2^n + 1}\left(\sum_{i=1}^{2^n} F_n(f_{\mathbb{A}};\ \boldsymbol{\gamma}_i, \boldsymbol{\alpha}) + F_n(f_{\mathbb{A}};\ \boldsymbol{\gamma}_0, \boldsymbol{\alpha})\right), \tag{33}$$

where $\boldsymbol{\gamma}_i$ are the believed extreme values of the set $\boldsymbol{\Gamma}$, and $\boldsymbol{\gamma}_0$ is the most believable (or probable under some instances) cost–prior product. In [35], the prior probabilities are assumed to be fixed, but they can be varied according to belief as well (although developing the belief functions will prove challenging).

As an example, consider the plot of two competing families of classification system in Fig. 11. Since we collected only finite data, the ROC 'curves' are actually a finite collection of ROC points. While our theory develops out of smooth manifolds, nevertheless, we can still calculate the functionals we require, since they operate on individual points on the ROC manifolds. The two curves in question cross more than once, and this is typical of many ROC curves, so deciding which family of classification system is best really boils down to which classification system within the family is best. Suppose our belief of the situation we are trying to classify is that the ratio of prior probabilities $\frac{Pr(E_1)}{Pr(E_2)}$ is $\frac{1}{2}$, with a range of ratios from $\frac{1}{3}$ to 1. Furthermore, our experts believe the most likely cost ratio is $\frac{C_{2|1}}{C_{1|2}} = 1$, with a range from $\frac{1}{2}$ to 2. Therefore, our prior–cost ratio is most likely $\frac{1}{2}$, with a range from $\frac{1}{6}$ to 2. We will refer to the two ROC curves as $f_{\mathbb{C}_1}$ and $f_{\mathbb{C}_2}$. Hence, the two classification systems shown in the figure yield scores of $F_2(f_{\mathbb{C}_1}) = F_2(f_{\mathbb{C}_2}) = 1.137$, indicating that the best classification systems in each family are equivalent with regard to the most believable prior–cost ratio. However, $S_2(f_{\mathbb{C}_1}) = 0.336 \geqslant 0.330 = S_2(f_{\mathbb{C}_2})$, indicating a preference of the best choice from $f_{\mathbb{C}_1}$, once belief regarding the range of the prior–cost ratio is taken into account. If our beliefs are actual probabilities from recorded data, the results are even stronger for selecting the best classification system represented in $f_{\mathbb{C}_1}$.

There are, of course, other suggestions for performance functionals regarding competing fusion rules. Consider

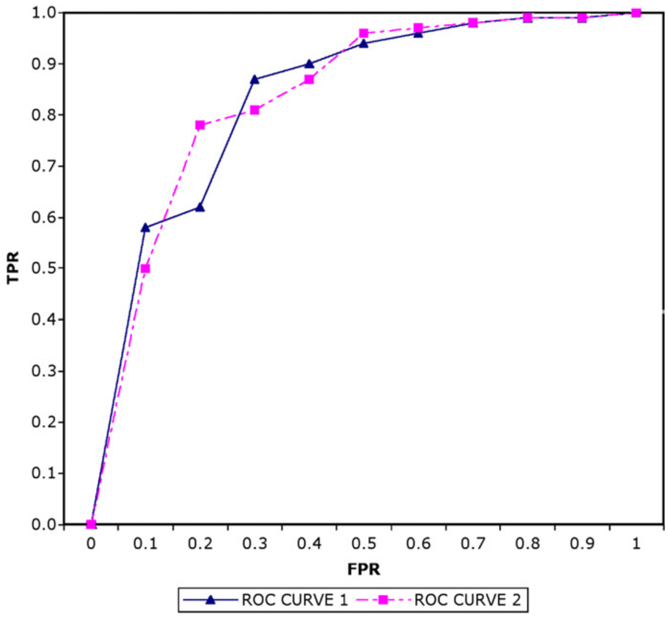

Fig. 11. ROC curves of two competing classifier systems.

