12-2005

The Application of Category Theory and Analysis of Receiver Operating Characteristics to Information Fusion

Steven N. Thorsen

THE APPLICATION OF CATEGORY THEORY AND ANALYSIS OF RECEIVER OPERATING CHARACTERISTICS TO INFORMATION FUSION

DISSERTATION

Steven N. Thorsen, Major, USAF

AFIT/DS/ENC/05-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

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

AFIT/DS/ENC/05-02

THE APPLICATION OF CATEGORY THEORY AND ANALYSIS OF RECEIVER OPERATING CHARACTERISTICS TO INFORMATION FUSION

DISSERTATION

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Steven N. Thorsen, B.A., M.A.

Major, USAF

December 2005

AFIT/DS/ENC/05-02

THE APPLICATION OF CATEGORY THEORY AND ANALYSIS OF RECEIVER OPERATING CHARACTERISTICS TO INFORMATION FUSION

Steven N. Thorsen, B.A., M.A.

Major, USAF

Approved:

Dr. Mark E. Oxley Dissertation Advisor	date
Dr. Lawrence K. Chilton Committee Member	date
Dr. Kenneth W. Bauer, Jr. Committee Member	date
Dr. Mark N. Goltz Dean's Representative	date

Accepted:

Robert A. Calico, Jr. Dean, Graduate School of Engineering and Management	Date

AFIT/DS/ENC/05-02

Abstract

Multisensor data fusion is presented in a rigorous mathematical format, with definitions consistent with the desires of the data fusion community. In particular, a model of event-state fusion is developed and described, concluding that there are two types of models on which to base fusion (in the literature referred to as within fusion and across fusion). Six different types of fusion are shown to exist, with respect to the model, using category theory. Definitions of fusion rules and fusors are introduced, along with the functor categories, of which they are objects. Defining fusors and competing fusion rules involves the use of an objective function of the researchers choice. One such objective function, a functional on families of classification systems, and in particular, receiver operating characteristics (ROCs), is introduced. Its use as an objective function is demonstrated in that the argument which minimizes it (a particular ROC), corresponds to the Bayes Optimal threshold, given certain assumptions, within a family of classification systems. This is proven using a calculus of variations approach using ROC curves as a constraint. This constraint is extended to ROC manifolds, in particular, topological subspaces of $\mathbb{R}^n$. These optimal points can be found analytically if the closed form of the ROC manifold is known, or calculated from the functional (as the minimizing argument) when a finite number of points are available for comparison in a family of classification systems. Under different data assumptions, the minimizing argument of the ROC functional is shown to be the point of a ROC manifold corresponding to the Neyman-Pearson criteria. A second functional, the ℓ_2 norm, is shown to determine the min-max threshold. Finally, more robust functionals can be developed from the offered functionals.

Acknowledgements

First and foremost, I owe a large debt of gratitude to my wife of nearly 20 years. Without her tireless support of me, from taking tremendous care of the children and I, to seeing me through my chemotherapy and radiation treatments, she has demonstrated the character and love of Jesus Christ. Without question, I have the best kids a dad could ask for, and I'm grateful to them as well for sacrificing time with me so that I could accomplish this dissertation.

Second, I would like to thank Dr. Mark Oxley for seeing me through this project. It is not an easy thing to guide and motivate a grad student through the end of his education plan, but my dear advisor was more than strength itself for me in my desperate hours. Through it all, he held the vision of my graduation more firmly than I myself. I appreciate his vast fount of knowledge and creativity in helping a student, "poor in spirit", to come through in the end.

Third, a big thank you to my committee members, Drs. Chilton and Bauer for their patient review and guidance in making this research more valuable than it could be on its own.

Finally, I would also like to thank the Air Force Research Laboratory (AFRL) and the Defense Advanced Research Projects Agency (DARPA) for their support of this project.

ANNUIT CŒPTIS

Steven N. Thorsen

Table of Contents

Page

List of Figures

List of Tables

List of Symbols

Symbol		Page
Neur	Category of Neural Networks	9
Conc	Category of Concepts	9
iff	if and only if	15
dom	Domain	20
cod	Codomain	20
GRP	Category of Groups	22
$\mathbf{CRng}$	Category of Commutative Rings	23
Ab	Category of Abelian Groups	24
Ban	Category of Banach Spaces	24
$\mathbf{Vect}_{\mathbb{K}}$	Category of Finite Dimensional Vector Spaces	24
$\mathbf{SET}$	Category of Sets	24
$\operatorname*{p\,lim}_{n\to\infty}$	convergence in probability	33
wlog	without loss of generality	36
$a.e.$	almost everywhere	39
$a.s.$	almost sure	39
L^{E}	Functor Category of Classification Systems	44
$\boldsymbol{FR}_{\mathbb{O}_n}(\mathcal{O}_0)$	Functor Category of Fusion Rules	46
$\mathbf{CAT}$	The small category whose objects are sets	52
$\mathbf{OC}_{\mathrm{L}^{\mathrm{E}}}$	Category of Operating Characteristic Families	53
$\mathbf{sgn}(Z)$	sign function	60
$\langle \cdot, \cdot \rangle$	Scalar Product	73
$\mathbf{FUS}_{\mathrm{L}^{\mathrm{E}}}$	Category of Fusion Processes	79

List of Abbreviations

Abbreviation		Page
AFRL	Air Force Research Laboratory	v
DARPA	Defense Advanced Research Projects Agency	v
JDL	Joint Directors of Laboratories Data Fusion Subpanel	1
ROC	Receiver Operating Characteristic	3
ART	Adaptive Resonance Theory	9
AUC	area under the ROC curve	10
ROCCH	ROC Convex Hull	10

THE APPLICATION OF CATEGORY THEORY AND ANALYSIS OF RECEIVER OPERATING CHARACTERISTICS TO INFORMATION FUSION

I. Introduction

1.1 Problem Statement

Data fusion as a science has been rapidly developing since the 1980's. Fusion literature encompasses many aspects of data fusion from mathematical techniques [8, 15, 55] to technologies, how to register and align data, as well as resource management of the assets to be used. The Joint Directors of Laboratories Data Fusion Subpanel (JDL) has put out guidance in the form of a functional model (which we will review later). What is missing? A clear definition of what fusion is in a mathematical sense. While many mathematical techniques have been developed and compiled , one look at the spread and variety of sub-processes such as *sensor* fusion, *data* fusion, and *classifier* fusion (all of which can be identified by other names) demonstrates the lack of unity within the science. As late as 2001, the Handbook of Multisensor Data Fusion [15], includes a recommendation that data fusion be defined as

> *the process of combining data or information to estimate or predict entity states.*

This is an improvement over the Handbook's previous version, but what are the mathematical formulations for fusion? How shall we define the technology? For example, does it matter how data or information are combined? What is meant by data or information? Does the estimation or prediction of entity states need to conform to some standards of accuracy or reliability to be called fusion? Are there clear delineations of different types of fusion or is all fusion the same? How can we mathematically define and compare dif-

ferences? This dissertation will explore these questions, but will focus on the following problem:

An entity (say some corporation) wants to combine some sets of constructed information (or data) into a new set of symbols which clarifies the object from which the information (or data) originated. The technology developed includes a finite number of algorithms to compute the combinations. The entity has two problems it would like to address:

1. In documenting its efforts, writing patent applications, conversing with the fusion community, and contracting for technologies from other entities, it needs a common framework (preferrably quantitative in nature) to accomplish these tasks.
2. How does the entity compete the algorithms to ensure it is getting the most for its investment?

In particular, we envision developing a rigorous mathematical lexicon for the US Air Force to use in creating documents contracting for fusion technologies. Although the definitions will be structured from abstract mathematical ideas, the vocabulary will be rather intuitive in nature. Furthermore, we present one concept of how to compete fusion technologies. Main mathematical results are identified as theorems, lemmas, and corollaries.

Since information fusion is a rapidly advancing science, researchers are daily adding to the known repertoire of fusion techniques (that is, fusion rules); however, a methodology to define what fusion is and when it has actually occurred has not been widely discussed or identified in the literature. An organization that is building a fusion system to detect or identify objects using existing assets or those yet to be constructed 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 with a myriad of schemes. To add to this, there exists the fusion of all these technologies which create even more classification systems. Since receiver operating characteris-

tic (ROC) curves can be developed for each system under test conditions, we propose a functional defined on ROC curves as a method of quantifying the performance of a classification system. 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), which do rely upon the qualitative differences. 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 scenario to those proposing different classification systems *a priori* the test. Therefore, even those systems relying upon class conditional density knowledge a priori can at best estimate the test scenario (and by extension the operational conditions the system will find itself 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 based on the test scenario. The desired scenario of the test organization may be examined under a range of assumptions (without actually retesting), and functional averages can be observed as well, so performance can be compared over a restricted range of assumed cost functions and prior probabilities. The result is a sound mathematical approach to comparing classification systems. The functional is scalable to any finite number of classes (the classical detection problem being two classes), with the development of ROC manifolds of dimension $n \geq 3$. The functional will operate on discrete ROC points in the n-dimensional ROC space as well. Ultimately, we will be able under certain assumptions and constraints, to compete families of classification systems, fusion rules, fusors, and fused families of classification 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 [10, 14, 17, 30, 34, 36, 41, 45]. The method of discovery of these prop-

erties 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 classification systems, the fact that the slope of a ROC curve at a point actually is the likelihood ratio which produces this point, seems to have been discovered in this manner [14]. Using cost functions in relation to ROC curves to analyze best performance has recently (2001) been recognized by Provost and Foster [42], based on work previously published by [17, 36, 48]. 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, we have ROC manifolds that are continuous and differentiable, but we apply variational calculus to a weighted distance functional on a specific family of manifolds, which has the effect of identifying the point on the ROC manifold which minimizes Bayes Cost. Under any particular assumption on prior probabilities and costs associated with errors in classification, 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 systems, but rather it is Bayes optimal with respect to the elements of the family of classification systems producing the ROC manifold. We believe this functional (which is really a family of functionals for each finite number of classes considered) eliminates the need to discuss classification system performance in terms of area under the ROC curve (AUC), which is so prevalently used in the medical community, or volume under the ROC surface (VUS) [12, 37], since these performance 'metrics' do nothing to describe a classification system's value under a specific cost-prior assumption. Any classification system used will be set at a particular threshold (at any one time), and so its performance will be measured by 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 knowledge revealed by the calculus of variations approach. Additionally, the novelty

of this proposal also relies on the fact that no class conditional densities are assumed (by the tester), and 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 maybe to minimize a Bayes cost functional, 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 category of fusion rules, fusors, and ultimately the set of families of classification systems. This partial ordering is a category in itself, but is also used to provide a mathematical definition of a fusor, which is derived from the fusion rules, and embodies mathematically the qualitativeness desired by researchers according to the application of the problem to which they are engaged. In other words, we have put to paper the definition of what makes a fusion rule based classification system "better" than the classification systems from which it was derived. An illustrative example and further applications of the functionals, with consideration of robustness, are put forth in the final section of this dissertation.

1.2 Literature Review

Our literature review consisted of three main areas: information or data fusion, category theory *with* data fusion, and ROC analysis. We were interested in how other researchers discussed and communicated their ideas of fusion, and in particular, whether mathematical descriptions of the overall fusion process are used (and not just a particular technique). Our decision to use category theory as the mathematical language prompted a search for the application of category theory to the science of information fusion. Finally, how do researchers ensure their results have the quality required to actually call what they are doing fusion? We decided to explore the world of ROC analysis since every classification system can generate at least one ROC, and this seemed a reasonable place to look for the type of functions (or functionals) which would be useful to provide a definition for quality of a particular fusion rule.

1.2.1 Data Fusion. As late as 1999, Dr. Wald in [54] described the challenges in the science of data fusion, posed by not having a language with common terms. These challenges are readily seen in the early results of the JDL definitions, where the language of what fusion was consisted of combining, integrating, estimating, predicting, scheduling, optimizing, and more! The earlier Handbook of Data Fusion [15] had this definition of fusion (from the JDL Data Fusion Lexicon):

> A process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats, and their significance. The process is characterized by continuous refinements of its estimates and assessments, and the evaluation of the need for additional sources, or modification of the process itself, to achieve improved results.

This definition was pruned in [15] to be:

> Data fusion is the process of combining data or information to estimate or predict entity states.

Dr. Wald correctly identified some of the problems and expressed the desire to have a more suitable definition with the following principles:

- The definition should not be restricted to data output from sensors alone;
- It should not be based on the semantic levels of the information;
- It should not be restricted to methods and techniques;
- It should not be restricted to particular system architectures.

He then went on to write a definition, "data fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application."

Here we have two definitions, which are very close, but still at odds. The first does not require a formal framework, which the second does, and also throws in the purpose for the fusion, but no necessity of the quality of the information (at least not explicitly stated).

The second requires tools for the alliance of data from different sources (without defining what is different about them), states the purpose much better, and allows the quality of the improvements to rest with the body of research. This 'greater quality' is still not defined.

While these two works focus on definitions, the vast majority of other data fusion papers and books focus on the use of particular mathematical techniques. Each author shows the cases in which his technique is optimal (see for example [6, 23, 24]), and compares against a single parameter, such as probability of detection, or uses a ROC curve. In those cases where ROC curves can be shown to be dominant in the compared technique, the fusion rule is proven, but in cases where ROC curves cross this comparison is not possible without further elaboration and theory development. Performance evaluation is also a concern in [33], where the use of *information measures of effectiveness (MOEs)* are discussed. The focus here is on multisource-multitarget statistics, referred to as FISST (finite set statistics). The use of information theory measurements are used, such as the Kullback-Liebler cross entropy or discrimination. The use of these measures seems to only pertain to the signal level of the classification system. In particular, the Kullback-Liebler discrimination uses the probability distribution associated with ground truth and the random variable representing a sensor system. Since we will show the classification system is a random variable made up sensors, processors, and classifiers, the information theory approach is useful for the development of better sensors (and possibly processors). The drawbacks are that it does not respect Bayesian principles, in that it does not allow for testing of different prior probabilities and costs. In the cases which they seem to be a good measurement, the label sets are simply the two-class case of classification systems. Extending the distribution functions to a joint distribution function of k classes will prove to be very cumbersome to the researcher. We admit the connections between the information theory measurements and the classification system measurements of probabilities of error need to be formally explored, but this is beyond the scope of this dissertation.

1.2.2 Category Theory and Fusion. Literature in the area of Category Theory and Fusion is very limited. There are only a few authors that have attempted to use

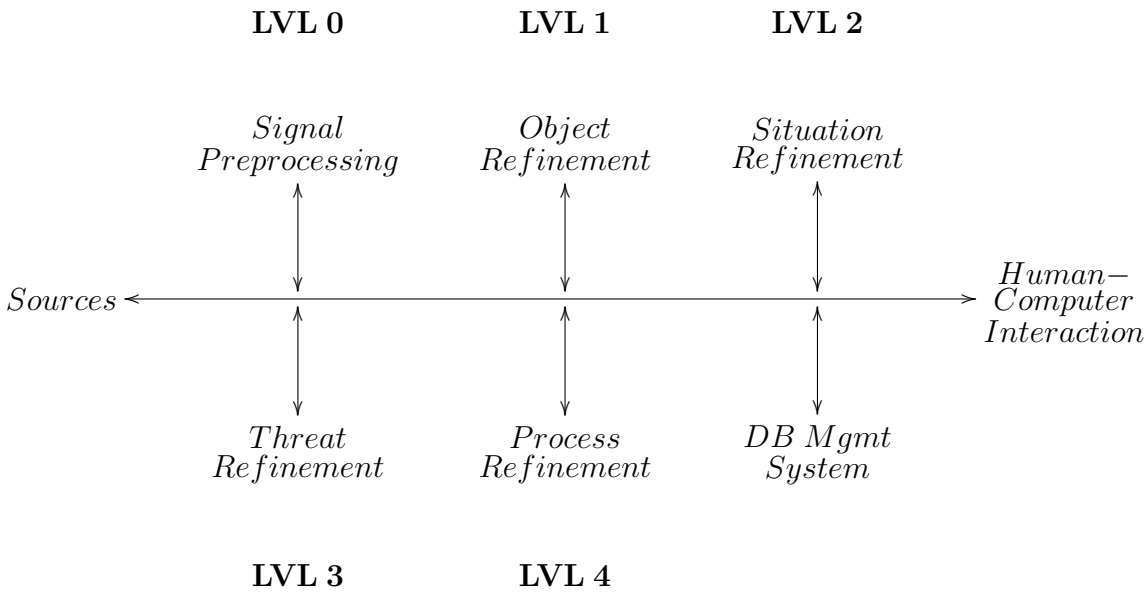


Figure 1.1: The Joint Directors of Laboratories Functional Model of Data Fusion [15].

the mathematics in this sense [7, 25–27]. Each of these works relies upon the use of formal systems, or systems constructed from first predicate logic. These are systems a computer can understand through the writing of software. These constructions require that theories can be written, using symbolic logic, which completely describe the target classes of interest. Also required are models of the environment, which incorporate these theories. The categories are actually categories whose objects are specifications (from the computer language Slang® by Specware). Each specification consists of a collection of pairs of theories and signatures (languages). The arrows of the category are mappings (not in the sense of functions) changing one specification into another, so that identities are clearly defined. In these papers fusion is an "operator" which returns the colimit of the objects. This turns out to be the disjoint union of the theories and languages. The operation is subject to a constraint that maintains the consistency of the category. We will show, as an example, after our development of a fusion definition, how this construction fits into our view.

Another set of interesting papers of category theory application has been written by Dr. M. J. Healy [19–22] of Boeing. Healy puts forth the notion of a category **Neur** of neural networks. The objects of this category are the nodes of the neural net, and the arrows are the primed paths (the identity arrows being clear). Composition is the usual composition of arrows, so that if one path is primed and a second is primed from the range of the first, then there is a primed path from the domain of the first to the range of the second. He then asserts that memories are the colimits of primed paths in [19]. Colimits, functors, and natural transformations in a different category show the shortcomings of adaptive resonance theory (ART) networks [22]. Colimits again play a pivotal role in [20], which expands the previous work, by enabling a new category **Conc** of concepts, which is like Kokar's work in that it relies upon theories and predicate logic, and defining functors between **Conc** and **Neur**.

With all these works pointing towards creating categories which then depend on colimits as their fusion, is it then true that colimits are the definition of fusion we're looking for? We don't think so based on the following:

- while colimits are optimal in an algebraic sense, there are still classification parameters to be considered. For example, just because a colimit can be calculated, doesn't equate to the new classification being correct! There is the possibility that error in the original data has skewed the colimit to producing something which performs worse than one of the systems we started with, thus ignoring the desired qualitative aspect to fusion we're looking for.
- the colimits developed were specific to formal methods used in designing computer systems. They are not applicable to other systems designed in different ways. According to Dr. Wald, then, this requires a particular system architecture; therefore, we cannot define fusion based on these cases alone.

1.2.3 ROC Analysis. Receiver Operating Characteristics (ROCs) play a significant role in determining the performance of classification systems. They have been used

extensively in the medical and psychological communities with regard to imaging and diagnoses [14, 48]. The definitions of ROCs and ROC curves and manifolds are presented in Section 2.2. There are in general two ways to look at the analysis of ROCs. The first is to consider an entire family of classification systems which create a ROC curve or manifold. The second is to consider that each classification system creates a ROC, that there are particular families of these classifications which can be constructed with meaning, and that there is a Bayesian interpretation of their significance with respect to the problem of classification.

We explore the first viewpoint by noting that in two-class problems, the ROC curve is an entity in two space, the basis of which is two error axes. Thus, the ROC curve can be used to calculate certain statistical properties of the original family of classification systems. One such measurement is the area under the ROC curve (AUC). The area under the ROC curve has been described in a couple of different ways:

- Given two instances of data, one from each of the two populations, the AUC is the probability of the system correctly identifying the class of each instance of data [14].
- The more general view is that AUC is a measure of how well a family of classification systems separate the conditional class distribution functions of the two classes.

We will point out that the emphasis on the family of classification systems is ours. Generally, researchers have regarded these curves as being derived from a single classification system. This is an incorrect view of the problem of ROC analysis. It is recognized that to generate a curve or a manifold, a parameter (which is possibly multi-dimensional) must be varied. This changes the classification system, so that it does not have the same performance as the original one.

The AUC is, in general, the measurement sought after by many researchers, and researchers have gone out of their way to estimate it by many means. These means include calculating the ROC convex hull (ROCCH) [41, 42], the Mann-Whitney test, and the Gini coefficient [16]. These efforts are based on the belief that the AUC divorces the problem

from finding out the costs involved in making the errors the ROC measures and in knowing the relative ratios of the classes in question. In other words, by using the AUC and the associated estimates, one does not need to concern oneself with prior probabilities and the costs of making certain errors in classification. We show in Section 4.3 that this is not the case. When one believes the AUC or any other measure based on ROCs (such as the Neyman-Pearson criterion) has divorced the problem from assuming particular costs and/or prior probabilities, one is deceived.