fusion rules as algorithms, divorcing them from the entire classification system. Mahler [36] recommends using mathematical information MoEs (measures of effectiveness) with respect to comparing performance of fusion algorithms (fusion rules). In particular, he refers to level 1 fusion MoEs as being traditionally 'localized' in their competence. His preferred approach is to use an information 'metric', the Kullback–Leibler discrimination functional,

$$K(f_{\mathrm{G}},f) = \int_X f_{\mathrm{G}}(\mathbf{x})\log_2\left(\frac{f_{\mathrm{G}}(\mathbf{x})}{f(\mathbf{x})}\right)\mathrm{d}\mathbf{x},$$

where f_{G} is a probability distribution of perfect or near perfect ground truth, f is a probability distribution associated with the fused output of the algorithm and X is the set of all possible measurements of the observation. This works fine, if such distributions are at hand. One drawback is that it measures the expected value of uncertainty and therefore its relationship to costs and prior probabilities is obscure (as was the case with the Neyman–Pearson criteria). The previous functionals we have forwarded for consideration operate on families of classification systems (in particular, ROC manifolds), not just systems which enjoy well-developed and tested probability distribution functions.

9. Conclusion

A fusion researcher should have a viable method of competing fusion rules. This is required to correctly define fusion, and to demonstrate improvements over existing methods. Every fusion system can generate a corresponding ROC curve, and under a mild assumption of smoothness of the ROC curve, a Bayes optimal threshold (BOT) can be found for each family of classification systems. Given additional assumptions on the a priori probabilities of the classes of interest, along with given thresholds for the conditional probabilities of errors, a functional can be generated over the ROC manifolds. Every such functional will generate a partial ordering on families of classification systems, categories of fusion rules, and ultimately categories of fusors, which can then be used to select the best fusor from among a finite collection of fusors. We demonstrate one such functional, the ROC functional, which is scalable to ROC manifolds in dimensions higher than 2, as well as to families of classification systems which do not generate ROC curves at all. The ROC functional, when populated with the appropriate data choices, will yield a value corresponding the Bayes optimal threshold with respect to the classification system family being examined. The Neyman–Pearson threshold of a classification system is shown to correspond to the output of the ROC functional with one such data choice (so that it corresponds with the Bayes optimal threshold under one set of assumptions). Ultimately, a researcher could choose a cost–prior ratio which seems most reasonable, then perturbate it, calculate the mean ROC functional value, and choose the classification system with the greatest average ROC functional value. This value would be a relative comparison of how robust that classification system is to changes compared with other classification systems (e.g., it would answer the question of how much change is endured before another classification system is optimal?). The relationship of the ROC functional to other functionals, including the loss comparison functional, is demonstrated. Finally, there are other functionals to choose from, one which we mentioned, the Kullback–Leibler discrimination functional, may be unrelated to the ROC functional, yet may be suitable in particular circumstances where prior probabilities and costs are not fathomable, but probability distributions for fusion system algorithms and ground truth are available.

Appendix A

This appendix contains definitions for the understanding of category theory. We have drawn upon the work of various authors in category theory literature [24–27] to present the definitions.

Definition 6 (*Category*). A category $\mathscr{C}$ is denoted as a 4-tuple, $(\mathbf{Ob}(\mathscr{C}), \mathbf{Ar}(\mathscr{C}), \mathbf{Id}(\mathscr{C}), \circ)$, and consists of the following:

(A1) A collection of objects denoted $\mathbf{Ob}(\mathscr{C})$.
(A2) A collection of arrows denoted $\mathbf{Ar}(\mathscr{C})$.
(A3) Two mappings, called domain (dom) and codomain (cod), which assign to an arrow $f \in \mathbf{Ar}(\mathscr{C})$ a domain and codomain from the objects of $\mathbf{Ob}(\mathscr{C})$. Thus, for arrow f, given by $O_1 \xrightarrow{f} O_2$, $\mathrm{dom}(f) = O_1$ and $\mathrm{cod}(f) = O_2$.

(A4) A mapping assigning each object $O \in \mathbf{Ob}(\mathscr{C})$ an unique arrow $1_O \in \mathbf{Id}(\mathscr{C})$ called the identity arrow, such that

$$O \xrightarrow{1_O} O,$$

and such that for any existing element, x, of O, we have that

$$x \overset{1_O}{\mapsto} x.$$

(A5) A mapping, $\circ$, called composition, $\mathbf{Ar}(\mathscr{C}) \times \mathbf{Ar}(\mathscr{C}) \xrightarrow{\circ} \mathbf{Ar}(\mathscr{C})$. Thus, given $f, g \in \mathbf{Ar}(\mathscr{C})$ with cod(f) = dom(g) there exists an unique $h \in \mathbf{Ar}(\mathscr{C})$ such that $h = g \circ f$.