The second viewpoint is present in the works of [2,3,41,42]. This viewpoint provides a way of working with prior probabilities and costs. It is the more valid viewpoint in our opinion, based on our theoretical developments. This is due to the fact that the problem of optimizing the ability to discriminate between multiple classes is an *optimization* problem with assumptions and constraints, not just a statistical problem. One cannot divorce the problem from the inherent prior probabilities and costs precisely because when you establish *any* criterion by which to make the discriminations, attached to it is an underlying cost-prior ratio, which is now simply hidden, so that one cannot escape from facing the costs and the prior probabilities of the problem. This is most clearly laid out in [40].

The problem is certainly expanded when one considers multiple class problems (problems where the classes in question number greater than two). A few papers have been written concerning this. In [16], the first viewpoint is used, and statistical estimates are developed to compare families of classification systems. In our view, this will lead to the selection of classification systems that are suboptimal to the problems where some knowledge regarding costs and prior probabilities exist. Also, we do not explore the inherently statistical nature of the work.

In [37], three classes, all mutually exclusive, are analyzed, and Volume Under the Curve (VUC) is explored as a measure of how well a family of classification systems performs. We have the same criticisms regarding the significance of this measure; however, the paper goes further, at least, in describing the geometry on which such a construct relies. There is no discrimination between the types of errors committed, since the axes

developed are each based on correct identifications, and not incorrect identifications. We show later that for k classes there are $k^2 - k$ error axes required for a full ROC manifold development and that only if errors within types have identical costs associated with them can we project the problem into k dimensions (so that for 3 classes you need 6 dimensions for a full ROC manifold, but could project into 3 dimensions only if the errors within classes have identical costs).

The authors' [41,42] show the greatest amount of promise in the field, by focusing on the optimization of costs. It is well known that if one considers a ROC curve as a function, with the independent variable being the false positive and the dependent variable the true positive, then under a mild assumption of smoothness, the ROC curve is differentiable, and one can show that to minimize the Bayes Cost function with two classes, one needs to find the point on the ROC curve which has a particular cost-prior ratio as a derivative. The only paper we found with a "proof" of this was [36], in which he claims the result can be shown. His own analysis fails to prove the achieved critical points are always a minimum. In fact, since the second derivative test is inconclusive, one must use the first derivative test to prove the minimum. The first derivative test is not available to us in the case of multivariate problems. We use calculus of variations and the global optimization theory of vector space methods to prove this not just for ROC curves, but we have extended the method to prove it for ROC manifolds, so that problems of multiple classes can be analyzed using the Bayesian methods. Our method involves only the geometry of the ROC manifold in ROC space, along with the same cost function (examined as the functional J in Chapter IV).

Much use is made in [41, 42] of the ROCCH. The usefulness of the ROCCH is apparent when one considers creating randomized decision rules from previously created families of classification systems. When two classification systems overlap, one can consider the ROCCH as a solution to which classification system to choose, since the ROCCH can be created under a convex combination of selecting probabilistically one family or the other. We show, however, that with respect to a particular cost-prior assumption, no ben-

efit is gained by doing this, since under these assumptions no greater cost benefit can be achieved along the extension of the convex hull than there is at the endpoints (which already exist). Under the first viewpoint, there is a benefit. That is the direct increase in the area under the curve, so that by randomly selecting the family from which to choose, one may increase the overall ability to separate the conditional class distribution functions, associated with the classification families. Similarly, in [12], the search is on to construct the convex polytope associated with three class (and n class) problems. The authors use the "trivial" classification systems to construct the best estimate of the convex polytope, but they also recognize the efforts of [37] and [16] in their approach.

ROCs also have an inherent application to detection problems involving electronics (thus the "receiver" in receiver operating characteristics). This history and analysis can be found in textbooks, particularly in [10, 30, 34, 45]. The emphasis in these texts are towards the development of classification systems and not the performance evaluation from ROCs only. In some respects the developments in these texts overlap with our development, but from the opposite approach. There are also some errors made in the texts, which are not apparent until you really dive into some of the analysis with respect to risk sets (particularly the min-max example in [45]). We need to point out two things with respect to this. First, our optimization is significantly different in its characterization of the problem. We use calculus of variations and properties of linear transformations to establish our optimization problem and we pick up on some differences that we feel are very important. Secondly, there is no connection to information fusion or category theory given in these texts. So our application is certainly new and independent and extends the field of knowledge. Overall, we believe our review to be sufficient to look into the use of ROCs, ROC curves, and ROC manifolds in order to produce a theory which is satisfactory to discriminating the performance of one classification system over another, or one family of classification systems over another. If an objective function can be produced on ROCs, ROC curves, and ROC manifolds, then we can define the qualitative

nature of fusion according to each application (that is, which fusion rules are better than the original classification systems, and which fusion rules are superior to others).

II. Background

2.1 Mathematical Formalisms and Definitions

This work is inherently an applied math dissertation. As such, the background material required in order to understand it is drawn from the areas of Topology, Category Theory, Probability Theory (measure-theoretic in scope), and some vector space knowledge. We assume a basic knowledge of vector spaces is understood by the reader, but certain useful definitions and theorems are stated concisely in this section to facilitate the readers understanding. Definitions of receiver operating characteristics (ROCs), ROC curves, and ROC manifolds are also given, along with a couple of convergence theorems useful to understanding the context of why ROC analysis has drawn such attention from researchers.

2.1.1 Topology Formalisms and Definitions.

Definition 1 (Preimage). Let f be a function with X the domain of f and Y the range. Then given $B \subset Y$, we denote the preimage of B over f by $f^{\natural}(B)$, where

$$f^{\natural}(B) = \{x \in X \,:\, f(x) \in B \subset Y\}. \tag{2.1}$$

The symbol $\natural$ is the natural symbol from music literature (also known as the becuadro) and is used precisely because we do not want to confuse the preimage of a set over a function with the inverse of the function, which is denoted as f^{-1}.

Definition 2 (Topology, Topological Space [38]). A topology τ on a set X is a collection of subsets of X such that:

i. $X, \emptyset \in \tau$.

ii. Arbitrary collections of sets of τ have their unions in τ.

iii. Finite collections of sets of τ have their intersections in τ.

The sets contained in τ are called the open sets of X. We say (X, τ) is a topological space.

Example 1. Here is an example from [38]. Given a set X, with an order relation $<$, and $a, b \in X$, the following types of sets are in the topology:

1. $(a, b) = \{x \mid a < x < b\}$;
2. $[a_0, b) = \{x \mid a_0 \leq x < b\}$, where a_0 is the smallest element of X (if one exists);
3. $(a, b_0] = \{x \mid a < x \leq b_0\}$, where b_0 is the largest element of X (if one exists);

The collection $\mathscr{B}$ of such sets for all $a, b \in X$ is the **order topology** on X.

Definition 3 (Hausdorff Space). A topological space (X, τ) is a Hausdorff space if for any two elements $x_1, x_2 \in X$ with $x_1 \neq x_2$, there exists open sets $U, V \in \tau$ such that $x_1 \in U$ and $x_2 \in V$, with $U \cap V = \emptyset$.

Example 2. The set of real numbers, $\mathbb{R}$, with the Euclidean metric of distance between two points, is an example of an Hausdorff Space.

Definition 4 (Metric, Metric Space). Let X be a set. Then for $x, y, z \in X$, if there exists a function d, such that $d : X \times X \to \mathbb{R}^+$, which satisfies the conditions:

i. $d(x, y) \geq 0$ (non-negativity);

ii. $d(x, y) = 0$ **iff** $x = y$ (positive definiteness);

iii. $d(x, y) = d(y, x)$ (symmetry);

iv. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality);

then d is a metric. We call (X, d) a metric space, though the notation is often suppressed to simply X.

Example 3. The Euclidean metric of distance, given $x, y \in X$, $d(x, y) = |x - y|$ is a metric. Every Euclidean metric induces a topology as well, since open sets can be defined in terms of Euclidean distance, and a basis for such topologies can easily be formed using open balls in X.

Definition 5 (Open Ball in $\mathbb{R}^n$). An open ball in $\mathbb{R}^n$ relative to a metric d is written $\mathcal{B}(x;r)$ where there is no misunderstanding of the metric. The meaning of the open ball is

$$\mathcal{B}(x;r) = \{y \in \mathbb{R}^n \mid d(x,y) < r\},$$

where $x \in \mathbb{R}^n$ is the center of the ball with r its radius.

Example 4. Let $(\mathbb{R}, d)$ be a metric space. Then for $\varepsilon > 0$ and $x \in \mathbb{R}$,

$$\mathcal{B}(x;\varepsilon) = \{y \,:\, d(x,y) < \varepsilon\}$$

forms an open ball in $\mathbb{R}$.

Definition 6 (m-Manifold [38]). Let $m \in \mathbb{N}$ be given. A topological space (X, τ) is an m-Manifold if it is a Hausdorff space and has a countable basis such that each neighborhood of a point $x_1 \in X$ is homeomorphic with an open subset in $\mathbb{R}^m$.

Example 5. In $\mathbb{R}^n$, $n \geq 3$, a 1-manifold is a curve, a 2-manifold is a surface, etc.

2.1.2 Probability Theory Formalisms and Definitions. Necessary to reading this dissertation is a common frame of reference with regard to category theory and probability theory. We will start with the latter and the reader can always familiarize himself with [5] for probability theory, or [43, 44] for elementary measure theory.

Definition 7 (Algebra or Field of Sets). Let X be an arbitrary set. A collection $\mathscr{B}$ of subsets of X is an algebra or a field if it satisfies three properties:

i. $X \in \mathscr{B}$.

ii. For any $B \in \mathscr{B}$, we have the set complement, $X \backslash B$, written $\tilde{B}$, also in $\mathscr{B}$.

iii. Given the finite collection $\{B_i \in \mathscr{B} \,:\, i = 1, 2, \ldots, n \in \mathbb{N}\}$, then $\bigcup_{i=1}^{n} B_i \in \mathscr{B}$.

Definition 8 (σ-algebra or σ-field). Let X be an arbitrary set. A collection of subsets, $\mathscr{B}$, of X is a σ-algebra, or σ-field, on X if $\mathscr{B}$ satisfies three properties:

i. $X \in \mathscr{B}$.

ii. For any $B \in \mathscr{B}$, we have $\tilde{B}$ is also in $\mathscr{B}$.

iii. Given the countably infinite collection $\{B_i \in \mathscr{B} : i = 1, 2, \ldots\}$, then $\bigcup_{i=1}^{\infty} B_i \in \mathscr{B}$.

We can see that a σ-field is a field of sets as well.

Example 6. Given a set X, the power set of X, $\mathscr{P}(X)$, is a σ-field.

Definition 9 (Positive Measure). Let X be a set and $\mathscr{B}$ be a σ-field over X, then any set function ν defined on $\mathscr{B}$ with range $\mathbb{R}[0, \infty]$ is called a positive measure on X if it is **countably additive**. That is, given a disjoint, countable collection, $\{B_i\}_{i=1}^{\infty}$, of sets in $\mathscr{B}$, then

$$\nu\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \nu(B_i).$$

Example 7. See [43] for the definition of outer measure on $\mathbb{R}$. Outer measure returns lengths of intervals on $\mathbb{R}$, so that outer measure is a positive measure on $\mathbb{R}$.

Definition 10 (Sample Space). Given a complex Γ of conditions, which allows any number of repetitions (an experiment, for example), there is a collection of elementary events, $\xi, \varsigma, \zeta, \ldots$, not necessarily countable, which is called the sample space, and will be denoted as Ω [28].

Example 8. An example of a complex of conditions Γ is "the tossing of a coin", while the sample space $\Omega = \{h, t\}$, where h is the event of getting a "head" as a result, and t the event of getting a "tail". If Γ is "the tossing of a coin two times", then
$\Omega = \{h, h, h, t, t, h, t, t\}$ is the sample space. If the event T implies that a tail results, then any of the last three elementary events of Ω has occurred. The complex of conditions will usually be described using language in order to identify meaning. This language then leads to the formation of the sample space and the σ-field so that a probability measure can be defined. All probabilistic mathematical language for the given problem flows from this beginning.

A σ-field, $\mathscr{B}$, can be developed on Ω, so that the pair $(\Omega, \mathscr{B})$ is a measurable space. Given a positive measure μ on $(\Omega, \mathscr{B})$, the triple $(\Omega, \mathscr{B}, \mu)$ is called a measure space.

Definition 11 (Measurable Function). Let $(X, \mathscr{B})$ be a measurable space and (Y, τ) be a topological space. A function f is called measurable if for each $E \in \tau$, we have that the preimage of E under f, denoted $f^{\natural}(E)$, is also an element of $\mathscr{B}$. We call f a $\mathscr{B}$-measurable function.

Example 9. Let μ be Lesbegue measure on $\mathbb{R}$. Consider any continuous function f with compact support. Since the preimage of an open set is open for continuous functions, and open sets are always contained in the Borel σ-field, we have that these functions are measurable.

Definition 12 (Finite, σ-Finite Measures). If $\mu(\Omega) < \infty$, then μ is a finite measure. A measure μ is σ-**finite** if there exists a sequence $\{B_n\}$ of elements of $\mathscr{B}$ such that $\Omega = \bigcup_{n=1}^{\infty} B_n$ and $\mu\big(B_n\big) < \infty$ for each $n \in \mathbb{N}$. Finite measures are clearly σ-finite as well.

Example 10. Lesbegue measure μ on $\mathbb{R}$ is an example of a σ-finite measure. Consider the countable balls with radius $\varepsilon > 0$ and centers $x \in \mathbb{Q}$. The union of these balls is $\mathbb{R}$, while the measure of each ball is finite.

Definition 13 (Probability Measure, Probability Space [49]). Given a measurable space $(\Omega, \mathscr{B})$, a positive measure μ with $\mu(\Omega) = 1$ is defined as a probability measure. A probability measure is a finite measure and therefore a σ-finite measure as well. The measure space $(\Omega, \mathscr{B}, \mu)$ is called a probability space.

Notice how the properties of the definitions flow:

> Since $\Omega \in \mathscr{B}$, then $\emptyset \in \mathscr{B}$. Therefore, since $\mu(\Omega) = \mu(\Omega \cup \emptyset)$, we have by the definition of a positive measure the property of countable additivity, so that $\mu(\Omega) = \mu(\Omega) + \mu(\emptyset)$. Since $\mu(\Omega) = 1$, we have that
>
> $$1 = 1 + \mu(\emptyset),$$
>
> so that $\mu(\emptyset) = 0$. Furthermore, for any $B \in \mathscr{B}$, we have $0 \leq \mu(B) \leq 1$, and since $\mu(\Omega) = \mu(B) + \mu(\tilde{B})$ we have that
>
> $$1 - \mu(B) = \mu(\tilde{B}).$$

Definition 14 (Bayes Theorem). Given a probability space $(\Omega, \mathscr{B}, \mu)$, and $B_1, B_2 \in \mathscr{B}$, the conditional probability of B_1 given B_2 is written

$$P(B_1|B_2) = \frac{\mu(B_1 \cap B_2)}{\mu(B_2)}.$$

With this in mind, Bayes Theorem states

$$P(B_1|B_2)\mu(B_2) = P(B_2|B_1)\mu(B_1), \tag{2.2}$$

so that

$$P(B_1|B_2) = \frac{\mu(B_1)}{\mu(B_2)}P(B_2|B_1). \tag{2.3}$$

The notation is written to emphasize that P is not a measure, but rather given an event B_2, then $P(\cdot|B_2)$ is a measure which is related to the measure μ by the definition. We refer to the left hand side of the equation as the conditional probability of the event B_1 given event B_2 has occurred, while the conditional probability on the right-hand side of the equation is referred to as the posterior probability. Each real number $\mu(B_1)$ and $\mu(B_2)$ is a prior probability.

Definition 15 (Random Variable). Given a probability space $(\Omega, \mathscr{B}, \mu)$ and a measurable space $(\acute{\Omega}, \acute{\mathscr{B}})$, we say $f : \Omega \longrightarrow \acute{\Omega}$ is a random variable if $f^\natural(E) \in \mathscr{B}$ for each $E \in \acute{\mathscr{B}}$. We say f is an $\acute{\Omega}$-valued random variable.

Note: It is true that a Φ-valued, measurable function, $g : \Omega \longrightarrow \Phi$, is a Φ-valued random variable for any topological space (Φ, τ), since there always exists a smallest σ-field containing the elements of τ [44]. The specific language "random variable", without the hyphenated prefix, is reserved for the case when $\Phi = \mathbb{R}$.

Definition 16 (Stochastic Process [9]). Let $(\Omega, \mathscr{B}, \mu)$ be a probability space and Θ a set of parameters which may be finite, countably infinite, or uncountable. Then a family of random variables indexed by Θ, $\mathbb{X} = \{X_\theta : \theta \in \Theta\}$ is a stochastic process. If Θ is countably infinite or finite, then $\mathbb{X}$ is a $\mathbf{discrete\ parameter\ process}$. If Θ is a continuous

parameter, then $\mathbb{X}$ is a **continuous parameter process**. If we fix $\omega \in \Omega$, and allow θ to vary, then the function $X_\cdot(\omega)$ is a sample function when Θ is uncountable. When Θ is countable or finite, then $X_\cdot(\omega)$ is a sample sequence.

2.1.3 Category Theory. This section draws upon definitions contained in [53]. Category theory is a branch of mathematics useful for determining universal properties of objects. The science of information fusion does not yet know of all the relationships involved between the classes of data and the mappings from one type of data to another. It has been our goal to try to engage the community to think in terms of generalities when studying fusion processes in order to abstract the processes and perhaps gain some clarity of thought, if not genuine insight. We have drawn upon the work of various authors in Category Theory literature [1, 29, 32, 35] to present the definitions.

Definition 17 (Category). A category $\mathcal{C}$ is denoted as a 4-tuple,

$$\mathcal{C} = \Big(\mathbf{Ob}(\mathcal{C}), \mathbf{Ar}(\mathcal{C}), \mathbf{Id}(\mathcal{C}), \circ\Big),$$

and consists of the following:

A1. A class of objects denoted $\mathbf{Ob}(\mathcal{C})$, so object $O \in \mathbf{Ob}(\mathcal{C})$.

A2. A class of arrows denoted $\mathbf{Ar}(\mathcal{C})$, so arrow $f \in \mathbf{Ar}(\mathcal{C})$.

A3. Two mappings, called Domain (**dom**) and Codomain (**cod**), which assign to an arrow $f \in \mathbf{Ar}(\mathcal{C})$ a domain and codomain from the objects of $\mathbf{Ob}(\mathcal{C})$. Thus, for arrow $f \in \mathbf{Ar}(\mathcal{C})$, there exist objects $O_1 = \mathbf{dom}(f)$ and $O_2 = \mathbf{cod}(f)$ and we represent the arrow f by the diagram

$$O_1 \xrightarrow{f} O_2 \, .$$

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

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

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

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

A5. A binary mapping, $\circ$, called composition, $\mathbf{Ar}(\mathcal{C}) \times \mathbf{Ar}(\mathcal{C}) \xrightarrow{\circ} \mathbf{Ar}(\mathcal{C})$. Thus, given $f, g \in \mathbf{Ar}(\mathcal{C})$ with **cod**$(f) =$ **dom**(g) there exists an unique $h \in \mathbf{Ar}(\mathcal{C})$ 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 $h \in \mathbf{Ar}(\mathcal{C})$ 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 arrow $1_A \in$ **Id**$(\mathcal{C})$ such that $1_A \circ g = g$ and $f \circ 1_A = f$.

Definition 18 (Subcategory). A subcategory $\mathcal{B}$ of $\mathcal{A}$ is a category whose objects are some of the objects of $\mathcal{A}$, i.e., **Ob**$(\mathcal{B}) \subset$ **Ob**$(\mathcal{A})$, and whose arrows are some of the arrows of $\mathcal{A}$, i.e., **Ar**$(\mathcal{B}) \subset$ **Ar**$(\mathcal{A})$, such that for each arrow $f \in$ **Ar**$(\mathcal{B})$, **dom**(f) and **cod**(f) are in $\mathbf{Ob}(\mathcal{B})$, along with each composition of arrows, and an identity arrow for each element of $\mathbf{Ob}(\mathcal{B})$.

Definition 19 (Discrete Category). A $\mathbf{discrete}$ category is a category whose only arrows are identity arrows, i.e., **Ar**$(\mathcal{C}) =$ **Id**$(\mathcal{C})$.

Definition 20 (Small Category). A category C is called a **Small Category** when the class **Ob**(C) is a set.

Note: A historical note on this is that while in this paper and in many works, the only categories considered are small categories, category theorists are proposing an axiomatic replacement for set theory as a mathematical foundation. In other words, all mathematical properties can be shown using an axiomatic category theory rather than the Zermelo-Fraenkel axioms for set theory. The belief is that the category theory approach will avoid

certain paradoxes which creep up in set theory, such as "the set whose members are not in a set".

Definition 21 (Functor). A **functor** $\mathfrak{F}$ between two categories $\mathcal{A}$ and $\mathcal{B}$ is a pair of maps $(\mathfrak{F}_{\mathbf{Ob}}, \mathfrak{F}_{\mathbf{Ar}})$

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

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

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

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

i) $\mathfrak{F}_{\mathbf{Ob}}(\mathrm{A}) = \mathrm{B}$.

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) $\mathfrak{F}_{\mathbf{Ar}}(1_{\mathrm{A}}) = 1_{\mathfrak{F}_{\mathbf{Ob}}(\mathrm{A})} = 1_{\mathrm{B}}$.