Axioms A3–A5 lead to the associative and identity rules:

- *Associative rule*. Given appropriately defined arrows f, g, and $h \in \mathbf{Ar}(\mathscr{C})$ we have that

  $(f \circ g) \circ h = f \circ (g \circ h)$.
- *Identity rule*. Given arrows $A \xrightarrow{f} B$ and $B \xrightarrow{g} A$, then there exists identity arrow 1_A such that $1_A \circ g = g$ and $f \circ 1_A = f$.

Definition 7 (*Subcategory*). A subcategory $\mathscr{B}$ of $\mathscr{A}$ is a category whose objects are some of the objects of $\mathscr{A}$ and whose arrows are some of the arrows of $\mathscr{A}$, such that for each arrow f in $\mathscr{B}$, dom(f) and cod(f) are in $\mathbf{Ob}(\mathscr{B})$, along with each composition of arrows, and an identity arrow for each element of $\mathbf{Ob}(\mathscr{B})$.

Definition 8 (*Discrete category*). A *discrete* category is a category whose only arrows are identity arrows.

A category of interest is the category *Set*, which has as its objects sets, that is $\mathbf{Ob}(Set)$ is a collection of sets, and its arrows, $\mathbf{Ar}(Set)$, the collection of all total functions defined on these sets, with its composition being the typical composition of functions. Clearly this construct has identity arrows and the associative rule applies, so it is indeed a category. The subcategories of interest to us are first, subcategories of particular types of data sets, denoted $\mathscr{D}$, whose objects are similar types of data and whose arrows consist of only the identity arrows, and second, subcategories of particular types of feature sets, denoted $\mathscr{F}$, whose objects are similar types of features, and whose arrows are only the identity arrows. The objects and arrows of these categories shall correspond to a particular sensor system, so they will represent all of the possible data (or features) that can be generated by the sensor or processor. For example, the data generated by a particular sensor system may be represented in an $N \times N$ real-valued matrix. In this case, $\mathscr{D} = (\mathbb{R}^{N \times N}, \mathbf{id}_{\mathscr{D}}, \mathbf{id}_{\mathscr{D}}, \circ)$ represents a discrete category, whose objects are $N \times N$ matrices over the field of real numbers, and whose arrows are only identity arrows, with composition, $\circ$, being the usual composition of functions.

A further categorical concept which will be useful is a *functor*.

Definition 9 (*Functor*). A *functor* $\mathfrak{F}$ between two categories $\mathscr{A}$ and $\mathscr{B}$ is a pair of maps $(\mathfrak{F}_{\mathbf{Ob}}, \mathfrak{F}_{\mathbf{Ar}})$

$$\mathbf{Ob}(\mathscr{A}) \xrightarrow{\mathfrak{F}_{\mathbf{Ob}}} \mathbf{Ob}(\mathscr{B}),$$

$$\mathbf{Ar}(\mathscr{A}) \xrightarrow{\mathfrak{F}_{\mathbf{Ar}}} \mathbf{Ar}(\mathscr{B}),$$

such that $\mathfrak{F}$ maps $\mathbf{Ob}(\mathscr{A})$ to $\mathbf{Ob}(\mathscr{B})$ and $\mathbf{Ar}(\mathscr{A})$ to $\mathbf{Ar}(\mathscr{B})$ while preserving the associative property of the composition map and preserving identity maps.

Thus, given categories $\mathscr{A}, \mathscr{B}$ and functor $\mathfrak{F} : \mathscr{A} \to \mathscr{B}$, if $A \in \mathbf{Ob}(\mathscr{A})$ and $f, g, h, 1_A \in \mathbf{Ar}(\mathscr{A})$ such that $f \circ g = h$ is defined, then there exists $\mathrm{B} \in \mathbf{Ob}(\mathscr{B})$ and $f', g', h', 1_{\mathrm{B}} \in \mathbf{Ar}(\mathscr{B})$ such that

(i) $\mathfrak{F}_{\mathbf{Ob}}(\mathscr{A}) = \mathscr{B}$.
(ii) $\mathfrak{F}_{\mathbf{Ar}}(f) = f'$, $\mathfrak{F}_{\mathbf{Ar}}(g) = g'$.
(iii) $h' = \mathfrak{F}_{\mathbf{Ar}}(h) = \mathfrak{F}_{\mathbf{Ar}}(f \circ g) = \mathfrak{F}_{\mathbf{Ar}}(f) \circ \mathfrak{F}_{\mathbf{Ar}}(g) = f' \circ g'$.
(iv) $\mathfrak{F}_{\mathbf{Ar}}(1_A) = 1_{\mathfrak{F}_{\mathbf{Ob}}(\mathscr{A})} = 1_{\mathrm{B}}$.

Definition 10 (*Natural transformation*). Given categories $\mathscr{A}$ and $\mathscr{B}$ and functors $\mathfrak{F}$ and $\mathfrak{G}$ with $\mathscr{A} \xrightarrow{\mathfrak{F}} \mathscr{B}$ and $\mathscr{A} \xrightarrow{\mathfrak{G}} \mathscr{B}$, then a *natural transformation* is a family of arrows $v = \{v_A \mid A \in \mathbf{Ob}(\mathscr{A})\}$ such that for each $f \in \mathbf{Ar}(\mathscr{A})$, $A \xrightarrow{f} A'$, $A' \in \mathbf{Ob}(\mathscr{A})$, the square

$$\begin{tikzcd}
\mathfrak{F}(\mathrm{A}) \arrow[r, "\nu_A"] \arrow[d, "\mathfrak{F}(f)"'] & \mathfrak{G}(\mathrm{A}) \arrow[d, "\mathfrak{G}(f)"] \\
\mathfrak{F}(\mathrm{A}') \arrow[r, "\nu_{\mathrm{A}'}"] & \mathfrak{G}(\mathrm{A}')
\end{tikzcd}$$

commutes. We then say the arrows v_A are the components of $v : \mathfrak{F} \to \mathfrak{G}$, and call v the natural transformation of $\mathfrak{F}$ to $\mathfrak{G}$.

Definition 11 (*Functor category $\mathscr{A}^{\mathscr{B}}$*). Given categories $\mathscr{A}$ and $\mathscr{B}$, the notation $\mathscr{A}^{\mathscr{B}}$ represents the category of all functors $\mathfrak{F}$ such that $\mathscr{B} \xrightarrow{\mathfrak{F}} \mathscr{A}$. This category has all such functors as objects and the natural transformations between them as arrows.

Definition 12 (*Product category*). Let $\{\mathscr{C}_i\}_{i=1}^{n}$ represent a finite collection of small categories (i.e., those which can be described using sets). Then

$$\prod_{i=1}^{n} \mathscr{C}_i = \mathscr{C}_1 \times \mathscr{C}_2 \times \cdots \times \mathscr{C}_n$$

is the corresponding product category.

Appendix B

Here we prove the extension of the calculus of variations approach to ROC manifolds. From Eq. (19) we generate the first variation of J_1. Let $\beta > 0$ be fixed, and let $\alpha \in [-\beta, \beta]$ be a family of real parameters. Let

$$\{\mathbf{R}(t,\alpha) = (X_1(t,\alpha),\dots,X_m(t,\alpha)) | \alpha \in [-\beta,\beta]\} \tag{34}$$

be a family of one-parameter trajectories which contains the optimal curve defined by the function $\mathbf{R}^*$. Furthermore, we assume $\mathbf{R}(t,0) = \mathbf{R}^*(t)$. Let $\mathbf{R}(t_f,\alpha) \in \mathfrak{M}$ for all $\alpha \in [-\beta,\beta]$. By the implicit function theorem there exists a function $T_f(\alpha)$ such that $\mathbf{R}(T_f(\alpha),\alpha) \in \mathfrak{M}$ for all α. Thus, $\mathbf{R}(t_f^*,0) = \mathbf{R}^*(t_f^*)$ so that $T_f(0) = t_f^*$. Since $\mathbf{R}^*$ minimizes J_1, then a necessary optimality condition is that the first variation of

$$J_1[\mathbf{R}(\cdot,\alpha)] = \int_0^{T_f(\alpha)} G\,\mathrm{d}t \tag{35}$$

is zero at $\alpha = 0$ (see [33]), that is,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} J[\mathbf{R}(\cdot,\alpha)]|_{\alpha=0} = 0. \tag{36}$$