We denote a functor $\mathfrak{F}$ between categories $\mathcal{A}$ and $\mathcal{B}$ with the diagram

$$\mathcal{A} \xrightarrow{\mathfrak{F}} \mathcal{B}\ .$$

Example 11. An elementary example of a functor is the forgetful functor. Let **GRP** be the category of groups which has as objects groups and as arrows morphisms between groups. Let $\mathscr{U}(\mathrm{G})$ denote the underlying set of elements of a given group G. Then the forgetful functor, $\mathfrak{F}$, maps groups to their underlying sets, and all arrows to the identity arrow on the underlying set.

Definition 22 (Natural Transformation). Given categories $\mathcal{A}$ and $\mathcal{B}$ and functors $\mathfrak{F}$ and $\mathfrak{G}$ with $\mathcal{A} \xrightarrow{\mathfrak{F}} \mathcal{B}$ and $\mathcal{A} \xrightarrow{\mathfrak{G}} \mathcal{B}$, then a **Natural Transformation** is a family of arrows

$\nu = \{\nu_{\mathrm{A}} : \mathrm{A} \in \mathbf{Ob}(\mathcal{A})\}$ such that for each $f \in \mathbf{Ar}(\mathcal{A})$, $\mathrm{A} \xrightarrow{f} \mathrm{A}'$, $\mathrm{A}', \mathrm{A} \in \mathbf{Ob}(\mathcal{A})$, the square diagram

$$\begin{array}{ccc} \mathrm{A} & & \mathfrak{F}(\mathrm{A}) \xrightarrow{\nu_{\mathrm{A}}} \mathfrak{G}(\mathrm{A}) \\ f \downarrow & & \mathfrak{F}(f) \downarrow \qquad\qquad \downarrow \mathfrak{G}(f) \\ \mathrm{A}' & & \mathfrak{F}(\mathrm{A}') \xrightarrow{\nu_{\mathrm{A}'}} \mathfrak{G}(\mathrm{A}') \end{array}$$

commutes. We say the arrows ν_{A} are the components of

$$\nu : \mathfrak{F} \longrightarrow \mathfrak{G} ,$$

and call ν the natural transformation of $\mathfrak{F}$ to $\mathfrak{G}$.

Example 12. This example is from [32]. Let $\mathbf{CRng}$ be the category of commutative rings, and $\mathbf{GL}_n(\cdot)$ be the category of general linear groups, which consists of all $n \times n$ invertible matrices over commutative ring $(\cdot)$. The determinant of the matrices is a natural transformation (since the matrices are calculated with the same formula regardless of the ring used) making the following square commute (K^*, K'^* are rings with their additive identity removed, so that all of the elements are invertible, and therefore they are objects of the category **GRP**.):

$$\begin{array}{ccc} K & & \mathbf{GL}_n(K) \xrightarrow{\mathrm{Det}_K} K^* \\ f \downarrow & & \mathbf{GL}_n(f) \downarrow \qquad\qquad \downarrow f^* \\ K' & & \mathbf{GL}_n(K') \xrightarrow{\mathrm{Det}_{K'}} K'^* \end{array}$$

This says that for every morphism, f of commutative rings, the determinant is natural among functors $\mathbf{CRng} \longrightarrow \mathbf{GRP}$.

Definition 23 (Functor Category $\mathcal{A}^{\mathcal{B}}$). Given categories $\mathcal{A}$ and $\mathcal{B}$, the notation $\mathcal{A}^{\mathcal{B}}$ represents the category of all functors $\mathfrak{F}$ such that $\mathcal{B} \xrightarrow{\mathfrak{F}} \mathcal{A}$. This category has all such functors as objects and the natural transformations between them as arrows. We can also

have that given objects $A \in \mathbf{OB}(\mathcal{A})$ and $B \in \mathbf{OB}(\mathcal{B})$, there exists functor categories denoted as $A^{\mathcal{B}}$, $\mathcal{A}^B$, and A^B as well.

Definition 24 (Product Category). Let $\{\mathcal{C}_i\}_{i=1}^n$ be a finite collection of small categories. Then the cartesian product

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

forms a category called the product category. For each $\mathbf{O} \in \mathbf{Ob}\left(\prod_{i=1}^{n} \mathcal{C}_i\right)$, then $\mathbf{O} = (O_1, O_2, \ldots, O_n)$ where $O_i \in \mathbf{Ob}(C_i)$ for $i = 1, 2, \ldots, n$. For each arrow $\boldsymbol{f} \in \mathbf{Ar}\left(\prod_{i=1}^{n} \mathcal{C}_i\right)$, then $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$ where $f_i \in \mathbf{Ar}(C_i)$ for $i = 1, 2, \ldots, n$. Given arrows $\boldsymbol{f}, \boldsymbol{g} \in \mathbf{Ar}\left(\prod_{i=1}^{n} \mathcal{C}_i\right)$, then the composition of these arrows mean $\boldsymbol{f} \circ \boldsymbol{g} = (f_1 \circ g_1, f_2 \circ g_2, \ldots, f_n \circ g_n)$

2.1.3.1 Category Examples.

Some examples of categories are:

Example 13. The category of all Abelian Groups, **Ab**. Here the objects are abelian groups and the arrows are all morphisms from one Abelian Group to another.

Example 14. The category **Ban** of Banach Spaces. Here the objects are Banach Spaces, and the arrows are all bounded linear transformations between them.

Example 15. The category $\mathbf{Vect}_{\mathbb{K}}$ of finite-dimensional Vector Spaces over the field $\mathbb{K}$. The objects are finite vector spaces and the arrows are all linear transformations between them.

Example 16. The category $\mathbf{SET}$, a small category whose objects are sets and arrows are the total functions between them.

Example 17. The category $\mathbf{CAT}$, a small category whose objects are small categories and arrows are the functors between them.

Some examples of functors are:

Example 18. $\mathfrak{F}$: **Ab** $\rightarrow$ **SET**, which is the forgetful functor which simply maps all non-identity arrows in the category **Ab** which map from an object to the identity arrow of that object, now considered as a set only within the category **SET**, rather than a group.

Example 19. $\mathfrak{G}$: **Ban(X)** $\rightarrow$ **Set**, the functor mapping all subspaces of a Banach space X to their respective subsets. Non-identity arrows are mapped to identity arrows, so this functor is also a "forgetful" functor.

2.2 Receiver Operating Characteristic (ROC) Background

2.2.1 Definition of ROC curve. Let $(\Omega, \mathscr{B}, \mu)$ be a probability space, L be a two-class label set, $\mathrm{L} = \{\ell_1, \ell_2\}$, and let $X(t, \cdot) : \Omega \rightarrow \mathrm{L}$ be a discrete random variable indexed by a **parameter** set **T**, where $t \in \mathbf{T}$ is a parameter, and **T** might be uncountable and multidimensional. We will refer to the sample function $X_t(\cdot) = X(t, \cdot)$ as a classifier of members of $\mathscr{B}$. Usually, **T** is homeomorphic to some subset of $\mathbb{R}^m$ for some $m \in \mathbb{N}$. We assume Ω can be partitioned into two sets of events, so $\Omega = \Omega_1 \cup \Omega_2$, where the first set Ω_1 corresponds to the label ℓ_1, and the second to the label ℓ_2. Thus, $\Omega_1 \cap \Omega_2 = \emptyset$ is assumed. Under the assumption of only two labels, we will assume that **T** is a one-dimensional parameter space.

Each classifier X_t can make a mistake in classification. There are two types of errors it can make. It can assign objects in class 1 to label ℓ_2, or it can assign objects in class 2 to label ℓ_1. Let $X_t^{\natural}(\ell_i)$ denote the pre-image of the label ℓ_i under classifier $X(t, \cdot)$. We can construct the two conditional probabilities of a classifier making these errors as

$$p_{1|2}(t) = P(\ell_1 \mid \Omega_2) = \frac{\mu(X_t^{\natural}(\ell_1) \cap \Omega_2)}{\mu(\Omega_2)}. \tag{2.4}$$

and

$$p_{2|1}(t) = P(\ell_2 \mid \Omega_1) = \frac{\mu(X_t^{\natural}(\ell_2) \cap \Omega_1)}{\mu(\Omega_1)}. \tag{2.5}$$

where $\mu(\Omega_2)$ and $\mu(\Omega_1)$ are the prior probabilities of their respective events and the $p_{i|j}(t)$ for $i, j = 1, 2$ are the conditional class probabilities of classifying an event as ℓ_i when event ℓ_j has occured. Two conjuctive, conditional class probabilities, constructed in the same manner, form the following relationships [11]:

$$p_{1|2}(t) + p_{2|2}(t) = 1 \tag{2.6}$$

$$p_{2|1}(t) + p_{1|1}(t) = 1 \tag{2.7}$$

For a specific $t \in \mathbf{T}$, the ordered pair, $(p_{1|2}(t), p_{1|1}(t))$ is called the receiver operating characteristic (ROC) of classifier $X(t, \cdot)$, when the dependent class is Ω_1. We will use the notation $(p_{1|2}(t), p_{2|1}(t))$ as the ROC, however, to better accommodate our description of the n-class problem. A set $\mathbb{X} = \{X(t, \cdot) : t \in \mathbf{T}\}$ is called a family of classification systems (alternatively, a classifier family). We say the set of triples formed by $\mathbb{X}$,

$$\tilde{f}_{\mathbb{X}} = \big\{(t, p_{1|2}(t), p_{2|1}(t)) : t \in \mathbf{T}, \Omega = \Omega_1 \cup \Omega_2\big\}$$

forms a ROC trajectory, when it is lower semi-continuous and monotonic, non-increasing (see [9] regarding distribution function properties for comparison). We say the set

$$f_{\mathbb{X}} = \big\{(p_{1|2}(t), p_{2|1}(t)) : t \in \mathbf{T}, \Omega = \Omega_1 \cup \Omega_2\big\}, \tag{2.8}$$

which is the projection of $\tilde{f}_{\mathbb{X}}$ from the space $\mathbf{T} \times \mathbb{R}[0, 1]^2$ into the space $\mathbb{R}[0, 1]^2$, is the ROC curve of family of classification systems $\mathbb{X}$, when its closure has endpoints in the compact interval $\mathbb{R}[0, 1]$, and it is lower semi-continuous, and monotonic non-increasing. We also call the ROC curve the ROC manifold (technically, a ROC 1-manifold, see Lemma 2.2.2), since this curve is homeomorphic to $\mathbb{R}^1$ for every open ball on the curve and it is a Hausdorff space.

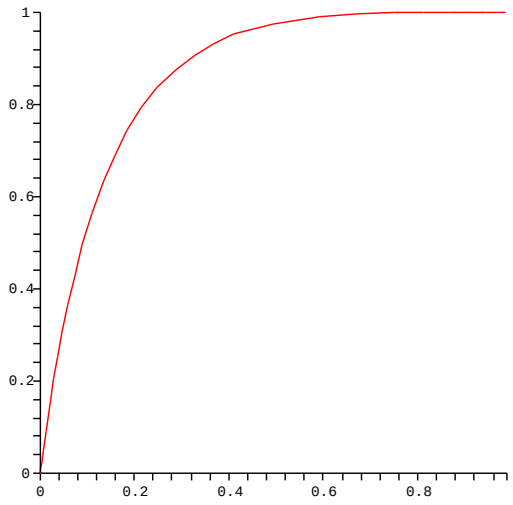


Figure 2.1: A Typical ROC Curve from Two Normal Distributions

Definition 25 (Proper ROC Curve [4]). Given a metric space $(\mathbf{T}, d)$ and a (one-dimensional) parameter $t \in \mathbf{T}$, a continuous ROC curve $f_{\mathbb{X}}$ as defined in Equation 2.8 is called a proper ROC curve when

1. $\lim_{t \to \infty} (p_{2|1}(t), p_{1|2}(t)) = (0, 1)$.
2. $\lim_{t \to -\infty} (p_{2|1}(t), p_{1|2}(t)) = (1, 0)$.

Typically, ROC curves are graphed using $(p_{2|1}(t), p_{1|1}(t))$ as coordinate pairs , producing a curve from $(0, 0)$ to $(1, 1)$. For multi-class problems (greater than two classes), this is not the best visualization scheme to follow.

2.2.2 ROC Space. Many publications refer to the real set product $\mathbb{R}([0, 1]) \times \mathbb{R}([0, 1])$ as ROC space. This terminology is unfortunate since $\mathbb{R}([0, 1])$ is not a 'space' in the sense of a linear space. We clarify here that by the term ROC space we mean the topological subspace $(\mathbb{R}^2([0, 1]), \tau)$ of $(\mathbb{R}^2, \tau)$ where τ is the Euclidean topology (the topology induced by the usual distance metric).

Lemma 1 (ROC 1-Manifold). *A proper ROC curve is a 1-manifold in ROC space.*

Proof: Let $\mathbf{S} = \big\{(P_{2|1}(A_\theta), P_{1|2}(A_\theta)) : \theta \in \Theta, \Omega = \Omega_1 \cup \Omega_2, A_\theta \in \mathbb{A}\big\}$ be a proper ROC curve, with $\{\Omega_1, \Omega_2\}$ a partition of Ω into two classes. Let $x(\theta) = P_{2|1}(A_\theta)$, $y(\theta) = P_{1|2}(A_\theta)$, and let $x = x(\theta)$,$y = y(\theta)$ for brevity of notation. Let $\varepsilon > 0$ be given. The norm $\|\cdot\|$ is Euclidean 2-norm. An open set A in **S** is open relative to the usual $\mathbb{R}^2$ topology. There is a countable basis for this topology which consists of the open balls of rational radius r about each coordinate point with rational first component. To show **S** is Hausdorff let (x, y) and (w, z) be two distinct points in **S**. Then we have that

$$\|(x, y) - (w, z)\| = \delta,$$

for some $\delta \in \mathbb{R}$. Let $\gamma = \frac{\delta}{2}$. Thus we have that

$$\mathcal{B}((x, y); \gamma) \cap \mathcal{B}((w, z); \gamma) = \emptyset$$

are two intersecting open sets containing the two distinct points.

Now let $(x, y) \in \mathbf{S}$ be given. Define a function $g : \mathcal{B}\big((x, y); \varepsilon\big) \to \mathcal{B}\big(x; \varepsilon\big) \subseteq \mathbb{R}^1$ by

$$g[(x, y)] = x, \forall\ (x, y) \in \mathcal{B}\big((x, y); \varepsilon\big),$$

where $\mathcal{B}\big(\cdot; \varepsilon\big)$ is an open ball of radius ε with center $\cdot$. Clearly, g is one-to-one, since for $z \in \mathbb{R}$ such that $x = z$ we have that there exists $y_2 \in \mathbb{R}$ with $g[(z, y_2)] = z$. Thus, if $(x, y) \neq (z, y_2)$, then either $x \neq z$ (which is a contradiction) or $y \neq y_2$. Suppose $y \neq y_2$. Then **S** is not a set representation of a function, which is a contradiction, since this is implicit in the definition of **S**. Therefore, $(x, y) = (z, y_2)$, and g is one-to-one. Thus, g has an inverse, g^{-1}.

Now, let $\xi > 0$ be given, with $\varepsilon > \xi > 0$. Then for $(x_2, y_2) \in \mathcal{B}\big((x, y); \varepsilon\big)$ such that

$$\|(x, y) - (x_2, y_2)\| < \xi$$

we have that

$$\begin{aligned} \|g[(x,y)] - g[(x_2,y_2)]\| &= \|x - x_2\| \\ &\leq \sqrt{(x-x_2)^2 + (y-y_2)^2} \\ &= \|(x-x_2, y-y_2)\| \\ &= \|(x,y) - (x_2,y_2)\| \\ &< \xi \end{aligned} \tag{2.9}$$

so that g is continuous as well. Since g is continuous over every compact subset of $\mathcal{B}\big((x,y);\varepsilon\big)$, g^{-1} is continuous on $g(\mathcal{B}\big((x,y);\varepsilon\big))$. Now, there exists an open set, $O \subseteq \mathbb{R}[0,1]$, such that $O \subseteq g[\mathcal{B}\big((x,y);\varepsilon\big)]$ and $g^{\natural}(O) = \mathcal{B}\big((x,y);\varepsilon\big)$. Hence, for all $o \in O$, $g^{-1}(o) \in \mathcal{B}\big((x,y);\varepsilon\big)$. Since, $g[g^{-1}(o)] \in g\Big(\mathcal{B}\big((x,y);\varepsilon\big)\Big)$ for all $o \in O$, we have that $\mathcal{B}\big((x,y);\varepsilon\big) \subseteq \mathbf{S}$ is homeomorphic to $O \subseteq \mathbb{R}^1$, with

$$g : \mathcal{B}\big((x,y);\varepsilon\big) \to O$$

being the homeomorphism, so that **S** is a 1-manifold in ROC-space. $\diamondsuit$

An example is seen in Figure 2.1. This proof can be extended to show that a ROC surface in n-space is a ROC $(n-1)$-manifold, the basis of the manifold being the points on the ROC surface corresponding to $(r_1, r_2, \ldots, r_{n-1}, x_n)$ where the first $n-1$ components are rational numbers, with x_n being the dependent component, along with rational radii in an $(n-1)$-ball open relative to the $\mathbb{R}^n$ topology.

2.2.3 ROC n-Space. We will retain the conventional language of ROC space and offer an extension to $n2$ dimensions. Suppose we have a multi-class label set (a label set with more than two labels). To construct a corresponding ROC space, in the case of $m > 2$ labels, we desire to have $n = m^2 - m$ axes, so we will designate this ROC space as a ROC n-space. This is due to the fact that when there are m classes, the number of possible types of classifications of the classification system is m^2 and the

number of conjunctive conditional probability equations is m (which also corresponds to the number of correct classifications), so that there are $m^2 - m$ degrees of freedom left after the application of the conjunctive equations (instead of the usual fact of contingency tables allowing $m^2 - 1$ degrees of freedom), which we have already seen in the case of $m = 2$ with the application of the conjuctive equations in Equations 2.6 and 2.7. So if we associate a correct classification with the m conjunctive equations, then we have $m^2 - m$ incorrect classifications corresponding to the degrees of freedom, each demanding its own axis in ROC n-space. If we were to allow all errors to have equal cost, then we can combine all errors within a class, and we would then have $n = m^2 - m(m-1) = m$ degrees of freedom, which is the same as the number of classes, each one requiring its own axis. When $m = 2$, we have that $n = 2$, which results in the typical ROC space of ROC curves.

In the case of three classes, $m = 3$, as an example, examine the conjunctive conditional probability equations (with notation suppressed with respect to the sample functions involved),

$$\begin{aligned} p_{1|1} + p_{2|1} + p_{3|1} &= 1 \\ p_{1|2} + p_{2|2} + p_{3|2} &= 1 \\ p_{1|3} + p_{2|3} + p_{3|3} &= 1 \end{aligned}$$

for $i, j = 1, 2, 3$. This system could be described by a 3×3 stochastic matrix. Notice that once the errors of each row are given, the correct classification is completely determined

by the equation. Additionally, the equations could be rewritten as

$$\begin{aligned} p_{2|1} + p_{3|1} &= 1 - p_{1|1} \\ p_{1|2} + p_{3|2} &= 1 - p_{2|2} \\ p_{1|3} + p_{2|3} &= 1 - p_{3|3} \end{aligned}$$

so that the ROC space needed to describe the system completely is now 3-space due to all costs being equal (in this case, cost $c_{i,j} = 1, \forall\ i, j$). There is a relationship between the dimensionality of a parameter set **T** and the dimensionality of the ROC manifold. Ultimately, we want to construct ROC manifolds which allow a unique optimization point to be embedded in the manifold, while maintaining independence of the conditional probabilities of $n - 1$ classes. If

$$r = \mathbf{dim}(\mathbf{T}) > n - 1,$$

then there are several optimal points embedded in the ROC r-manifold, so that a unique solution cannot be found analytically. If $r < n - 1$, then a unique optimization point embedded in the ROC r-manifold can be found, but independent control over all of the conditional probabilities is lost and information corresponding to each class is incomplete. Therefore, when we refer to ROC n-space, the ROC manifolds assumed to inhabit it are ROC $(n - 1)$-manifolds unless otherwise declared. This means the parameter space **T** is assumed to be of dimension $n - 1$, and this guarantees a unique optimization point with respect to the assumptions on prior probabilities and costs.

2.2.4 Convergence of Receiver Operating Characteristic (ROC) curves. Albert Einstein once said, "Not everything that can be counted counts, and not everything that counts can be counted." [47] Part of the reason we use ROC curves is due to their inherent dependency upon probability theory. Some sets have measure (they count, but may not be countable), and some have none (they don't count, though they may be countable). The

ROC curve is a graph of tradeoffs of the errors made by families of classification systems. Virtually all ROC manifolds are estimates of performance and do not meet the theoretical constraints we have defined. However, Alsing, in his Ph.D. dissertation [4], put forth a theorem which shows that estimates of ROC curves, created from calculating the true positive and false positive rates, converge to **the** ROC curve of a family of classification systems. We rely upon this convergence when we discuss the theory, because without it, not much makes sense. Therefore, we refer to **the** ROC curve. Alsing's proof of ROC convergence focused on two things:

1. $\widehat{p}_{i|j}^{(n)}(t)$ are estimates (random variables) which depend upon the actual (he says finite) data collected during the test.
2. There is a collection of metrics which show, for $\widehat{\mathbf{P}}_n(t) = \left(\widehat{p}_{1|2}^{(n)}(t), \widehat{p}_{1|1}^{(n)}(t)\right)$ and metric d in his collection, that

$$\lim_{n\to\infty} d\left(\widehat{\mathbf{P}}_n(t), \mathbf{P}(t)\right) = 0$$

   for some **P**. This **P** is referred to as **the** [emphasis mine] ROC curve.

With this proof we can theorize more about the actual underlying ROC curves and compare the systems they represent without much worry over our goals. After all, if we have a non-continuous collection of ROC points from a family of classification systems, we can approximate the underlying continuous ROC curve by connecting the points with straight lines. We can then imagine a sequence of such ROC curves *converging* to the ROC curve in order to talk about where the optimal points on the curve are, and perhaps to compare one curve generated by a family of classification systems to a finite number of other curves generated by different families of classification systems.

There are a few problems with Alsing's approach and proof. First, it relies upon the assumption of convergence of the sets of finite feature vectors to the sample space. There are two errors with this statement. First, I believe he means to say that in some way he can take countably increasing random samples of feature vectors, which are converging

to the test sample space. This convergence is described as being in the Hausdorff metric. This is the second error, since although the Hausdorff metric is calculated using the closure of each set relative to the other, there is no way one can accomplish this with even a countable collection of random variables. Additionallly, he relies upon 'balanced' samples of class 1 and class 2 objects of detection. This is unrealistic and unnecessary to the proof when using the Law of Large Numbers [5]. Therefore, Hausdorff measure is not a sufficient constraint for the theorem, the sets of random samples need to be constructed appropriately, and there is no need for these sets to strive for balance if they are truly random samples. Furthermore, there needs to be a proof showing convergence of ROC curves when the test sample spaces are a collection of countably increasing nested sample spaces with the population sample space as the union of the collection. Together these two proofs would demonstrate that as you increase the number of random samples from a test sample space, the conditional probabilities (and the ROC manifolds) converge to the expected values almost surely, and that as you nest your sample spaces in a countably infinite fashion, your conditional probabilities (and the ROC manifold) converges almost surely. This is important if you are going to use ROC manifolds from a test as a measure of performance. Thus, if we set up the test to reflect as accurately as possible the real world, and we take enough random samples, we can have confidence in using the ROC curve as a performance characteristic of families of classification systems participating in the same procedure.

Alsing begins his proof by showing that $\widehat{p}_{i|j}^{(n)}(\cdot)$ (my notation, not his) is a consistent estimator. He shows it is a consistent estimator of the mean. This is true due to the weak law of large numbers (in his proof he applies Chebyshev's inequality), so that for each $t \in \mathbf{T}$ we have that

$$\operatorname*{p\,lim}_{n\to\infty} \widehat{p}_{i|j}^{(n)}(t) = \pi_{i,j}, \tag{2.10}$$

for some mean value $_{i,j}$, where $\operatorname*{p\,lim}_{n\to\infty}$ denotes the limit in probability. Alsing does not characterize the values $\pi_{i,j}$ beyond this or $\mathbf{P}(\cdot) = \Big(p_{1|2}(\cdot), p_{1|1}(\cdot)\Big)$ (the ROC curve), and fails to connect the expected values of his random variables to the actual conditional prob-

abilities he's trying to prove his estimates to converge to. Moreover, his collection of metrics seems to require the measure space $(\mathbf{T}, \mathscr{T}, \mu)$ from which we draw the parameter t be such that $\mu(\mathbf{T}) < \infty$ (that is, a finite measure space). This is certainly not correct for $\mathbf{T} = \mathbb{R}$ and μ being Lebesgue measure. Even the simplest of toy problems has $\mathbf{T} = \mathbb{R}$, so then no positive, translation-invariant measure can be used to scale $\mathbf{T}$ down to a finite measured set.

Therefore, we offer a proof of convergence that characterizes better the nature of ROC convergence, and we extend the result to problems with classes greater than 2.

Theorem 1 (Extension of Alsing's ROC Convergence, Convergence of ROC manifolds). *Let $k \in \mathbb{N}$ be given, $m = k^2 - k$. Given denumerable, nested partitions of random samples, whose union is a sample population, the sequence of ROC $(m-1)$-manifolds, constructed from sample functions with parameter set $\boldsymbol{\Theta}$, a σ-finite measure space, converges to a ROC manifold.*

Proof: Let $(\Omega, \mathscr{B}, \mu)$ be a probability space and $(\boldsymbol{\Theta}, \mathscr{A}, \mu)$ be a σ-finite measure space of parameters. Let $\{\Omega_j\}_{j=1}^{k}$ be a partition of Ω into k classes. Let $O_n \in \mathscr{B}$ for each $n \in \mathbb{N}$. Let $O_n \uparrow \Omega$ as $n \to \infty$, *i.e.*, $O_1 \subseteq O_2 \subseteq \ldots O_n \subseteq \ldots \subseteq \Omega$ and $\bigcup_{n=1}^{\infty} O_n = \Omega$. For each n let $\{O_{n,j}\}_{j=1}^{k}$ be a partition of O_n into the k classes. We assume $O_{n,j} \neq \emptyset$ for each $n, j \in \mathbb{N}$ and that $O_{n,j} \uparrow \Omega_j$ as $n \to \infty$ for each $1 \leq j \leq k$. Let

$$\alpha = \min_{\substack{0<i<n \\ 1<j<k}} \{\mu^2(O_{i,j}), \mu(O_{i,j})\mu(O_j)\}.$$

Let $\mathbb{A}$ be a family of classification systems of Ω, so that for each parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $A_{\boldsymbol{\theta}} : \Omega \to \mathbf{O}$ defines a discrete, $\mathscr{B}$-measurable random variable.

Denote by $A_{\boldsymbol{\theta}}^{\natural}(k)$ the preimage of class k under $A_{\boldsymbol{\theta}}$. Let O_n be the sample space of the nth instantiation of data. Now fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k^2-k})$ and let $\Delta = \{\delta_1, \delta_2, \ldots\}$ be a discrete index set. Then for each $O_{n,j}$ we can construct a new probability space, $(O_{n,j}, \mathscr{B}_{n,j}, \mu_{n,j})$, where $\mathscr{B}_{n,j}$ is a σ-field on $O_{n,j}$, with $\mathscr{B}_{n,j} \subseteq \mathscr{B}$, and $\mu_{n,j}(B) = \frac{\mu(B)}{\mu(O_{n,j})}$ for each $B \in \mathscr{B}_{n,j}$. Let $C_{i,j,n} = A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j}$

Now, for each $\delta_r \in \Delta$, construct the random variable

$$X_{\delta_r}^{i|j}(\boldsymbol{\theta}, \cdot) = \mathrm{I}_{C_{i,j,n}}(\cdot),$$

where I is an indicator function. This random variable essentially tells us whether or not an $\omega \in O_{n,j}$ is classified as an error or not. Then the expected value of the random variable is

$$\begin{aligned}
\mathrm{E}\{X_{\delta_r}^{i|j}\} &= \mathrm{E}\{\mathrm{I}_{C_{i,j,n}}\} \\
&= \int_{O_{n,j}} \mathrm{I}_{C_{i,j,n}} d\mu_{n,j} \\
&= \frac{\mu(C_{i,j,n})}{\mu(O_{n,j})} \\
&= \mathrm{P}(A_{\boldsymbol{\theta}}^{\natural}(i)|O_{n,j}) \\
&= \widehat{p}_{i|j}(\theta).
\end{aligned}$$

Now let

$$\breve{p}_{i|j}^{(m)}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{r=1}^{m} X_{\delta_r}^{i|j}.$$

Since $\{X_{\delta_r}^{i|j}\}_{r=1}^{\infty}$ are independent identically distributed random variables, by the strong Law of Large Numbers [5], we have that

$$\breve{p}_{i|j}^{(m)}(\boldsymbol{\theta}) \to \mathrm{P}(A_{\boldsymbol{\theta}}^{\natural}(i)|O_{n,j}) = \widehat{p}_{i|j}^{(n)}(\boldsymbol{\theta}) \text{ almost surely } \mu \text{ as } m \to \infty.$$

Consider wlog that error $\breve{p}^{(m)}_{k-1|k}$ is the dependent variable with respect to the classification system. Then fixing the parameter $\boldsymbol{\theta} \in \Theta$ and letting

$$\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) = \big(\breve{p}^{(m)}_{2|1}((\boldsymbol{\theta}), \breve{p}^{(m)}_{3|1}(\boldsymbol{\theta}), \ldots, \breve{p}^{(m)}_{k|1}(\boldsymbol{\theta}), \ldots, \breve{p}^{(m)}_{1|k}(\boldsymbol{\theta}), \ldots, \breve{p}^{(m)}_{k-1|k}\big), \tag{2.11}$$

we have that the set of $(k^2 - k)$-vectors

$$\Big\{\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\Big\} \tag{2.12}$$

forms an estimate of the ROC manifold. We assume it is a proper ROC manifold, and that $(\boldsymbol{\Theta}, \mathscr{A}, \zeta)$ is a σ-finite measure space. Let

$$\boldsymbol{P}^{(n)}(\boldsymbol{\theta}) = \big(\widehat{p}^{(n)}_{2|1}(\boldsymbol{\theta}), \widehat{p}^{(n)}_{3|1}(\boldsymbol{\theta}), \ldots, \widehat{p}^{(n)}_{k|1}(\boldsymbol{\theta}), \ldots, \widehat{p}^{(n)}_{1|k}(\boldsymbol{\theta}), \ldots, \widehat{p}^{(n)}_{k-1|k}\big) \tag{2.13}$$

and now consider the product measure $\mu \times \zeta$. Thus, by Fubini's Theorem [44] we have that

$$\begin{aligned}
\lim_{m\to\infty} \int_{\boldsymbol{\Theta}\times\boldsymbol{\Omega}} \big|\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) - \boldsymbol{P}^{(n)}(\boldsymbol{\theta})\big| d(\mu \times \zeta) & \\
&= \lim_{m\to\infty} \int_{\boldsymbol{\Theta}} \left[\int_{\boldsymbol{\Omega}} \big|\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) - \boldsymbol{P}^{(n)}(\boldsymbol{\theta})\big| \, d\mu \right] d\zeta \\
&= \lim_{m\to\infty} \int_{\boldsymbol{\Omega}} \left[\int_{\boldsymbol{\Theta}} \big|\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) - \boldsymbol{P}^{(n)}(\boldsymbol{\theta})\big| \, d\zeta \right] d\mu \\
&= 0,
\end{aligned}$$

so that $\boldsymbol{P}^{(m)}(\boldsymbol{\theta}) \to \boldsymbol{P}^{(n)}(\boldsymbol{\theta})$ almost everywhere $\mu \times \zeta$. $\diamondsuit$

Next, we offer a continuation of the idea of convergence by now considerinig a ROC manifold convergence. This convergence is similar to the convergence of distribution functions with the exceptions that, 1) because ROCs are inherently connected to probability measures, any convergence can only be as strong as convergence almost everywhere,

and 2) the converging sequence of ROCs is constructed by building up smaller probability spaces into a universal one (universal with respect to **the** population).

Theorem 2 (ROC Convergence). *Let $k \in \mathbb{N}$ be given, $m = k^2 - k$. Given denumerable, nested partitions of random samples within denumerable, nested partitions of sample populations, whose union is **the** population, the sequence of ROC $(m-1)$-manifolds, constructed from sample functions with parameter set Θ, converges to the ROC manifold.*

Proof: Let the assumptions be the same and the estimates be the same as the results in Theorem 1. Let $\widehat{p}_{i|j}^{(n)}(\theta) = P(A_{\boldsymbol{\theta}}^{\natural}(i)|O_{n,j})$ be the estimate of the conditional probability, $p_{i|j}(\theta) = P(A_{\boldsymbol{\theta}}^{\natural}(i)|\Omega_j)$. Now consider the following two notes:

1. Since $O_{n,j} \uparrow O_j$, $\exists\ N_1(\theta) \in \mathbb{N}$ such that for $n \geq N_1$ we have that

$$|\mu(O_j) - \mu(O_{n,j})| < \frac{\varepsilon\alpha}{2\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) + 1}$$

   for each $i, 1 \leq i \leq k$ and each $j, 1 \leq j \leq k$.

2. Consider that $(A_{\boldsymbol{\theta}}^{\natural}(i) \cup O_{n,j}) \subseteq (A_{\boldsymbol{\theta}}^{\natural}(i) \cup O_j)$. Since

$$(A_{\boldsymbol{\theta}}^{\natural}(i) \cup O_{n,j}) = \mu(A_{\boldsymbol{\theta}}^{\natural}(i)) + \mu(O_{n,j}) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j})$$

   and

$$(A_{\boldsymbol{\theta}}^{\natural}(i) \cup O_j) = \mu(A_{\boldsymbol{\theta}}^{\natural}(i)) + \mu(O_j) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j),$$

   then by the monotonicity of μ we have that

$$\mu(A_{\boldsymbol{\theta}}^{\natural}(i)) + \mu(O_{n,j}) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j}) \leq \mu(A_{\boldsymbol{\theta}}^{\natural}(i)) + \mu(O_j) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j),$$

   so that

$$\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j}) \leq \mu(O_j) - \mu(O_{n,j}).$$

Thus, since the left side of the equation is non-negative, we have that

$$|\mu(A^{\natural}_{\boldsymbol{\theta}}(i) \cap O_j) - \mu(A^{\natural}_{\boldsymbol{\theta}}(i) \cap O_{n,j})| \leq |\mu(O_j) - \mu(O_{n,j})|.$$

Now $\exists\ N_2 \in \mathbb{N}$ such that for $n \leq N_2$ we have that

$$|\mu(O_j) - \mu(O_{n,j})| < \frac{\varepsilon\ \alpha}{2|\mu(O_j)|},$$

for all j. Thus,

$$|\mu(A^{\natural}_{\boldsymbol{\theta}}(i) \cap O_j) - \mu(A^{\natural}_{\boldsymbol{\theta}}(i) \cap O_{n,j})| < \frac{\varepsilon\ \alpha}{2|\mu(O_j)|},$$

for all j.

Now let $N = \max\{N_1, N_2\}$. Then for each θ, and $n \leq N$, we have that

$$\begin{aligned}
\left|p_{i|j}(\boldsymbol{\theta}) - \widehat{p}_{i|j}^{(n)}(\boldsymbol{\theta})\right| &= \left|\frac{\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j)}{\mu(O_j)} - \frac{\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j})}{\mu(O_{n,j})}\right| \\
&= \left|\frac{\mu(O_{n,j})\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) - \mu(O_j)\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j})}{\mu(O_j)\mu(O_{n,j})}\right| \\
&\leq \frac{\left|\mu(O_{n,j})\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) - \mu(O_j)\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j)\right|}{\left|\mu(O_j)\right|\left|\mu(O_{n,j})\right|} \\
&\quad + \frac{\left|\mu(O_j)\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) - \mu(O_j)\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j})\right|}{\left|\mu(O_j)\right|\left|\mu(O_{n,j})\right|} \\
&\leq \frac{\left|\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j)\right|\left|\mu(O_{n,j}) - \mu(O_j)\right|}{\alpha} \\
&\quad + \frac{\left|\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j) - \mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_{n,j})\right|\left|\mu(O_j)\right|}{\alpha} \\
&< \frac{\left|\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j)\right|\varepsilon\,\alpha}{2(\left|\mu(A_{\boldsymbol{\theta}}^{\natural}(i) \cap O_j)\right| + 1)\alpha} + \frac{\left|\mu(O_j)\right|\,\varepsilon\,\alpha}{2\left|\mu(O_j)\right|\alpha} \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \qquad (2.14)
\end{aligned}$$

This convergence occurs almost everywhere, $a.e.$, since it cannot be shown to occur over sets of μ-measure zero. This is equivalent to almost sure, $a.s.$, convergence and convergence with probability 1 (also known as convergence in law), since we are using probability measure μ. Recall that class $p_{k-1|k}$ is the dependent class conditional probability with regard to the classification system. Let

$$\boldsymbol{Q}^{(n)}(\boldsymbol{\theta}) = \left(\widehat{p}_{2|1}^{(n)}(\boldsymbol{\theta}), \widehat{p}_{3|1}^{(n)}(\boldsymbol{\theta}), \ldots, \widehat{p}_{k|2}^{(n)}(\boldsymbol{\theta}), \ldots, \widehat{p}_{1|k}^{(n)}(\boldsymbol{\theta}), \ldots, \widehat{p}_{k-2|k}^{(n)}(\boldsymbol{\theta})\right) \qquad (2.15)$$

Then the set of $k^2 - k - 1$-vectors over all $\boldsymbol{\Theta}$,

$$\Big\{\boldsymbol{Q}^{(n)}(\boldsymbol{\theta}) \,:\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\Big\}, \tag{2.16}$$

defines the nth ROC estimate of the ROC manifold. We assume it is a proper ROC manifold. Thus for each $n \in \mathbb{N}$ there exists a continuous real-valued function in k^2-k-1 variables. Let

$$g_n\big(\boldsymbol{Q}^{(n)}(\boldsymbol{\theta})\big) = \widehat{p}^{\,n}_{k-1|k}(\boldsymbol{\theta})$$

be such a function. Let

$$\boldsymbol{Q}(\boldsymbol{\theta}) = \big(p_{2|1}(\boldsymbol{\theta}, p_{3|1}(\boldsymbol{\theta}), \ldots, p_{k|2}(\boldsymbol{\theta}), \ldots, p_{1|k}(\boldsymbol{\theta}), \ldots, p_{k-2|k}(\boldsymbol{\theta})\big) \tag{2.17}$$

for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and set

$$g\big(\boldsymbol{Q}(\boldsymbol{\theta})\big) = p_{k-1|k}(\boldsymbol{\theta})$$

It is clear from Theorem 1, that:

1. g_n is continuous on $\boldsymbol{\Theta}$;
2. $\Big|g_n\big(\boldsymbol{Q}^{(n)}(\boldsymbol{\theta})\big)\Big| \leq 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; and
3. $g_n\big(\boldsymbol{Q}^{(n)}(\boldsymbol{\theta})\big) \to g\big(\boldsymbol{Q}(\boldsymbol{\theta})\big)$ $a.e.$ for fixed $\boldsymbol{\theta}$.

Then for $\varepsilon > 0$ given, let $\mathrm{B}(\boldsymbol{\Theta}; \varepsilon)$ be an open ε-ball in $\boldsymbol{\Theta}$. Thus, by the Dominated Convergence Theorem, we have that

$$\lim_{n\to\infty} \int_{\mathrm{B}(\boldsymbol{\Theta};\varepsilon)} \Big|g_n - g\Big| d\zeta = 0. \tag{2.18}$$

so that $\lim_{n\to\infty} g_n = g$ $a.e.$ on $\mathrm{B}(\boldsymbol{\Theta}; \varepsilon)$. This convergence is uniform $a.e.$ over compact subsets of $\boldsymbol{\Theta}$. $\diamondsuit$

III. A Category Theory Description of Fusion

3.1 Probabilistic Construction of the Event-Label Model

Let $\mathscr{C}$ be a complex of conditions [28] for a repeatable experiment, and let Ω be a set of outcomes of this experiment with $\mathrm{T} \subset \mathbb{R}$ being a bounded interval of time. Interval T sorts Ω such that we call $\mathrm{E} \subseteq \Omega \times \mathrm{T}$ an **event-state**. An event-state is then comprised of event-state elements, $e = (\omega, t) \in \mathrm{E}$, where $\omega \in \Omega$ and $t \in \mathrm{T}$. Thus e denotes a state ω at an instant of time t. Let $\Omega \times \mathrm{T}$, be the set of all event-states for an event over time interval T. Let $\mathscr{E}$ be a σ-field on $\Omega \times \mathrm{T}$, and μ be a probability measure defined on the measurable space $(\Omega \times \mathrm{T}, \mathscr{E}, \mu)$. Then the triple $(\Omega \times \mathrm{T}, \mathscr{E}, \mu)$ forms a probability space [5].

The design of a classification system involves the ability to detect (or sense) the occurrence of an event in Ω, and process the event into a label of set L. For example, design a system that detects airborne objects and classifies them friendly or unfriendly. To do this a classification system relies on several mappings, which are composed, to provide the user an answer (from the event, to the label). Since $\mathscr{E}$ is a σ-field on $\Omega \times \mathrm{T}$, then let $\mathrm{E} \in \mathscr{E}$ be any member of $\mathscr{E}$. Then a sensor, s, is defined as a mapping from E into a (raw) data set D. We denote this with the diagram

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

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

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

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

diagram

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

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

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

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

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

Thus, A is an L-valued random variable which maps members $\mathrm{E} \in \mathscr{E}$ into the label set L and is diagrammed by

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

Consider the simple model of a multi-sensor system using two sensors in Figure 3.1. The sets E_i, for $i \in \{1, 2\}$, are sets of event-states. The label set L_i can be as simple as

$$\mathrm{E}_1 \xrightarrow{s_1} \mathrm{D}_1 \xrightarrow{p_1} \mathrm{F}_1 \xrightarrow{c_1} \mathrm{L}_1$$

$$\mathrm{E}_2 \xrightarrow{s_2} \mathrm{D}_2 \xrightarrow{p_2} \mathrm{F}_2 \xrightarrow{c_2} \mathrm{L}_2$$

Figure 3.1: Simple Model of a Dual-Sensor System.

the two-class set {target, non-target} or could have a more complex structure to it, such as the *types* of targets and non-targets, paired with a ranking of measure, for example [56], in order to define the battlefield more clearly for the warfighter. Now the diagram in Figure 3.1 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. Figure 3.1 models two sensors with differing fields of view. Performing fusion along any node or edge in this graph could possibly result in an elevated level of fusion [15]–that of situa-

tion refinement or threat refinement, since we are not fusing common information about a particular event or events, but we may be fusing situations.