We use the operator notation

$$\delta = \frac{\mathrm{d}}{\mathrm{d}\alpha}|_{\alpha=0}$$

for brevity. Applying Leibniz's rule to $J_1[\mathbf{R}(\cdot,\alpha)]$ in Eq. (35) we get the derivative to be

$$\delta J[\mathbf{R}^*] = G^*|_{t=t_f^*}\delta T_f + \int_0^{t_f^*} (\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} + \nabla_{\mathbf{y}} G^* \cdot \delta\dot{\mathbf{R}})\,\mathrm{d}t, \tag{37}$$

where G^* is a suppressed notation for $G(t,\mathbf{R}^*(t),\dot{\mathbf{R}}^*(t))$. Now integrating by parts yields

$$\begin{aligned}\delta J[\mathbf{R}^*] = {} & G^*|_{t=t_f^*}\delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_0^{t_f^*} \\ & + \int_0^{t_f^*} (\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} - \frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R})\,\mathrm{d}t.\end{aligned} \tag{38}$$

At $\alpha = 0$ the necessary optimality condition implies

$$\begin{aligned}\delta J[\mathbf{R}^*] = {} & G^*|_{t=t_f^*}\delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_{t=t_f^*} \\ & + \int_0^{t_f^*} \left(\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} - \frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}\right)\mathrm{d}t = 0.\end{aligned} \tag{39}$$

Since this must be true over all admissible variations $(\delta\mathbf{R}, \delta T_f)$, we have the Euler equations

$$\nabla_{\mathbf{x}} G^* - \frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathbf{y}} G^* = \mathbf{0} \tag{40}$$

for all $t \in [0, t_f^*]$ and a transversality condition

$$G^*|_{t=t_f^*}\delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_{t=t_f^*} = 0. \tag{41}$$

Since G is independent of $\mathbf{x}$ then $\nabla_{\mathbf{x}} G^* = \mathbf{0}$. Using this in solving the Euler Eq. (40), yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathbf{y}} G^* = \mathbf{0}, \tag{42}$$

hence, for $i = 1,\dots, m$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{sgn}[\dot{X}_i^*(t)] = 0 \quad \text{for all } t \in (0, t_f). \tag{43}$$

Thus, integrating for each $i = 1,\dots,m$, we have

$$\mathbf{sgn}[\dot{X}_i^*(t)] = \mathrm{K}_i. \tag{44}$$

Hence, $\mathbf{sgn}(\dot{X}_i^*(t)) = K_i$ for some constant $K_i \in \mathbb{R}$. For $i = 1,\dots,m$, we have that $\Delta X_i^*(t) \geqslant 0$ and $\Delta t > 0$ for all t, so that $K_i = 0$ or 1. We make the assumption that $X_i^*(t_f) \neq 0$ for some i, since to say otherwise would indicate we have the perfect classification system. Thus, we have that $K_i = 1$ for at least one i. Hence, we have that $\{1,2,\dots, m\} = N_1 \cup N_0$ is a partition such that

$$\mathbf{sgn}(\dot{X}_i^*(t)) = 1$$

for all $i \in N_1$, and

$$\mathbf{sgn}(\dot{X}_i^*(t)) = 0$$

for all $i \in N_0$. Since, $\mathbf{R}(T_f(\alpha),\alpha)$ terminates on $\mathfrak{M}$ for all α, then $\Psi(\mathbf{R}(T_f(\alpha),\alpha)) = 0$ for all α. Let $\mathbf{R}^*(t_f^*) = (x_1^*,\dots,x_m^*) \in \mathfrak{M}$. Hence,

$$X_m(T_f(\alpha),\alpha) = f(X_1(T_f(\alpha),\alpha),\dots,X_{m-1}(T_f(\alpha),\alpha)) \tag{45}$$

for all α. Taking the variation of each side of Eq. (45), we have

$$\dot{X}_m^*(t_f^*)\delta T_f + \delta X_m(t_f^*) = \sum_{i=1}^{m-1} \frac{\partial f(x_1^*,\dots,x_{m-1}^*)}{\partial x_i}[\delta T_f + \delta X_i(t_f^*)]. \tag{46}$$

Expanding Eq. (46) and defining $H_i(t) = \delta X_i(t)$, we have

$$\begin{aligned}\dot{X}_m(t_f^*)\delta T_f + H_m(t_f^*) = {} & \sum_{i=1}^{m-1} \frac{\partial f(x_1^*,\dots,x_{m-1}^*)}{\partial x_i}\dot{X}_i(t_f^*)\delta T_f \\ & + \sum_{i=1}^{m-1} \frac{\partial f(x_1^*,\dots,x_{m-1}^*)}{\partial x_i}H_i(t_f^*).\end{aligned} \tag{47}$$