There are at least two other possible scenarios that Figure 3.1 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 classification 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, $\mathrm{E} = \mathrm{E}_1 \cap \mathrm{E}_2$. Hence, there are two fundamental 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). This is commonly referred to as Level 1 fusion (or Level 0 fusion) Object Assessment. 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. These are the higher Levels 2 and 3, Situation Assessment and Impact Assessment. We distinguish between the two types of fusion scenarios discussed by calling them **event-state fusion** and **situation-state fusion** respectively. Therefore, Figure 3.2 represents an Event-State-to-Label model of a dual sensor process. The only

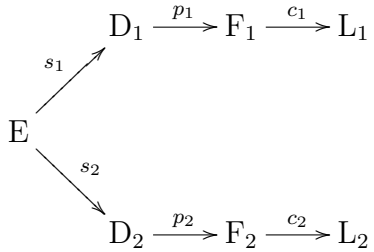


Figure 3.2: Two Classification Systems with Overlapping Fields of View.

restriction necessary for the usefulness of this model is that a common field of view, E, be used. Consequently, D_1 and D_2 could actually be the same data set under the model, while s_1 and s_2 could be different sensors. We will refer to a finite number of families of

classification systems, such as the two in Figure 3.2, which we wish to explore the fusion of, as a fixed classification category. For $\mathscr{E}$ considered as a category of sets, and a fixed label set L, we note that $\mathrm{L}^{\mathscr{E}}$, is the functor category of all such classification systems, so that our fixed classification category is a subcategory of $\mathrm{L}^{\mathscr{E}}$. Each classification system or set of sample functions comprises a fixed branch of $\mathrm{L}^{\mathscr{E}}$ (*i.e.*, a functor or a family of functors). Equally true is the fact that if we want to compete classification systems, we must test them over the same sample space as well. Therefore, we choose the functor category L^{E}, with a fixed L and a fixed E, to compete the classification systems over. Our convergence theorems allow us to treat E as if it were the sample population, with the caveat that our test then is only as good as it is representative of the operational circumstances of the real-world population.

It is also important to note that when we want to fuse classification systems (or families of classification systems heretofore denoted as sample functions), we must be fusing systems which are originally yielding values from the same label set L, and not just the same set up to isomorphism. We will later show that there are two kinds of fusion with regard to these label sets, but for right now, we consider fusing only those branches which produce values in the same exact set. Additional considerations and techniques must be used to fuse across different label sets.

3.2 Construction of a family of classification systems

3.2.1 Single Parameter. Now suppose we have a parameter $\theta \in \Theta$, which is possibly multidimensional. Then it is common that there is a family, $\{c_\theta : \theta \in \Theta\}$, of classifiers so that for each $\theta \in \Theta$, each composition,

$$c_\theta \circ p \circ s$$

describes an event-state model on fixed $\mathrm{E} \in \mathscr{E}$, and fixed sets D, F, and L. The corresponding family

$$\mathbb{A} = \{A_\theta \mid \theta \in \Theta\},$$

where $A_\theta = c_\theta \circ p \circ s$, is a family of classification systems. Thus, Θ acts as an indexing set for defining $\mathbb{A}$ which also could be thought of as a collection of sample functions or sample sequences depending on whether or not the parameter set is countable.

3.2.2 Multiple Parameter . One can extend the ideas in Section 3.2.1 to include other index sets Γ and Δ, so that the composition

$$c_\theta \circ p_\delta \circ s_\gamma,$$

where $\theta \in \Theta, \delta \in \Delta, \gamma \in \Gamma$, is a classifier, $A_{(\theta,\delta,\gamma)}$. In this case, we must look at the triple (θ, δ, γ) as the parameters for the ROC manifold. If we have a two-class label set, then this presents us with the case of degeneracies. For example, suppose we calculate the optimal point on the ROC 1-manifold. Then we have three parameters representing each point on the curve, so that there may be multiple triples which optimize, none better than the others. This fact alone may make it difficult to calculate an optimal triple, since no inverse function mapping ROC points on the curve to the product space $\Theta \times \Delta \times \Gamma$ exists. To eliminate degeneracies, given k classes, we require $k^2 - k - 1 = m - 1$ parameters. Any more than this yields such degeneracies, while any fewer results in either a smaller dimensional ROC manifold, or a set of ROCs which is not a manifold and possibly a suboptimal choice of operating parameters (suboptimal with respect to a ROC $(m-1)$-manifold).

3.3 Defining Fusion Rules from the Event-Label Model

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

Definition 26 (Fusion Rule of n Fixed Branches of Families of Classification Systems). Let $\mathfrak{S}_n$ be a fixed classification category with n branches. For each $i = 1, \ldots, n$, let $\mathcal{O}_i \in \mathbf{CAT}$ be a small category of data corresponding to the ith branch's source of data to

be fused (this could be raw data, features, or labels). Then the product

$$\pi(n) = \prod_{i=1}^{n} \mathcal{O}_i$$

is a product category. For any particular category of data, $\mathcal{O}_0$, the exponential, $\mathcal{O}_0^{\pi(n)}$, 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 $\mathbf{Ob}(\mathcal{O}_0)$, and maps data arrows in $\mathbf{Ar}(\pi(n))$ to arrows in $\mathbf{Ar}(\mathcal{O}_0)$. These fusion rules are functors, $\mathfrak{R}$, which make up the objects of the category. The arrows of the functor category are all the natural transformations between them. We designate $\boldsymbol{FR}_{\mathbb{O}_n}(\mathcal{O}_0)$ to be this functor category of fusion rules.

If the $\mathcal{O}_i$ are categories generated from sensor sources (*i.e.*, outputs), then we call $\mathcal{O}_0^{\pi(n)_1}$ a category of data-fusion rules and use the symbols $\mathcal{D}_0^{\pi(n)_1}$. The fusion rule branch would then be diagrammed like this:

$$E \xrightarrow{<s_1,s_2,\ldots,s_n>} \pi(n)_1 \xrightarrow{r} \mathrm{D}_0 \xrightarrow{p} \mathrm{F} \xrightarrow{c_\phi} \mathrm{L}, \tag{3.1}$$

where D_0 is the receiving category, r is the fusion rule, and $< s_1, s_2, \ldots, s_n >$ is the unique arrow generated by the product $\pi(n)_1$. If the categories are generated by processor sources, then call $\mathcal{O}_0^{\pi(n)_2}$ a category of feature-fusion rules and use the symbols $\mathcal{F}_0^{\pi(n)_2}$. fusion rule branch would then be diagrammed like this:

$$E \xrightarrow{<s_1,s_2,\ldots,s_n>} \pi(n)_1 \xrightarrow{<p_1,p_2,\ldots,p_n>} \pi(n)_2 \xrightarrow{r} \mathrm{F}_0 \xrightarrow{c_\phi} \mathrm{L}, \tag{3.2}$$

where $\pi(n)_1$ is the first product of data categories, $\pi(n)_2$ is the second product of feature categories, r is again the fusion rule, and $< p_1, p_2, \ldots, p_n >$ is the unique arrow generated by the product $\pi(n)_2$. Finally, if they have classifiers as sources, then call them label-fusion rules (or, alternatively, decision-fusion rules) and use the symbols $\mathcal{L}_0^{\pi(n)_3}$. This

fusion rule branch would be diagrammed like this:

$$E \xrightarrow{<s_1,s_2,...,s_n>} \pi(n)_1 \xrightarrow{<p_1,p_2,...,p_n>} \pi(n)_2 \xrightarrow{<c_1,c_2,...,c_n>} \pi(n)_3 \xrightarrow{r_\phi} \mathrm{L}, \tag{3.3}$$

where r_ϕ is a fusion rule for each parameter (in order to generate an appropriate family of classification systems), and $< c_1, c_2, \ldots, c_n >$ is the unique arrow generated by the product $\pi(n)_3$. (We removed the parameters from the classifiers and replaced them with a single, possibly vector valued, parameter on the fusion rule).

A fusion rule could be a Boolean rule, a filter, an estimator, or an algorithm. Notice that our definition of fusion rule does not include a qualitative component; there is no necessary condition of "betterness" for a fusion rule. The result of applying a fusion rule to an existing set of fundamental branches could result in output considerably worse than existed previously. This does not affect the definition. First we define fusion rules as the key component of the fusion process. Next, we pare down the category to a subcategory which does include a qualitative component, with one suggested way of accomplishing this. We now desire to show how defining a fusor (see Definition 30) as a fusion rule with a constraint changes the Event-State model into an Event-State Fusion model. Continuing to consider the two families of classification systems in Figure 3.2, it is evident that a fusion rule can be designed which would apply to either the data sets, the feature sets, or the label sets (though special care needs to be taken with this case, when the actual labels are not the same). Given a fusion rule $\mathfrak{R}$ for the two data sets as in Figure 3.2, our model becomes that of Figure 3.3. 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

$$\mathrm{L} = \mathrm{L}_1 = \mathrm{L}_2 = \{\mathrm{Target}, \mathrm{Nontarget}\},$$

where the targets and non-targets are well-defined across classification systems (*i.e.*, each classification is identifying targets that satisfy the same definition of what a target is). In

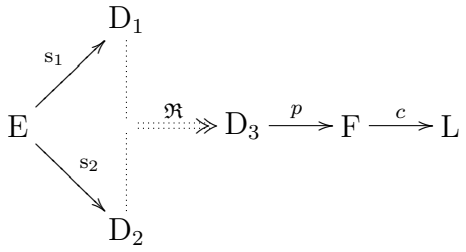


Figure 3.3: Fusion Rule Applied on Data Categories from Two Fixed Branches.

a **within**- fusion scenario (see Definition 34 as opposed to Definition 35), the data sets (or feature sets) are the same, $\mathrm{D}_1 = \mathrm{D}_2 = \mathrm{D}_3$. This is true in the case that the sensors used are the same type (that is, they collect the same types of measurements, but from possibly different locations relative to the overlapping field of view). In the case where the data sets (or feature sets) are truly different, a composite data set (and/or feature 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 functor.

Now at this point we may consider, in what way is the process modeled in Figure 3.3 *superior* to the original processes shown in Figure 3.2 when $\mathrm{L} = \mathrm{L}_1 = \mathrm{L}_2$ (we will deal with the case $\mathrm{L}_1 \neq \mathrm{L}_2$ later)? One way of comparing performance in such systems is to compare the processes' receiver operating characteristics (ROC) curves, which we will do in the Chapter IV.

3.4 Fusion Rules

3.4.1 Object-Fusion. There are, of course, multiple descriptions in the literature to "types" of fusion. There is *data*-fusion, *feature*-fusion, and *decision*-fusion. There is data in-feature out fusion [8] and many more. We would like to codify what should be meant by these expressions by introducing, in its most basic form, a vernacular for fusion which is intuitive, yet has its definition rooted in mathematics. We start by assuming we have a finite number of objects we wish to fuse together. What does the finite set of fusion rules look like? How can we describe in an observational way what is going on? Once

the definition of fusion is established, we can move on to labeling types of fusion under certain model asumptions.

Definition 27 (Object-Fusion Category). Let $\{\mathcal{O}_i \mid i \in \{1, \ldots, m\}\}$ be a finite sequence of non-empty categories (possibly discrete). Then

$$\prod_{i=1}^{m} \mathcal{O}_i$$

defines a product category (see Definitions 24 and 26). Let

$$\pi(m) = \prod_{i=1}^{m} \mathcal{O}_i$$

for fixed $m \in \mathbb{N}$. Then for a fixed category $\mathcal{O}$, we have that

$$\mathbf{FR}_{\pi(m)}(\mathcal{O}) = \mathcal{O}^{\pi(m)}$$

is a functor category. The functor category $\mathbf{FR}_{\pi(m)}(\mathcal{O})$ is called an $\pi(m)$-Fusion category relative to $\mathcal{O}$ to denote the functors are fusing m $\mathcal{O}_i$-objects, and as necessary, their accompanying arrows into a single object and arrow in $\mathcal{O}$. When the relationship of all the $\mathcal{O}_i$ objects can be made clear, by simply calling them "objects", then we call $\mathbf{FR}_{\pi(m)}(\mathcal{O})$ the Object-Fusion category relative to $\mathcal{O}$ (regardless of the value of m).

It's important to note in our definition of fusion rules we did not put forward the notion of defining fusion rules in terms of performance. We will need a second mathematical definition later to narrow the category of fusion rules down to a subcategory of fusion rules, which can be ordered according to their performance in some manner. First we'll consider further delineating the types of fusion rules within the Event-State model.

3.4.2 Types of Fusion Rules. We consider digraph G, as depicted in Figure 3.4, consisting of sample functions which are compositions of random variables. E is an event in the $\sigma-\mathrm{field}$, $\mathscr{E}$. The sets D_1 and D_2 are objects of a finite collection of categories

of data sets, while the sets F_1 and F_2 are objects of a finite collection of categories of feature sets. The label sets L_1 and L_2 are the objects of a finite collection of categories of label sets (and we still require that $\mathrm{L}_1 = \mathrm{L}_2$). Figure 3.5 shows the nodes in digraph

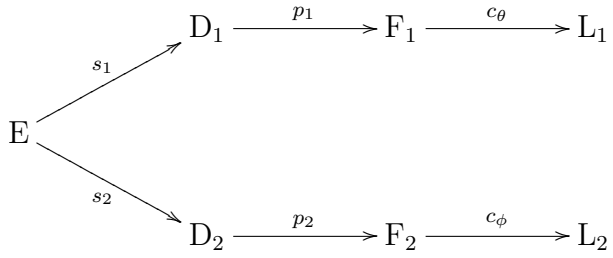


Figure 3.4: Digraph G.

G along which fusion rules can be applied. With the use of category theory, we can also

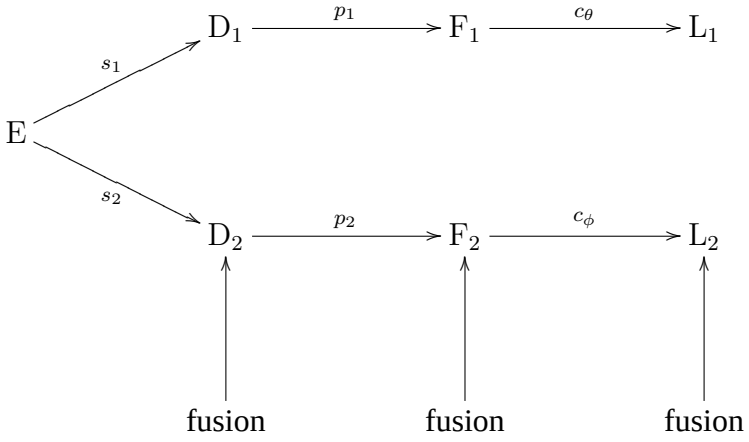


Figure 3.5: Known Fusion Rule Nodes of Digraph G.

describe that there should theoretically be nodes along the arrows of digraph G for fusion rules as well, though we have no example at this time of a rule or algorithm that does this without using the pointwise outputs of the arrows. Figure 3.6 shows all available fusion rule nodes applicable (at least theoretically) to the event-state decision model. This leads to a theorem regarding the types of fusion available under the model.

Theorem 3 (Six Categories of Object-Fusion under digraph G). *Let G be a digraph with an initial vertex and n branches with k vertices to each branch, so that there are*

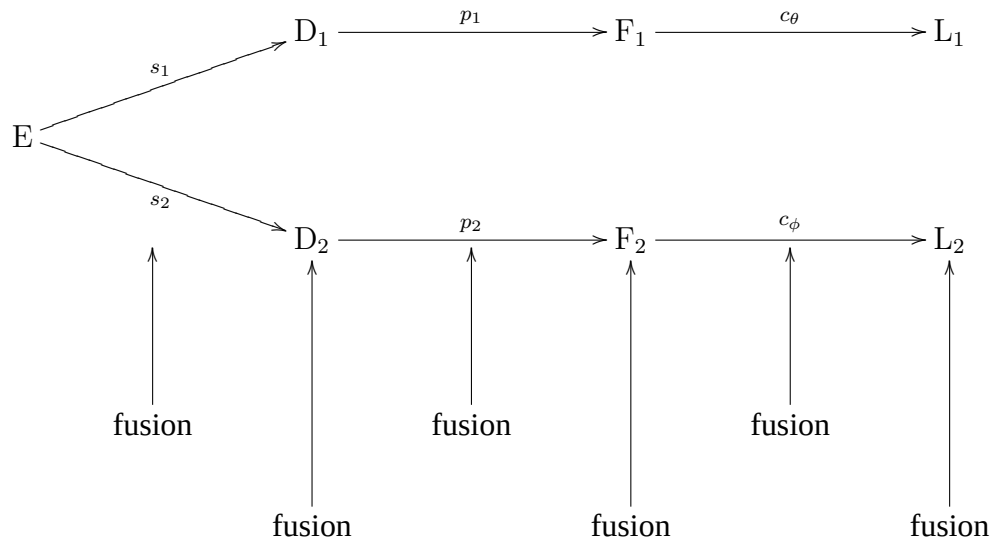


Figure 3.6: Theoretical Fusion Rule Nodes of Digraph G.

$nk-(n-1) = n(k-1)+1$ *total vertices and* $n(k-1)$ *edges. Then there exists* $2(k-1)$ *categories of Object-Fusion that can be performed on any event-state decision model that* **G** *represents.*

Proof: Excluding the event E, there are an equal number of edges and vertices to each branch. The initial vertex represents the event set while the composition of arrows (edges) along each branch represent the classification system. Fusion rules are objects within functor categories, so that if we label the non-initial vertices matrix style with rows representing branches and k columns representing the vertices:

$$\begin{array}{cccc} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & \vdots \\ \vdots & \cdot & \cdot & \vdots \\ v_{n1} & \cdots & \cdots & v_{nk} \end{array} \tag{3.4}$$

The components of column j can be considered as being i categories from a finite subcategory of the category $\mathbf{CAT}$. Suppose that in column j we have categories $\mathcal{O}_i$, $i = 1, 2, \ldots, n$. Then

$$\pi(i)_j = \prod_{i=1}^{n} \mathcal{O}_i$$

is a product category for the jth column. Let $\mathcal{O}$ be any category. Then the functor category $\mathbf{FR}_{\pi(i)_j}(\mathcal{O}) = \mathcal{O}^{\pi(i)_j}$ is the $\pi(i)_j$-Fusion category relative to $\mathcal{O}$. Furthermore, in addition to labeling the non-initial vertices as matrix components, we can create a matrix from the edges in the same manner, and without loss of generality, the result is k more fusion categories, so that the total number of fusion categories is $2(k-1)$. $\diamondsuit$

When the number of vertices per branch, $k = 4$, as in digraph G_0 (see Figure 3.4), then we have six ($2 \cdot (4-1) = 6$) categories of Object-Fusion. Adopting the labeling scheme used by our model, we can label each category's "objects" as Sensor-, Data-, Processor-, Feature-, Classifier-, or Label- (or Decision-)Fusion.

3.4.3 Comparison of Desarathys paradigm with Fusion Categories. The chart in Table 3.1 shows the relationship between these categories and Desarathy's breakdown of the types of fusion.

Desarathy's I/O taxonomy	Category Theory Approach
No taxonomy	Sensor-Fusion
Data In-Data Out	Data-Fusion
Data In-Feature Out	Processor-Fusion or Data-Fusion
Feature In-Feature Out	Feature-Fusion
Feature In-Decision Out	Classifier-Fusion or Feature-Fusion
Decision In-Decision Out	Label-Fusion (also called Decision-Fusion)

Table 3.1: Desarathy's I/O Fusion categorization from [15] .

Definition 28 (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 well-defined label set.

Suppose we have a fixed classification category L^{E}, and let A be an object in this category. Then for L consisting of k labels, there exists a vector in $(n = k^2 - k)$-ROC space described by an n-vector v_A, where

$$v_A = (p_{2|1}(A), \ldots, p_{k|1}(A), \ldots, p_{k-1|k}(A)).$$

The proof is self-evident since E is a sample space. We call this vector the **operating characteristic** vector, and we let

$$V = \left\{ v_A \mid A \in \mathbf{Ob}(\mathrm{L}^{\mathrm{E}}) \right\} \tag{3.5}$$

and

$$\mathcal{V} = \mathbf{OC}_{\mathrm{L}^{\mathrm{E}}} = \left(\mathscr{P}(V), \mathbf{Ar}(V), \mathbf{Id}(V), \circ \right), \tag{3.6}$$

where $\mathscr{P}(V)$ is the power set of V. The category $\mathbf{OC}_{\mathrm{L}^{\mathrm{E}}}$ is the category of operating characteristic families with undetermined non-identity arrows (we will determine them presently). Now, consider the category

$$\mathcal{C} = (\mathscr{P}(\mathbf{Ob}(\mathrm{L}^{\mathrm{E}})), \mathbf{Id}(\mathrm{L}^{\mathrm{E}}), \mathbf{Id}(\mathrm{L}^{\mathrm{E}}), \circ)$$

whose objects are sets of classification systems. Then $\mathbb{A} \in \mathbf{Ob}(\mathcal{C})$ for each family of classification systems $\mathbb{A}$. Let

$$\mathfrak{F} : \mathcal{C} \longrightarrow \mathcal{V} \tag{3.7}$$

be an operating characteristic functor, which maps power sets of classification systems to the set of operating characteristics associated with them. Let

$$\xi : \mathcal{V} \longrightarrow \mathcal{P} \tag{3.8}$$

be a functor where $\mathcal{P}$ is a poset, thought of as a category induced by a partial order, $\geq$, of its elements. Then ξ is a functor taking objects consisting of sets of operating characteristics into a value of $\mathcal{P}$. We do not need to define the rule at this point. Let $\mathbb{A}_0, \mathbb{A}_1 \in \mathcal{C}$, such that

$$\mathfrak{F}(\mathbb{A}_0) = f_{\mathbb{A}_0}$$

and

$$\mathfrak{F}(\mathbb{A}_1) = f_{\mathbb{A}_1}$$

where the outputs are families of operating characteristics. Then the diagram

$$\begin{array}{ccc} f_{\mathbb{A}_0} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_0}) = p_0 \\ {\scriptstyle g}\,\vdots & & \downarrow{\scriptstyle |\vee} \\ f_{\mathbb{A}_1} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_1}) = p_1 \end{array}$$

where $p_0, p_1 \in \mathcal{P}$, commutes for some unique (up to isomorphism) g. This g is an induced partial order on $\mathcal{V}$. Thus, for every pair of families of classification systems, $\mathbb{A}_0, \mathbb{A}_1 \in \mathcal{C}$, we have that the rectangle

$$\begin{array}{ccccc} \mathbb{A}_0 & \xrightarrow{\mathfrak{F}} & \mathfrak{F}(\mathbb{A}_0) = f_{\mathbb{A}_0} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_0}) = p_0 \\ {\scriptstyle |\curlyvee}\,\vdots & & {\scriptstyle g}\,\vdots & & \downarrow{\scriptstyle |\vee} \\ \mathbb{A}_1 & \xrightarrow{\mathfrak{F}} & \mathfrak{F}(\mathbb{A}_1) = f_{\mathbb{A}_1} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_1}) = p_1 \end{array} \tag{3.9}$$

commutes when we impose the criterion $\mathbb{A}_0 \succeq \mathbb{A}_1$ iff $(\xi \circ \mathfrak{F})(\mathbb{A}_0) \geq (\xi \circ \mathfrak{F})(\mathbb{A}_1)$, so that the functor $\xi \circ \mathfrak{F}$ is a natural transformation. It is precisely the arrows like g, which make

such rectangles commute, that belong in the category $\mathcal{V}$. It is also the arrows induced from the partial order $\succeq$, which provide unique maps from one classification family to another, which will allow us to define the fusion process in Section IV.

IV. An Optimization for Competing Fusion Rules

4.1 Bayes Optimal Threshold (BOT) in a family of classification systems

4.1.1 Two-class BOT. Let $(\mathrm{E}, \mathscr{E}, \mu)$ be a probability space and $\mathbb{A}$ be a family of sample functions (classification systems) with parameter space Θ. Let

$$\{\mathrm{E}_1, \mathrm{E}_2 \,:\, \mathrm{E}_1, \mathrm{E}_2 \in \mathscr{E}\}$$

be a partition of E, and $\mathrm{L} = \{\ell_1, \ell_2\}$. 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, if $A_{\theta^*} \in \mathbb{A}$ with $\theta^* \in \Theta$ minimizes the quantity

$$\begin{aligned} \mu\big((A_\theta^\natural(\ell_1) \cap \mathrm{E}_2) \cup (A_\theta^\natural(\ell_2) \cap \mathrm{E}_1)\big) &= \mu(A_\theta^\natural(\ell_1) \cap \mathrm{E}_2) + \mu(A_\theta^\natural(\ell_2) \cap \mathrm{E}_1) \\ &= p_{1|2}(A_\theta)\mu(\mathrm{E}_2) + p_{2|1}(A_\theta)\mu(\mathrm{E}_1), \end{aligned} \tag{4.1}$$

where $\mu(\mathrm{E}_1)$ and $\mu(\mathrm{E}_2)$ are the prior probabilities of class 1 and class 2, respectively. Then θ^* is the BOT for the family of classification systems $\mathbb{A}$.

4.1.2 N-class BOT. Now let us keep the assumptions of the previous section with the exception that we now have k classes to consider, and

$$\{\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k \,:\, \mathrm{E}_i \in \mathscr{E} \;\forall\; i = 1, 2, \ldots, k\}$$

is a new partition of E into k classes, with $\mathrm{L} = \{\ell_1, \ell_2, \ldots, \ell_k\}$ a label set corresponding to the partition of classes. Then the corresponding Bayes Optimal Threshold, $\theta^* \in \Theta$, where Θ is now $k-1$ dimensional would be the parameter which minimizes

$$B_{err} = \sum_{i=1}^{k} \sum_{j=1}^{k} (1 - \delta_{i,j}) p_{i|j}(A_\theta)\mu(\mathrm{E}_j), \tag{4.2}$$

where

$$\delta_{i,j} = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases}$$

.

4.2 An Optimization over ROC m-manifolds for competing fusion rules

4.2.1 ROC m-manifold optimization. The method used in this section applies and extends that of [31]. Let $k \in \mathbb{N}$, $k \geq 1$ be given, with $m = k^2 - k$. Let

$$x_{m+1} = f(x_1, x_2, \ldots, x_m)$$

be the equation of the ROC m-manifold. Then define

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

Let $\mathfrak{M} = \{(x_1, x_2, \ldots, x_{m+1}) : \Psi(x_1, x_2, \ldots, x_{m+1}) = 0\}$ be the ROC m-manifold. Assume $\mathbf{R}(0) = (0, 0, \ldots, 0, 0)$. Then there is $t_f \in [0, 1]$ such that $\mathbf{R}(t_f) \in \mathfrak{M}$, with t_f dependent upon the particular $\mathbf{R}$. We assume all first-order partial derivatives exist and are continuous for Ψ. For each $t \in [0, 1]$ let $\mathbf{R}(t) = (X_1(t), X_2(t), \ldots, X_{m+1}(t))$ be a smooth trajectory that starts at the initial point $(0, 0, \ldots, 0, 1)$ and terminates on the manifold $\mathfrak{M}$. Choose weights $a_i > 0$ for $i = 1, 2, \ldots, m + 1$ such that $\sum_{i=1}^{m+1} a_i = 1$, and let $\| \cdot \|_{\mathrm{W}}$ represent the weighted $\ell_1(\mathbb{R}^{m+1})$ norm defined on $\mathbf{V} = (v_1, v_2, \ldots, v_{m+1})$ by

$$\|\mathbf{V}\|_{\mathrm{W}} = \sum_{i=1}^{m+1} a_i \mid v_i \mid . \tag{4.3}$$

Define the functional J

$$J[\mathbf{R}] = \int_0^{t_f} \|\dot{\mathbf{R}}(t)\|_{\mathrm{W}} \, dt. \tag{4.4}$$

Theorem 4 (Thorsen-Oxley). *Given k classes, and a ROC m-manifold, where $m+1 = k^2 - k$ is the number of possible types of errors in classification, and given weights $a_i = c_i\alpha_i$, a cost times a prior probability (where $\sum_{i=1}^{m+1} \alpha_i = 1$), then the Bayes Optimal Threshold corresponds to the point, $\mathbf{p}$, on the ROC m-manifold, f, where*

$$\nabla f(\mathbf{p}) = \frac{-1}{a_{m+1}}\big(a_1, a_2, \ldots, a_{m+1}\big). \tag{4.5}$$

Proof: For ease of notation, define

$$G(t, \mathbf{R}(t), \dot{\mathbf{R}}(t)) = \|\dot{\mathbf{R}}(t)\|_{\mathrm{W}}$$

and let $Y_i(t) = \dot{X}_i(t)$ for each i. Hence we write Equation 4.4 as

$$J[\mathbf{R}] = \int_0^{t_f} G\, dt \tag{4.6}$$

and we will suppress the integrand variables. We would like to minimize J, so let's find $\mathbf{R}(t)$ with initial and terminal points as discussed which minimizes the functional.

Let $\alpha \in [-\beta, \beta]$ where $\beta \in \mathbb{R}, \beta > 0$ be a family of real parameters. Let

$$\{\mathbf{R}(t, \alpha) = (X_1(t, \alpha), X_2(t, \alpha), \ldots, X_{m+1}(t, \alpha)) \,:\, \alpha \in [-\beta, \beta]\} \tag{4.7}$$

be a family of one-parameter trajectories which contains the optimal curve $\mathbf{R}^*(t)$. Furthermore we assume that at $\alpha = 0$ $\mathbf{R}(t, 0) = \mathbf{R}^*(t)$. Let $\mathbf{R}(t_f, \alpha) \in \mathfrak{M}$. By the Implicit Function Theorem, there is 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 $T_f(0) = t_f^*$. Assume $\mathbf{R}^*(t)$ minimizes J, then a necessary optimality condition is that the first variation of

$$J[\mathbf{R}(\cdot, \alpha)] = \int_0^{T_f(\alpha)} G\, dt \tag{4.8}$$

be equal to zero at $\alpha = 0$. That is,

$$\frac{d}{d\alpha} J[\mathbf{R}(\cdot, \alpha)]_{\alpha=0} = 0. \tag{4.9}$$

We use the notation

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

for brevity. Applying Leibniz's rule to the derivative of Equation 4.8 we get

$$\delta J[\mathbf{R}^*] = G^*|_{t=t_f^*} \, \delta T_f + \int_0^{t_f^*} (\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} + \nabla_{\mathbf{y}} G^* \cdot \delta\dot{\mathbf{R}}) dt. \tag{4.10}$$

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

$$\delta J[\mathbf{R}^*] = G^*|_{t=t_f^*} \, \delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_0^{t_f^*} + \int_0^{t_f^*} (\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} - \frac{d}{dt} \nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}) dt. \tag{4.11}$$

At $\alpha = 0$ we have the necessary optimality condition

$$\delta J[\mathbf{R}^*] = G^*|_{t=t_f^*} \, \delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_{t=t_f^*} + \int_0^{t_f^*} (\nabla_{\mathbf{x}} G^* \cdot \delta\mathbf{R} - \frac{d}{dt} \nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}) dt = 0. \tag{4.12}$$

Since this must be true over all admissible variations, we have the Euler Equations

$$\nabla_{\mathbf{x}} G^* - \frac{d}{dt} \nabla_{\mathbf{y}} G^* = \mathbf{0}. \tag{4.13}$$

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

$$G^*|_{t=t_f^*} \, \delta T_f + [\nabla_{\mathbf{y}} G^* \cdot \delta\mathbf{R}]_{t=t_f^*} = 0. \tag{4.14}$$

Solving the Euler Equation 4.13, we have $\nabla_{\mathbf{x}} G^* = \mathbf{0}$, which implies

$$\frac{d}{dt} \nabla_{\mathbf{y}} G^* = \mathbf{0}, \tag{4.15}$$

hence,

$$\frac{d}{dt}\,\mathbf{sgn}(Y_i^*(t)) = 0 \tag{4.16}$$

for $i = 1, 2, \ldots, m+1$, where $\mathbf{sgn}(Z)$ returns the value of $-1, 0,$ or 1, depending on the sign of the function Z. Thus, for each $i = 1, 2, \ldots, m+1$, we have

$$\mathbf{sgn}(Y_i^*(t)) = k_i \tag{4.17}$$

Hence, $\mathbf{sgn}(Y_i^*(t)) = k_i$ for some $k_i \in \{-1, 0, 1\}$. Thus, for all i, we have $\triangle X_i^*(t) > 0$ for all t and $\triangle t > 0$ for all t, so that $k_i = 1$. It is clear that $\triangle X_i^*(t) = 0$ is not optimal given the initial and terminal conditions. Thus, we have that

$$\mathbf{sgn}(Y_1^*(t)) = \mathbf{sgn}(Y_2^*(t)) = \ldots = \mathbf{sgn}(Y_m^*(t)) = -\,\mathbf{sgn}(Y_{m+1}^*(t)) = 1.$$

Now $\mathbf{R}(T_f(\alpha), \alpha)$ terminates on $\mathfrak{M}$, so $\Psi(\mathbf{R}(T_f(\alpha), \alpha)) = 0$ for all α. Let $\mathbf{R}^*(t_f) = (x_1^*, x_2^*, \ldots, x_{m+1}^*) \in \mathfrak{M}$. Hence,

$$X_{m+1}(T_f(\alpha), \alpha) = f(X_1(T_f(\alpha), \alpha), \ldots, X_m(T_f(\alpha), \alpha)) \tag{4.18}$$

for all α. Taking the variation of each side, we have

$$Y_{m+1}^*(t_f^*)\delta T_f + \delta X_{m+1}(t_f^*) = \sum_{i=1}^{m} \frac{\partial f(x_1^*, \ldots, x_m^*)}{\partial x_i}[\delta T_f + \delta X_i(t_f^*)] \tag{4.19}$$

Expanding Equation 4.19 and defining $H_i(t) = \delta X_i(t)$, we have

$$\begin{aligned} Y_{m+1}(t_f^*)\delta T_f + H_{m+1}(t_f^*) &= \sum_{i=1}^{m} \frac{\partial f(x_1^*, \ldots, x_m^*)}{\partial x_i} Y_i(t_f^*)\delta T_f \\ &+ \sum_{i=1}^{m} \frac{\partial f(x_1^*, \ldots, x_m^*)}{\partial x_i} H_i(t_f^*). \end{aligned} \tag{4.20}$$

Rearranging terms, rewriting in vector notation, and letting $f^* = f(x_1^*, \ldots, x_m^*)$ we have

$$\begin{aligned}(\frac{\partial f^*}{\partial x_1}, \ldots, \frac{\partial f^*}{\partial x_{m-1}}, -1) \quad &\cdot \quad (H_1(t_f^*), \ldots, H_{m-1}(t_f^*), H_m(t_f^*)) \\ &+ \quad (\frac{\partial f^*}{\partial x_1}, \ldots, \frac{\partial f^*}{\partial x_{m-1}}, -1) \cdot \dot{\mathbf{R}}^*(t_f^*)\delta T_f = 0, \end{aligned} \tag{4.21}$$

which can be rewritten

$$\nabla\Psi^* \cdot \mathbf{H}(t_f^*) + \nabla\Psi^* \cdot \dot{\mathbf{R}}^*(t_f^*)\delta T_f = 0. \tag{4.22}$$

From Equation 4.14 we write

$$\nabla_{\mathbf{y}} G^*|_{t_f^*} \cdot \mathbf{H}(t_f^*) + G^*|_{t_f^*}\delta T_f = 0. \tag{4.23}$$

Since both Equations 4.22 and 4.23 must be true over all variations and all possible one-parameter families, we have

$$\kappa\nabla_{\mathbf{y}} G^*|_{t_f^*} = \nabla\Psi^*|_{t_f^*} \tag{4.24}$$

for some $\kappa \in \mathbb{R}$. Hence, for $i = 1, 2, \ldots, m+1$ we have

$$\frac{\partial\Psi}{\partial x_i}|_{t=t_f^*} = \kappa a_i \ \mathbf{sgn}(Y_i^*)(t_f^*). \tag{4.25}$$

In the case of $i = m+1$ we have that

$$-1 = \frac{\partial\Psi^*}{\partial x_{m+1}}|_{t=t_f^*} = \kappa a_{m+1}. \tag{4.26}$$

Thus, we have that $\kappa = \frac{-1}{a_{m+1}}$. Hence for $i = 1, 2, \ldots, m$ we have that

$$\frac{\partial\Psi^*}{\partial x_i}|_{t=t_f^*} = \frac{-a_i}{a_{m+1}}. \tag{4.27}$$

This leads to the result that

$$\nabla\Psi^*|_{t_f^*} = \frac{-1}{a_{m+1}}(a_1, a_2, \ldots, a_{m+1}) \tag{4.28}$$

is a normal to the ROC m-manifold $\mathfrak{M}$ at the terminal point of $\mathbf{R}^*(t_f^*)$, the smooth trajectory minimizing J! This is a global minimum, since we are optimizing a convex functional [31]. This agrees with the limited approach based on observation taken by Haspert [18].

The equation of the plane perpendicular to this normal and tangent to the ROC manifold at the optimal point is

$$a_1(x_1 - x_1^*) + a_2(x_2 - x_2^*) + \ldots + a_{m+1}(x_{m+1} - x_{m+1}^*) = 0. \tag{4.29}$$

$\diamondsuit$

To find $(x_1^*, x_2^*, \ldots, x_{m+1}^*)$, generate the plane $a_1x_1 + a_2x_2 + \ldots + a_{m+1}x_{m+1} = 0$, which passes through the origin, and translate it to the ROC m-manifold towards the point $(1, 1, \ldots, 1)$ until the plane rests tangent to the ROC manifold at a single point. This point, $(x_1^*, x_2^*, \ldots, x_{m+1}^*)$, is the terminal point of $\mathbf{R}^*(t_f^*)$. Now, recall that $m = k^2 - k$. We associate the k with the number of classes in a classification problem, so that there is a label set of interest with cardinality k, and a sample space with a partition of cardinality k associated with these labels. Let $l = 1, 2, \ldots, k$, and for each l let $r = 1, 2, \ldots, k$, $r \neq l$. Then associate with each $i = 1, 2, \ldots, m$ an unique (l, r) pair. Designate for each x_i that $x_i = p_{l|r}(A_\theta)$ for some $A_\theta \in \mathbb{A}$, a family of sample functions (classification systems). Thus, the i variables represent the error axes of the k-class classification problem. Similarly, we can designate costs (or losses) for each error by allowing $a_i = c_{l,r}$ for each i. Then the sum,

$$\sum_{l=1}^{k}\sum_{r=1}^{k}\big(1 - \delta_{l,r}\big)c_{l,r}p_{l|r}(A_\theta), \tag{4.30}$$

is the equation for Bayes Risk, or Bayes Error in the case where $c_{l,r} = 1$ for each (l, r) pair. We then have for the specific x_i^*, that there exists $\theta^* \in \Theta$, such that

$$x_i^* = p_{l,r}(A_{\theta^*}), \tag{4.31}$$

so that θ^* is the Baye's Optimal Threshold (or Baye's Optimal Risk Threshold) for the k-class classification problem. When these correspondences can be made (along with the appropriate dimensionality of the parameter space Θ and the ROC manifold $\mathfrak{M}$), it is clear that our optimal trajectory, $\mathbf{R}^*(t_f^*)$, terminates on the ROC manifold corresponding to the Baye's Optimal Threshold (or Baye's Optimal Risk Threshold). Therefore, we have shown how a ROC manifold can be analyzed to find the point corresponding to the Baye's Optimal Threshold.

4.2.2 ROC 1*-manifold optimization (Optimizing the ROC curve).* Here we demonstrate the optimization of ROC 1-manifolds, referred to in this section as ROC curves. We demonstrate that the technique shown in the previous section applies to the case of the two-class problem, with the ROC curves having the axes *typical* in the literature-a true positive axis in the vertical direction and a false positive axis in the horizontal axis. We will only consider ROC curves that are smooth (differentiable) over the entire range, i.e., we consider the set

$$\begin{aligned} \mathrm{C}^1([0,1],\mathbb{R}) \;=\; & \{f : [0,1] \to \mathbb{R} \;:\; f \text{ is differentiable at each } x \in (0,1) \\ & \text{and its derivative } f' \text{ is continuous at each } x \in [0,1]\}. \end{aligned}$$

Given a diagram describing the family of classification systems $\mathbb{A} = \{A_\theta : \theta \in \Theta\}$, with Θ a continuous parameter set (assumed to be one dimensional), and $(\mathrm{E}, \mathscr{E}, \mu)$ a probability space of features, there is a set $\tau_{\mathbb{A}} = \{(\theta, p_{2|1}(A_\theta), p_{1|1}(A_\theta)) : \theta \in \Theta\}$ which is called the *ROC trajectory* for the classification system family $\mathbb{A}$. The projection of the ROC trajectory onto the $\big(p_{2|1}, p_{1|1}\big)$- plane is the set $f_{\mathbb{A}} = \{(p_{2|1}(A_\theta), p_{1|1}(A_\theta)) : \theta \in \Theta\}$ which

is the ROC curve of the classification system family $\mathbb{A}$. Hence, for $h \in [0,1]$ such that $h = p_{2|1}(A_\theta)$ for some $\theta \in \Theta$, we have that

$$[p_{2|1}]^\natural(\{h\}) = \{A_\theta\},$$

that is, the pre-image of h under $p_{2|1}(\cdot)$ 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 $h^* = p_{2|1}(A_{\theta^*}) \in [0,1]$, which may not be unique, unless the function $p_{2|1}(\cdot)$ 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 $\alpha, \beta \geq 0$. Among all smooth curves whose endpoints lie on the point $(0,1)$ and the ROC curve given by $y = f(X(t))$, find the curve, defined by the trajectory $\mathbf{R}(t) = \big(X(t), Y(t)\big)$, for which the functional
>
> $$\mathrm{J}[\mathbf{R}] = \int_0^h \|\mathbf{R}\|_W \, dt = \int_0^h [\alpha|\dot{X}(t)| + \beta|\dot{Y}(t)|]dt \tag{4.32}$$
>
> has a minimum subject to the constraints:
>
> $$\begin{array}{rcl} \mathbf{R}(0) & = & (0,1) \\ \mathbf{R}(h) & = & (h, f(h)), \end{array} \tag{4.33}$$
>
> for some $h \in [0,1]$ that depends on $\mathbf{R}$. We let $X(t) = t$ due to the constraints and denote $W = \dot{X}(t)$ and $Z = \dot{Y}(t)$, so that $\dot{X}(t) = 1$, and Equation 4.32 becomes
>
> $$\mathrm{J}[\mathbf{R}] = \int_0^h [\alpha + \beta|Z(t)|]dt. \tag{4.34}$$
>
> Observe that $h = p_{1|2}(A_\theta)$, $f(h) = p_{1|1}(A_\theta)$ for some $\theta \in \Theta$, and $\beta = \mu(\mathrm{E}_1) = 1 - \alpha$ with $\alpha = \mu(\mathrm{E}_2)$, the prior probability of a class 2 occurrence.