Rearranging terms, rewriting in vector notation, and letting $f^* = f(x_1^*,\dots,x_{m-1}^*)$ we have

$$\begin{aligned}& \left(\frac{\partial f^*}{\partial x_1},\dots,\frac{\partial f^*}{\partial x_{m-1}},-1\right)\cdot\left(H_1(t_f^*),\dots,H_{m-1}(t_f^*),H_m(t_f^*)\right) \\ & + \left(\frac{\partial f^*}{\partial x_1},\dots,\frac{\partial f^*}{\partial x_{m-1}},-1\right)\cdot\dot{\mathbf{R}}^*(t_f^*)\delta T_f = 0\end{aligned} \tag{48}$$

which can be rewritten as

$$\nabla\Psi^* \cdot \mathbf{H}(t_f^*) + \nabla\Psi^* \cdot \dot{\mathbf{R}}^*(t_f^*)\delta T_f = 0. \tag{49}$$

From Eq. (41) we write

$$\nabla_{\mathbf{y}} G^*|_{t_f^*} \cdot \mathbf{H}(t_f^*) + G^*|_{t_f^*}\delta T_f = 0. \tag{50}$$

Since both Eqs. (49) and (50) must be true over all variations and all possible one-parameter families, we have

$$\nabla\Psi^*|_{t_f^*} = \lambda\nabla_{\mathbf{y}} G^*|_{t_f^*} \tag{51}$$

for some $\lambda \in \mathbb{R}$. Hence, for $i = 1,\dots,m$ we have

$$\left.\frac{\partial\Psi}{\partial x_i}\right|_{t=t_f^*} = \lambda a_i\mathbf{sgn}(\dot{X}_i^*(t_f^*)). \tag{52}$$

In the case of $i = m$ we have that

$$-1 = \frac{\partial \Psi^*}{\partial x_3}\Big|_{t=t_f^*} = \lambda a_3. \tag{53}$$

Thus, we have that $\lambda = \frac{-1}{a_m}$. Hence, for $i = 1,\ldots,m$ we have that

$$\frac{\partial \Psi^*}{\partial x_i}\bigg|_{t=t_f^*} = \frac{-a_i}{a_m}. \tag{54}$$

This is a global minimum since we are optimizing a convex functional [34]. This solution agrees with the limited approach, based on observation, made by Haspert [37].

References

[1] S. Thorsen, M. Oxley, Comparing fusors within a category of fusors, in: Proceedings of the Seventh International Conference of Information Fusion, ISIF, Stockholm, Sweden, 2004, pp. 435–441.

[2] L. Wald, Some terms of reference in data fusion, IEEE Transactions on Geoscience and Remote Sensing 37 (3) (1999) 1190–1193.

[3] D.M. Green, J.A. Swets, Signal Detection Theory and Psychophysics, John Wiley and Sons, New York, 1966.

[4] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[5] L.L. Scharf, Statistical Signal Processing, Addison-Wesley, MA, 1991.

[6] F.J. Provost, T. Fawcett, Robust classification systems for imprecise environments, in: AAAI/IAAI, 1998, pp. 706–713. Available from: <http://citeseer.ist.psu.edu/provost98robust.html>.

[7] L.C. Ludeman, Random Processes Filtering, Estimation, and Detection, John Wiley and Sons, New Jersey, 2003.

[8] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley and Sons, New York, 2001.

[9] H.L. Van Trees, Detection, Estimation, and Modulation Theory, John Wiley and Sons, New York, 2001.

[10] C.E. Metz, Basic principles of ROC analysis, Seminars in Nuclear Medicine 8 (4) (1978) 283–298.

[11] F.J. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (3) (2001) 203–231. Available from: <http://citeseer.ist.psu.edu/provost01robust.html>.

[12] J.A. Swets, Measuring the accuracy of diagnostic systems, Science 240 (1988) 1285–1293.

[13] C. Ferri, J. Hernandez-Orallo, M.A. Salido, Volume under the ROC surface for multi-class problems, Exact computation and evaluation of approximations, 2003. Available from: <http://citeseer.ist.psu.edu/ferri03volume.html>.

[14] D. Mossman, Three-way ROCs, Medical Decision Making 19 (1) (1999) 78–89.

[15] S.N. Thorsen, M.E. Oxley, Describing data fusion using category theory, in: Proceedings of the Sixth International Conference on Information Fusion, ISIF, Cairns Australia, 2003, pp. 1202–1206.

[16] S.N. Thorsen, M.E. Oxley, Multisensor fusion description using category theory, in: IEEE Aerospace Conference. Proceedings, 2004, pp. 2016–2021.

[17] M.E. Oxley, S.N. Thorsen, Fusion and integration: What's the difference? in: P. Svensson, J. Schubert (Eds.), Proceedings of the Seventh International Conference on Information Fusion, International Society of Information Fusion, CA, 2004, pp. 429–434. Available from: <http://www.fusion2004.foi.se/papers/IF04-0429.pdf>.

[18] S.A. DeLoach, M.M. Kokar, Category theory approach to fusion of wavelet-based features, in: Proceedings of the Second International Conference on Fusion (Fusion 1999), vol. I, 1999, pp. 117–124.

[19] M.M. Kokar, J.A. Tomasik, J. Weyman, Data vs. decision fusion in the category theory framework, in: Proceedings of the Fourth International Conference on Fusion (Fusion 2001), vol. I, 2001.

[20] M. Healy, T. Caudell, Y. Xiao, From categorical semantics to neural network design, in: IEEE Transactions on Neural Networks, Proceedings of International Joint Conference on Neural Networks (IJCNN '03), 2003, pp. 1981–1986.

[21] M. Healy, Category theory applied to neural modeling and graphical representations, in: IEEE Transactions on Neural Networks, INNS-ENNS Proceedings of International Joint Conference on Neural Networks (IJCNN '00), vol. 3, 2000, pp. 35–40. Available from: <http://citeseer.ist.psu.edu/healy00category.html>.

[22] M. Healy, T. Caudell, A categorical semantic analysis of art architectures, in: IEEE Transactions on Neural Networks, Proceedings of International Joint Conference on Neural Networks (IJCNN '01), vol. 1, 2001, pp. 38–43.

[23] M. Healy, Colimits in memory: Category theory and neural systems, in: IEEE Transactions on Neural Networks, Proceedings of International Joint Conference on Neural Networks (IJCNN '01), vol. 1, 1999, pp. 492–496.

[24] S. MacLane, Categories for the Working Mathematician, second ed., Springer, New York, 1978.

[25] C. McClarty, Elementary Categories, Elementary Toposes, Oxford University Press, New York, 1992.

[26] J. Adámek, H. Herrlich, G. Strecker, Abstract and Concrete Categories, John Wiley and Sons, New York, 1990.

[27] F.W. Lawvere, S.H. Schanuel, Conceptual Mathematics, A First Introduction to Categories, Cambridge University Press, Cambridge UK, 1991.

[28] P. Billingsley, Probability and Measure, third ed., John Wiley and Sons, New York, 1995.

[29] L. Xu, A. Krzyák, Y.C. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, in: IEEE Transactions on Systems, Man, and Cybernetics, vol. XXII, 1992, pp. 418–435.

[30] D.L. Hall, J. Llinas, Handbook of Multisensor Data Fusion, CRC Press, Florida, 2001.

[31] C. Schubert, M.E. Oxley, K.W. Bauer, A comparison of ROC curves for label-fused within and across classifier systems, in: Proceedings of the Ninth International Conference on Information Fusion, International Society of Information Fusion, 2005.

[32] S.G. Alsing, The evaluation of competing classifiers, Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH, March 2000.

[33] I.M. Gelfand, S.V. Fomin, Calculus of Variations, Dover, New York, 2000.

[34] D.G. Luenberger, Optimization by Vector Space MethodsWiley Professional Paperback Series, John Wiley and Sons, New York, 1969.

[35] N. Adams, D. Hand, Improving the practice of classifier performance assessment, Neural Computation XII (2000) 305–311.

[36] R.P. Mahler, An Introduction to Multisource–Multitarget Statistics and its Applications, Lockheed Martin, MN, 2000.

[37] J.K. Haspert, Optimum ID sensor fusion for multiple target types, Technical Report IDA Document D-2451, Institute for Defense Analyses, March 2000.