The functional J, when minimized, identifies the trajectory with smallest arclength (measured with respect to the weighted 1-norm). The constraints of Equation 4.33 require that the curve must begin at $(0,1)$ and terminate on the ROC curve. The integrand of Equation 4.34 can be written in a suppressed form

$$\mathrm{G}(t, X(t), Y(t), \dot{X}(t), \dot{Y}(t)) = \mathrm{G}(t, X, Y, W, Z), \tag{4.35}$$

so that the partial derivatives are more easily understood. In the case where $X(t) = t$, then $\dot{X}(t) = 1$ and we have that Equation 4.35 can be further suppressed:

$$\mathrm{G}(t, Y, Z) \tag{4.36}$$

Any $\mathbf{R}$ that minimizes J, subject to the constraints 4.33, necessarily must be a solution to Euler's Equation [13]

$$\frac{\partial}{\partial Y}\mathrm{G}(t, Y, Z) - \frac{d}{dt}\frac{\partial}{\partial Z}\mathrm{G}(t, Y, Z) = 0 \quad \text{for all } t \in (0, h). \tag{4.37}$$

From Equation 4.32 we have $\mathrm{G}(t, Y, Z) = \alpha + \beta|Z|$, so that $\frac{\partial}{\partial Y}\mathrm{G} = 0$ and $\frac{\partial}{\partial Z}\mathrm{G} = \beta\,\mathbf{sgn}(Z)$. Hence, we have that $\mathbf{R}$ solves the Euler equation

$$-\frac{d}{dt}\mathbf{sgn}(Z(t)) = 0 \quad \text{for all} t \in (0, h). \tag{4.38}$$

Integrating this equation reveals that $\mathbf{sgn}(Z(t))$ is constant for all $t \in [0, h]$. Since $Y(t) \leq 1$ for all $t \in (0, h)$, and $Y(0) = 1$, from Constraints 4.33, then $\mathbf{sgn}(Z(t))$ must be 0 or -1, since the trajectory is moving either constantly across to the curve or constantly downward from the point $(0, 1)$. Now, if $\mathbf{sgn}(Z(t)) = 0$ for all t, then $1 = Y(0) = Y(h) = Y(1)$ due to the smoothness of the ROC curve. Substituting this solution into the functional J in Equation 4.32 yields

$$\mathrm{J}[\mathbf{R}] = \alpha h = \mu(\mathrm{E}_2)p_{1|2}(A_\theta), \tag{4.39}$$

with $p_{1|2}(A_\theta) = 1$. Thus, $\mathrm{J}[\mathbf{R}] = \mu(\mathrm{E}_2)$ and the weighted (1-norm) arclength of curve $\mathbf{R}$ is therefore $\mu(\mathrm{E}_2)$. On the other hand, if $\mathbf{sgn}(Z(t)) = -1$ for all $t \in (0, h)$, then

$|Z(t)| = -Z(t)$ and substituting this into J directly in Equation 4.32 yields

$$\begin{aligned} \mathrm{J}[\mathbf{R}] &= \int_0^h \big[\alpha - \beta Z(t)\big] dt \qquad (4.40) \\ &= \alpha h + \beta[Y(0) - Y(h)] \\ &= \alpha h + [1 - Y(h)]\beta \\ &= p_{1|2}(A_\theta)\mu(\mathrm{E}_2) + \big(1 - p_{1|1}(A_\theta)\big)\,\mu(\mathrm{E}_1) \\ &= p_{1|2}(A_\theta)\mu(\mathrm{E}_2) + p_{2|1}(A_\theta)\mu(\mathrm{E}_1). \qquad (4.41) \end{aligned}$$

Notice that Equation 4.41 is identical to the unminimized Bayes Optimal Threshold equation. Therefore, $h = h^*$ which minimizes Equation 4.41 corresponds to the BOT, θ^*, of the family of classification systems, $\mathbb{A}$. The transversality condition [13] of this problem is

$$\alpha + \beta|Z(t)| \mid_{t=h^*} + \beta(f'(t) - Z(t))\mathbf{sgn}(Z(t)) \mid_{t=h^*} = 0 \qquad (4.42)$$

so that

$$f'(h^*) = \frac{\alpha}{\beta} \qquad (4.43)$$

which is

$$f'(h^*) = \frac{\mu(\mathrm{E}_2)}{\mu(\mathrm{E}_1)}. \qquad (4.44)$$

So the transversality condition tells us that the BOT of a family of classification systems corresponds to a point on the ROC curve which has a derivative equal to the ratio of prior probabilities,

$$\frac{\mu(\mathrm{E}_2)}{\mu(\mathrm{E}_1)}.$$

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 $\alpha = C_{1|2}\mu(\mathrm{E}_2)$ and $\beta = C_{2|1}\mu(\mathrm{E}_1)$ where $C_{1|2}$ and $C_{2|1}$ are the costs of

making each error, or we could specify a cost-prior ratio

$$\frac{C_{1|2}\mu(\mathrm{E}_2)}{C_{2|1}\mu(\mathrm{E}_1)},$$

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 (the point $(0, 1)$ for the *typical* two-class ROC manifold classification problem, the origin in the $k > 2$ class case.). Then for each point on the ROC curve, take the 2-norm of the vector which minimizes the distance from this point to the linear variety. If we knew the function generating the ROC curve (or a ROC manifold), we could calculate the optimal ROC directly, but this is not the case in practice.

It is still possible that many ROC curves could be constructed so that the point on the ROC curve corresponding to the BOT for each one has the same distance to the linear variety. This could be a rather large equivalence class of families of classification systems. 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 using 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. AUC equivalence classes are over the entire range of possible priors and therefore 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 classification system 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.

4.3 A Category of Fusors

4.3.1 A Functional for Comparing Families of Classification Systems. We desire a method to compete families of classification systems with the specific intent to compete fusion rules. We show explicitly how to do this with $n = 2$ classes. Although we are proposing one specific functional on the ROC curve to do this, 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 $\mathrm{X} = C([0,1]^{m-1}, \mathbb{R}) \cap C^1((0,1)^{m-1}, \mathbb{R})$, recognizing that we are competing ROC curves, which are by definition a subset of X. The functional

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

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 29 (ROC curve Functional). Let $(\gamma_1, \gamma_2, \alpha, \beta)$ be given data. Let

$$\mathbf{y_0} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mathbf{\Gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix},$$

and

$$V_{\mathbf{\Gamma}} = \{\mathbf{v} \mid \mathbf{v} = k\mathbf{\Gamma}, \forall\, k \in \mathbb{R}\}.$$

Then $V_{\mathbf{\Gamma}} + \mathbf{y}_0$ is a linear variety through the supremum ROC point, $(0, 1)$, over all possible ROC curves, under the data. Let $f \in \mathrm{X}$ 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 $z_{\mathbf{\Gamma}} = \min_{\substack{\mathbf{v} \in \mathbf{V}_{\mathbf{\Gamma}} \\ \mathbf{y} \in \mathbf{T}}} \|\mathbf{v} + \mathbf{y}_0 - \mathbf{y}\|_2$. Then define

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

by

$$F_2(f; \gamma_1, \gamma_2, \alpha, \beta) = \sqrt{2} - z_{\mathbf{\Gamma}}, \forall\ f \in \mathrm{X}. \tag{4.45}$$

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 in ROC space (*i.e.* , $\alpha = 1$ and $\beta = 0$ in the 2-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 [45]. Given a family of classification systems $\mathbb{A} = \{A_\Theta : \theta \in \Theta\}$, the N-P criteria could be written as

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

Theorem 5 (ROC Functional-Neyman-Pearson Equivalence). *Let γ_1 be the dependent constraint, and $\sum_{i=1}^{2} \gamma_i \leq 1$. 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 $\alpha \leq \alpha_0$.*

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

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

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 $\beta_N \in f([0, \alpha_0])$ as the optimal point in the image of f under the N-P criteria. Then $z_N = 1 - \beta_N$ 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 $\beta_N \geq \beta, \forall \; \beta = f(\alpha), \forall \; \alpha \leq \alpha_0$. Hence, $1 - \beta_N \leq 1 - \beta, \forall \; \beta = f(\alpha), \forall \; \alpha \leq \alpha_0$. Then for

$$\mathbf{y}_N = \begin{pmatrix} \alpha_N \\ \beta_N \end{pmatrix},$$

we have that

$$\begin{aligned}\left[(1-\beta_N)^2\right]^{\frac{1}{2}} &= \left\| \begin{pmatrix} \alpha_N \\ 1 \end{pmatrix} - \begin{pmatrix} \alpha_N \\ \beta_N \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} \alpha_N \\ 1 \end{pmatrix} - \mathbf{y}_N \right\| \\ &\leq \left\| \begin{pmatrix} \alpha_N \\ 1 \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right\|\end{aligned}$$

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

$$\min_{\alpha \leq \alpha_0} \left\| \begin{pmatrix} \alpha \\ 1 \end{pmatrix} - \mathbf{y}_N \right\| \leq \min_{\substack{\alpha \leq \alpha_0 \\ \mathbf{y} \in [0,\alpha_0] \times f([0,\alpha_0])}} \left\| \begin{pmatrix} \alpha \\ 1 \end{pmatrix} - \mathbf{y} \right\| \tag{4.46}$$

$$= \min_{\substack{\alpha \leq \alpha_0 \\ \mathbf{y} \in [0,\alpha_0] \times f([0,\alpha_0])}} \left\| \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \mathbf{y} \right\| \tag{4.47}$$

$$= \min_{\substack{\mathbf{v} \in V_\Gamma \\ \mathbf{y} \in [0,\alpha_0] \times f([0,\alpha_0])}} \|\mathbf{v} + \mathbf{y}_0 - \mathbf{y}\| \tag{4.48}$$

On the other hand,

$$\min_{\substack{\mathbf{v} \in V_\Gamma \\ \mathbf{y} \in [0,\alpha_0] \times f([0,\alpha_0])}} \|\mathbf{v} + \mathbf{y}_0 - \mathbf{y}\| \leq \min_{\mathbf{v} \in V_\Gamma} \|\mathbf{v} + \mathbf{y}_0 - \mathbf{y}_N\| \tag{4.49}$$

$$\leq \left\| \begin{pmatrix} \alpha_N \\ 1 \end{pmatrix} - \mathbf{y}_N \right\| \tag{4.50}$$

$$= \left\| \begin{pmatrix} 0 \\ 1-\beta_N \end{pmatrix} \right\| \tag{4.51}$$

$$= 1 - \beta_N. \tag{4.52}$$

Therefore, we have that

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

But $z_\Gamma = 1-\beta_R$, where β_R is the optimal point in the image of f under the ROC functional, so that $\beta_R = \beta_N$. 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. $\diamondsuit$

This idea can be extended to the $k > 2$-class problem by setting a maximum acceptable error rate α_m for each of the $m-1$ independent error axes, where $m = k^2 - k$.

4.3.2 The Calculation and Scalability of the ROC Functional. The calculation and scalability of the functional is straightforward. Suppose we have k classes. In the two-class case, one axis is chosen as $P_{1|1}$, but in the k-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 (the supremum ROC) under this scheme. We choose the conditional class probability $p_{k|k-1}$ to be the dependent one. 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 $p_{k|k-1}$. Let $\mathbf{q} = (q_1,\ldots,q_m)$. Then let

$$Q = \left\{\mathbf{q} \;\middle|\; q_r = p_{i|j}, r = 1,2,\ldots,m, p_{i|j} \leq \alpha_r, r \neq m; i,j = 1,2,\ldots,k; i \neq j\right\} \quad (4.53)$$

be the set of points comprising the ROC curve within the constraints. Then we have that $\mathbf{y}_0 = (0,0,\ldots,0)$ and $\mathbf{N} = \frac{-1}{\gamma_m}(\gamma_1,\ldots,\gamma_m)$, so that if we are given the ROC curve represented by the set Q, call it f_Q, we have that

$$F_n(f_Q;\mathbf{d}) = \sqrt{2} - \min_{\mathbf{q}\in Q}\left\{\frac{\langle \mathbf{q}-\mathbf{y}, -\mathbf{N}\rangle}{\|-\mathbf{N}\|}\right\} = \sqrt{2} - \min_{\mathbf{q}\in Q}\Big\{\,\langle \mathbf{q}, -\mathbf{n}\rangle\,\Big\},$$

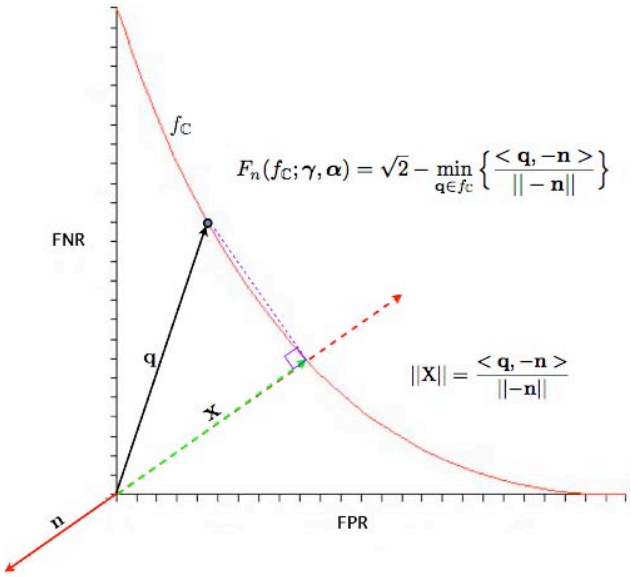


Figure 4.1: Geometry of calculating the ROC functional, F_2, for a point (with vector $\mathbf{q}$) on ROC curve $f_{\mathbb{C}}$.

with $\mathbf{n}$ the unit normal in the direction of $\mathbf{N}$, when Q is not empty, and

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

otherwise. The notation $\langle\cdot,\cdot\rangle$ is the scalar product. Figure 4.1 shows the geometry of the ROC functional calculation where the number of classes is $n = 2$, and the given data is $(\boldsymbol{\gamma},\boldsymbol{\alpha})$.

4.4 The Min-Max Threshold

Suppose we are given a ROC m-manifold with $m = k^2 - k$. This can be viewed as a set of class conditional probability m-vectors $\mathcal{A} = \{\boldsymbol{\alpha}(t) : t \in \mathbf{T}\}$ that form a continuously differentiable, non-increasing function in ROC m-space. Let

$$\mathcal{A}^c = \{\mathbb{I} - \boldsymbol{\alpha}(t) : \boldsymbol{\alpha}(t) \in \mathcal{A}\}, \tag{4.54}$$

where $\mathbb{I}$ is the appropriate identity vector. We associate with each of the m errors a cost c_i, $i = 1, 2, \ldots, m$, so that $\alpha_i(t)$ corresponds to c_i. There are also k prior probabilities, p_k, each prior having $k-1$ copies, so that we can enumerate them and allow each $\alpha_i(t)$ to correspond to p_i, $i = 1, 2, \ldots, m$ where $p_i = p_k$ for each i and some particular k. Let $\mathscr{B} = \{e_1, e_2, \ldots, e_m\}$ be the standard basis for the linear space $\mathbb{R}^m$. Then put $\mathfrak{T} = \sum_{i=1}^{m} c_i e_i$. Then for any $\boldsymbol{\nu} \in \mathbb{R}^m$ we have that

$$\mathfrak{T}\boldsymbol{\nu} = (c_1\nu_1, c_2\nu_2, \ldots, c_m\nu_m)^T \tag{4.55}$$

Now a risk is a decision error times the cost of such an error, so that in our vernacular a risk is $r_i = c_i\alpha_i$. Hence, $\mathfrak{T}\boldsymbol{\alpha}$ is a risk vector and $\mathfrak{T}(\mathcal{A})$ is a risk set. Let $a_i \in \mathbb{R}$ such that $\sum_{i=1}^{m} a_i = 1$. Let

$$\mathscr{R} = \left\{ \mathbf{r} \in \mathbb{R}^m \,:\, \mathbf{r} = \sum_{i=1}^{m} (\mathfrak{T}\boldsymbol{\alpha}_j) a_i, \forall\, \boldsymbol{\alpha}_j \in \mathcal{A} \cup \mathcal{A}^c \right\}.$$

Then $\mathscr{R}$ is a convex risk set. Let $\mathscr{P} = \{\mathbf{p}_i \,:\, i = 1, 2, \ldots, m\}$. $\mathscr{P}$ is a convex set. Now consider also that

$$\langle \mathbf{r}, \mathbf{p} \rangle = \langle \mathfrak{T}\boldsymbol{\alpha}, \mathbf{p} \rangle = \langle \boldsymbol{\alpha}, \mathfrak{T}^*\mathbf{p} \rangle = \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \tag{4.56}$$

showing that $\mathfrak{T}$ is a self-adjoint linear operator on $\mathbb{R}^m$. Since $\mathbb{R}^m$ is a reflexive, normed space, and $\mathscr{R}$,$\mathscr{P}$ are convex subsets of $\mathbb{R}^m$ and $\mathbb{R}^{m*}$ respectively, we have by the Min-Max theorem [31]

$$\min_{\mathbf{r}\in\mathscr{R}} \left[\max_{\mathbf{p}\in\mathscr{P}} \langle \mathbf{r}, \mathbf{p} \rangle \right] = \max_{\mathbf{p}\in\mathscr{P}} \left[\min_{\mathbf{r}\in\mathscr{R}} \langle \mathbf{r}, \mathbf{p} \rangle \right] \tag{4.57}$$

and this occurs where $\mathbf{r}$ and $\mathbf{p}$ are aligned, so that

$$\langle \mathbf{r}_*, \mathbf{p}_* \rangle = \|\mathbf{r}_*\| \, \|\mathbf{p}_*\| \tag{4.58}$$

for the unique $\mathbf{r}_*$ and $\mathbf{p}_*$ which makes Equation 4.57 hold true. Now define $\tilde{\mathcal{A}} = \mathrm{conv}(\mathcal{A} \cup \mathcal{A}^c)$, so then $\tilde{\mathcal{A}}$ is the convex hull of the ROC monifold and its "compliment". Thus, $\tilde{\mathcal{A}} = \mathscr{R}$. Furthermore, we have that $\mathfrak{T}\big(\mathscr{P}\big)$ is a convex subset of $\mathbb{R}^{m*}$, and $\tilde{\mathcal{A}} \subset \mathbb{R}^m$. Thus, the Min-Max theorem applies so that

$$\min_{\boldsymbol{\alpha}\in\mathcal{A}} \big[\max_{\mathfrak{T}\mathbf{p}\in\mathfrak{T}(\mathscr{P})} \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \big] = \max_{\mathfrak{T}\mathbf{p}\in\mathfrak{T}(\mathscr{P})} \big[\min_{\boldsymbol{\alpha}\in\mathcal{A}} \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \big] \tag{4.59}$$

which only occurs where $\boldsymbol{\alpha}$ and $\mathfrak{T}\mathbf{p}$ are aligned

$$\langle \boldsymbol{\alpha}_{**}, \mathfrak{T}\mathbf{p}_{**} \rangle = \|\boldsymbol{\alpha}_{**}\| \, \|\mathfrak{T}\mathbf{p}_{**}\| \, . \tag{4.60}$$

Therefore, we have that

$$\begin{aligned}
\min_{\mathbf{r}\in\mathscr{R}} \big[\max_{\mathbf{p}\in\mathscr{P}} \langle \mathbf{r}, \mathbf{p} \rangle \big] &= \min_{\mathfrak{T}\boldsymbol{\alpha}\in\mathfrak{T}(\mathcal{A})} \big[\max_{\mathbf{p}\in\mathscr{P}} \langle \mathfrak{T}\boldsymbol{\alpha}, \mathbf{p} \rangle \big] && (4.61)\\
&= \min_{\boldsymbol{\alpha}\in\tilde{\mathcal{A}}} \big[\max_{\mathbf{p}\in\mathscr{P}} \langle \mathfrak{T}\boldsymbol{\alpha}, \mathbf{p} \rangle \big] && (4.62)\\
&= \min_{\boldsymbol{\alpha}\in\tilde{\mathcal{A}}} \big[\max_{\mathbf{p}\in\mathscr{P}} \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \big] && (4.63)\\
&= \min_{\boldsymbol{\alpha}\in\tilde{\mathcal{A}}} \big[\max_{\mathfrak{T}\mathbf{p}\in\mathfrak{T}(\mathscr{P})} \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \big] && (4.64)\\
&= \max_{\mathfrak{T}\mathbf{p}\in\mathfrak{T}(\mathscr{P})} \big[\min_{\boldsymbol{\alpha}\in\tilde{\mathcal{A}}} \langle \boldsymbol{\alpha}, \mathfrak{T}\mathbf{p} \rangle \big]. && (4.65)
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathbf{r}_*\| \, \|\mathbf{p}_*\| &= \|\boldsymbol{\alpha}_{**}\| \, \|\mathfrak{T}\mathbf{p}_{**}\| && (4.66)\\
&= \|\boldsymbol{\alpha}_*\| \, \|\mathfrak{T}\mathbf{p}_*\| \, . && (4.67)
\end{aligned}$$

So,

$$\|\mathbf{r}_*\| = k \, \|\boldsymbol{\alpha}_*\| \, , \tag{4.68}$$

where

$$k = \frac{\|\mathfrak{T}\mathbf{p}_*\|}{\|\mathbf{p}_*\|}. \tag{4.69}$$

The point of this section is that the minimax point on the hull of the ROC manifold is now shown to be the point with minimum ℓ_2-norm. This point corresponds to the minimax point of the convex risk set generated by the self-adjoint linear transformation $\mathfrak{T}$ on the ROC manifold. This leads to the conclusion that when a researcher is testing two or more families of classification systems, if he has good knowledge of the prior probabilities, then the ROC functional, F_k, is the preferred functional with which to establish which fusion rules are fusors. On the other hand, if prior probabilities are not understood well, the minimax threshold may be the threshold he would want to compare in order to establish the partial ordering over the fusion rules (and for defining the fusors). In this case, the researcher would want to compare values of the functional

$$G_k(\mathbb{A}_j) = \min_{\boldsymbol{\alpha} \in \mathbb{A}_j} \|\boldsymbol{\alpha}\|_2 \,, \tag{4.70}$$

for each family of classification systems $\mathbb{A}_j$. There is one caveat to the solution here. This is based on research in [42], where it is shown that if the solution to Equation 4.70 is not on the ROC convex hull, then a random decision rule can be developed using the two closest points which are on the convex hull, with this random decision rule being optimal to the optimizing argument of the functional $G_k(\mathbb{A}_j)$. In other words, its 2-norm would be smaller then what the family $\mathbb{A}_j$ can produce.

4.4.1 Defining Fusors. We are now in a position to define a way in which we can compete fusion rules. Suppose we have a fixed classification system such as that in Figure 3.2. Each branch of the system (whether fixed, or associated with a fusion rule) has a ROC manifold 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, take 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 categories (or the feature or label categories 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 branch'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. Since the subcategory of fusors is partially ordered, the best choice for a fusor 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 BOTs for each system. Then calculate 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.

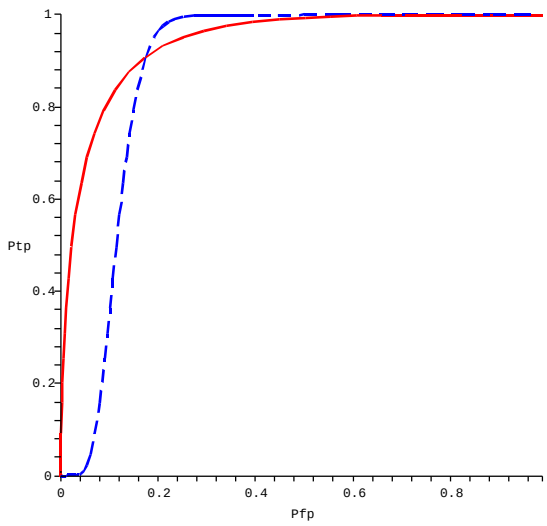


Figure 4.2: ROC Curves of Two Competing Classification Systems.

Definition 30 (Fusor over ROC Manifolds). Let $\mathbb{I} \subset \mathbb{N}$ be a finite subset of the natural numbers, with $\max \mathbb{I} = n$. Given $\{\mathbb{A}_i\}_{i \in \mathbb{I}}$ a finite collection of similar families of classification systems, let $\mathcal{O}_0^{\pi(n)}$ be the category of fusion rules associated with the product of n data sets. Let F_m be the ROC functional on the associated ROC manifolds of the families of classification systems, both original and fused, where $m = k^2 - k$, with k being the

number of classes of interest in the classification problem. Let $\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}\right)$ be the established data for the problem. 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 $\mathfrak{R} \in \mathbf{Ob}(\mathcal{O}_0^{\pi(n)})$, we say that

$$\mathbb{A}_i \succeq \mathbb{A}_j \iff F_m(f_{\mathbb{A}_i}) \geq F_m(f_{\mathbb{A}_j}) \tag{4.71}$$

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 $\mathcal{O}_0^{\pi(n)}$, and whose arrows are induced by the ROC functional, ξ, such that given objects $\mathfrak{R}$ and $\mathfrak{S}$ of this subcategory, then there exists an arrow, $\mathfrak{R} \xrightarrow{\gtrsim} \mathfrak{S}$ if and only if $\mathbb{A}_{\mathfrak{R}} \succeq \mathbb{A}_{\mathfrak{S}}$ if and only if $p_{\mathfrak{R}} \geq p_{\mathfrak{S}}$. This can be seen in the commutativity of the rectangle constructed from Equation 3.9,

$$\begin{array}{ccccccc}
\mathfrak{R} & \longrightarrow & \mathbb{A}_{\mathfrak{R}} & \xrightarrow{\mathfrak{F}} & \mathfrak{F}(\mathbb{A}_{\mathfrak{R}}) = f_{\mathbb{A}_{\mathfrak{R}}} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_{\mathfrak{R}}}) = p_{\mathfrak{R}} \\
\wr\vee & & \wr\curlyvee & & g & & \wr\vee \\
\mathfrak{S} & \longrightarrow & \mathbb{A}_{\mathfrak{S}} & \xrightarrow{\mathfrak{F}} & \mathfrak{F}(\mathbb{A}_{\mathfrak{S}}) = f_{\mathbb{A}_{\mathfrak{S}}} & \xrightarrow{\xi} & \xi(f_{\mathbb{A}_{\mathfrak{S}}}) = p_{\mathfrak{S}}
\end{array}$$

where we can see that in order for the rectangle to commute, that $\gtrsim$ must be a partial order.

We are now in a position to define the fusion processes.

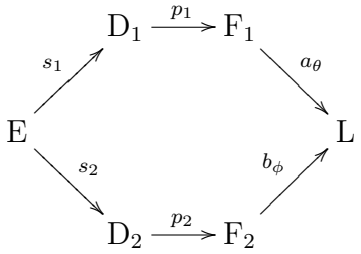
Definition 31 (Fusion-Rule Process). Given a fixed classification problem defined by the category L^{E}, a fusion-rule process is an element of $\mathbf{Ob}(\mathrm{L}^{\mathrm{E}})$.

We didn't really whittle this down from the category of classification systems, because a fusion rule could be the rule "choose classification system X", which doesn't necessarily give a performance improvement. The next definition is the one of interest, since it defines the fusion with the necessary addition of a qualitative element.

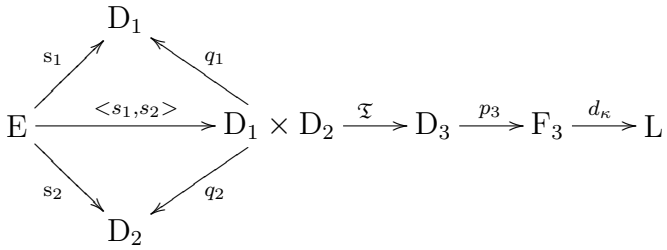
Definition 32 (Fusion Process). Given a fixed classification problem defined by the category L^{E}, and a natural transformation from this category to a category defined by a poset

$\mathcal{P} = (\mathrm{X}, \geq)$, let $\mathbf{FUS}_{\mathrm{L}^{\mathrm{E}}}$ be the subcategory of classification systems induced by the partial ordering. This category has as objects precisely those objects of L^{E} which have an arrow pointing to every fixed branch. We then say a fusion process is an element of $\mathbf{Ob}(\mathbf{FUS}_{\mathrm{L}^{\mathrm{E}}})$, and we can call this category the category of fusion processes.

We have now given a definition of the fusion process which contains everything necessary. As an example, suppose we start with the system



with L a k-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_k 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_k being created under the assumptions and data of the researcher's choice). Then, given fusion rules $\mathfrak{S}$, such as that in Figure 4.3, and $\mathfrak{T}$ and a second fusion system



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 Figure 4.2). Then we have that if $F_k(f_{\mathfrak{S}}) \geq F_k(f_{\mathbb{A}})$ and $F_k(f_{\mathfrak{S}}) \geq F_k(f_{\mathbb{B}})$ and similarly, if $F_k(f_{\mathfrak{T}}) \geq F_k(f_{\mathbb{A}})$ and $F_k(f_{\mathfrak{T}}) \geq F_k(f_{\mathbb{B}})$ then we say that $\mathfrak{S}$, $\mathfrak{T}$ are fusors. Furthermore, suppose $F_k(f_{\mathfrak{S}}) \geq F_k(f_{\mathfrak{T}})$. 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. Figure 4.3 is a diagram showing all branches and products (along with the associated projectors) in category theory notation.

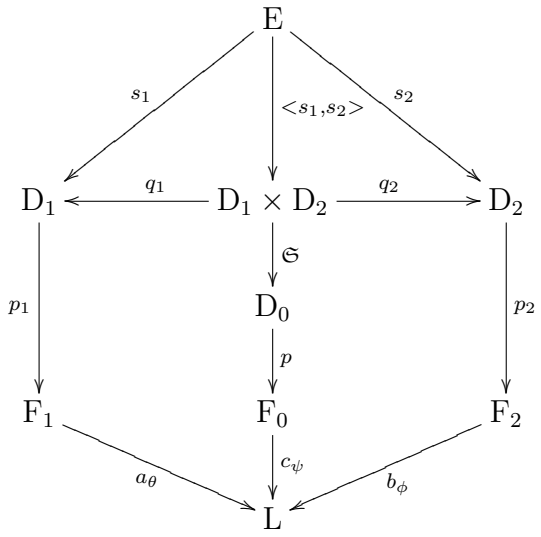


Figure 4.3: Data Fusion of Two Classification Systems.

4.4.2 Fusing Across Different Label Sets. Up to this point, we have considered fusing only those branches of our fixed classification category. This category had a fixed event set and a fixed label set. Sometimes researchers have reason to fuse classification systems which classify events into different label sets before fusion takes place. For example, consider the classification of a mammogram by two classification systems, A_1 and A_2. The first system detects microcalcifications in the breast and returns a result of cancer or non-cancer. The second system detects irregular masses and returns a result of cancer or non-cancer. While the label sets look the same (in fact, bijective), they are not equal. The first partitions the event set into two sets, one where microcalcifications are present and one where they are not. Obviously, irregular masses can occur in either set, so that the cancer label of system A_1 does not correspond with the cancer set of system A_2. We would still like to fuse the results, but now we must consider carefully what should the label set be? It would be prudent to put the label set again as cancer and non-cancer,

which is isomorphic to both the original label sets. The new label set could still be cancer or no cancer, however, these labels induce a new partition of the event space since we now consider cancerous results to be those where microcalcifications **or** irregular masses are returned by the systems. This leads to two definitions developed by Drs. Oxley, Bauer, Schubert, and myself [46].

Definition 33 (Consistent Functor Category of Classification Systems). A functor category of classification systems, $\mathscr{L}^{\mathrm{E}}$, is called consistent when there exists:

1. a probability space $(\mathrm{E}, \mathscr{E}, \mu)$,
2. a finite label set $\mathscr{L} = \{\ell_1, \ell_2, ..., \ell_M\}$,
3. a classification system $\tau \in \mathscr{L}^{\mathrm{E}}$,

such that the set of sets

$$\mathscr{E}_{\mathscr{L}} = \left\{\tau^{\natural}(\{\ell_i\}) \,:\, \ell_i \in \mathscr{L}, i = 1, 2, \ldots, M\right\} \subset \mathscr{E}$$

forms a partition of E. That is, for $\tau^{\natural}(\ell_i) = \mathrm{E}_i$ we have that $\bigcup_{i=1}^{M} \mathrm{E}_i = \mathrm{E}$ and $\mathrm{E}_i \cap \mathrm{E}_j = \emptyset$ for all $i \neq j$. In practice, the classification system τ referred to above, in a consistent, fixed classification system is called the "truth" classifier.

It should be clear from the definition above that

$$\mathrm{P}(\tau^{\natural}(\{\ell_i\})|\mathrm{E}_i) = 1. \tag{4.72}$$

Definition 34 (Within-Fusion Rule). Let S be a fixed classification system with N fixed branches. Assume the following:

- $(\mathrm{E}, \mathscr{E}, \mu)$ is a probability space;
- $\mathscr{L} = \{\ell_1, \ell_2, ..., \ell_M\}$ is a finite label set;
- $\mathscr{L}^{\mathrm{E}}$ is consistent;

- $\mathscr{E}_{\mathscr{L}} = \{\mathrm{E}_{\ell_1}, \mathrm{E}_{\ell_2}, ..., \mathrm{E}_{\ell_M}\} \subset \mathscr{E}$ is the partition of E with respect to $\mathscr{L}$ and truth classifier τ;

Let $\mathbb{A}_{\mathfrak{R}}$ represent the branch generated by fusion rule $\mathfrak{R}$. If for each $m = 1, 2, ..., M$, the fixed branches $\mathbb{A}_1, \mathbb{A}_2, ..., \mathbb{A}_N : \mathrm{E} \to \mathscr{L}$ are designed to map E_{ℓ_m} to ℓ_m, then the fusion rule $\mathfrak{R}$ is said to be a *within-fusion rule.* Furthermore, $\mathbb{A}_{\mathfrak{R}} : \mathrm{E} \to \mathscr{L}$ is designed to map E_m to ℓ_m for each $m = 1, 2, ..., M$.

Definition 35 (Across-Fusion Rule). Let S be a fixed classification system with N fixed branches. Assume the following:

- $(\mathrm{E}, \mathscr{E}, \mu)$ is a probability space;
- $\mathscr{L} = \{\ell_1, \ell_2, ..., \ell_M\}$ is a finite label set, and $\mathcal{L}$ is the power set of $\mathscr{L}$ so that $(\mathscr{L}, \mathcal{L})$ is a measurable space;
- $\mathscr{L}^{\mathrm{E}}$ is consistent;
- $\mathscr{E}_{\mathscr{L}} = \{\mathrm{E}_{\ell_1}, \mathrm{E}_{\ell_2}, ..., \mathrm{E}_{\ell_M}\} \subset \mathscr{E}$ is a partition of E with respect to $\mathscr{L}$ and truth classifier τ;
- $\mathscr{L}^{(0)}, \mathscr{L}^{(1)}, \ldots, \mathscr{L}^{(N)} \subset \mathcal{L}$ are (possibly different) partitions of $\mathscr{L}$, which allow for their functor categories to be consistent, each under a different truth classifier, say τ_n for $n = 0, 1, \ldots, N$;
- for each $n = 0, 1, \ldots, N$, let $M^{(n)} = card(\mathscr{L}^{(n)}) \leq M$, and $\mathscr{L}^{(n)}$ correspond to the label set $\mathrm{L}^{(n)} = \{\omega_1^{(n)}, \omega_2^{(n)}, \ldots, \omega_{M^{(n)}}^{(n)}\}$ in a one-to-one fashion;
- for each $n = 0, 1, \ldots, N$, $\mathscr{E}^{(n)} \subset \mathscr{E}$ is the partition of E with respect to $\mathscr{L}^{(n)}$ (and $\mathrm{L}^{(n)}$) and truth classifier τ_n;

If the families of classification systems,

$$\mathbb{A}_1 : \mathrm{E} \to \mathrm{L}^{(1)}, \mathbb{A}_2 : \mathrm{E} \to \mathrm{L}^{(2)}, \ldots, \mathbb{A}_N : \mathrm{E} \to \mathrm{L}^{(N)},$$

are designed to map each partition set of $\mathscr{E}^{(n)}$ to the corresponding $\omega_j^{(n)} \in \mathrm{L}^{(n)}$ for every $n = 1, 2, \ldots, N$, and $j \leq M^{(n)}$, then the fusion rule $\mathfrak{R}$ that combines the collection of

such systems (yielding a new family of classification systems),

$$\mathbb{A}_0 = \mathfrak{R}(\mathbb{A}_1, \mathbb{A}_2, ..., \mathbb{A}_N),$$

is said to be an *across-fusion rule.* Furthermore, $\mathbb{A}_0 : \mathrm{E} \rightarrow \mathrm{L}^{(0)}$ is designed to map partition sets in $\mathscr{E}^{(0)}$ to the corresponding $\omega_j^{(0)} \in \mathrm{L}^{(0)}$, for $j \leq M^{(0)}$.

The diagram of across-fusion, where $\mathbb{A}_{\mathfrak{R}}$ represents the branch which is essentially a fused branch, is shown in Figure 4.4. If the partitions are equal among the families of

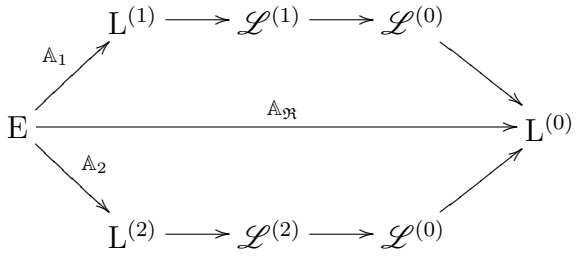


Figure 4.4: Example of Across-Fusion.

classification systems and if the partitions are each injective to $\mathscr{L}$, that is,

$$\mathscr{L}^{(1)} = \mathscr{L}^{(2)} = \ldots = \mathscr{L}^{(N)} = \{\{\ell_1\}, \{\ell_2\}, ..., \{\ell_M\}\}$$

so that

$$\mathrm{L}^{(1)} = \mathrm{L}^{(2)} = \ldots = \mathrm{L}^{(N)} = \{\ell_1, \ell_2, ..., \ell_M\} = \mathscr{L},$$

then there is no need to consider other partitions of $\mathscr{L}$, since clearly

$$\mathrm{L}^{(0)} = \{\omega_1^{(0)}, \ldots, \omega_M^{(0)}\} = \{\ell_1, \ldots, \ell_M\} = \mathscr{L},$$

where $\omega_j^{(0)} = \ell_j$ for all $j = 1, \ldots, M$. Therefore, within-fusion is a special case of across-fusion.

4.5 Changing Assumptions, Robustness, and Example

While we have suggested a family of functionals to use as a way of competing classification systems and fusors, this family is not the only choice available. 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 [3], the loss comparison of a classification system (LC) is denoted by

$$LC = \int I(c_1)L(c_1)dc_1, \tag{4.73}$$

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 $\mathrm{H}_{\boldsymbol{\gamma}}$ by
>
> $$\mathrm{H}_{\boldsymbol{\gamma}} = \left\{\mathbb{A}_j \in R \;\middle|\; F_2(f_{\mathbb{A}_j}; \boldsymbol{\gamma}, \boldsymbol{\alpha}) \geq F_2(f_{\mathbb{A}_i}; \boldsymbol{\gamma}, \boldsymbol{\alpha}), \forall\, i \neq j, i = 1, 2, \ldots, k\right\}.$$
>
> Then, for $\mathbb{A}_i$ we have that
>
> $$LC(\mathbb{A}_i) = \int_{\boldsymbol{\Gamma}} I_{\mathrm{H}_{\boldsymbol{\gamma}}}(\mathbb{A}_i) W(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \tag{4.74}$$
>
> 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

$$S_n(f_{\mathbb{A}}) = \frac{1}{2^n+1}(\sum_{i=1}^{2^n} F_n(f_{\mathbb{A}};\boldsymbol{\gamma}_i,\boldsymbol{\alpha}) + F_n(f_{\mathbb{A}};\boldsymbol{\gamma}_0,\boldsymbol{\alpha})), \tag{4.75}$$

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 [3], 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 systems in Figure 4.5. 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 systems is best really boils down to which classification system is best. Suppose our belief of the situation we are trying to classify is that the ratio of prior probabilities $\frac{\mu\{\ell_1\}}{\mu\{\ell_2\}}$ is $\frac{1}{2}$, with 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, $S_2(f_{\mathbb{C}_1}) = 0.336 \geq 0.330 = S_2(f_{\mathbb{C}_2})$, 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 $f_{\mathbb{C}_1}$ as the best classification system.

There are, of course, other suggestions for performance functionals regarding competing fusion rules. Consider fusion rules as algorithms, divorcing them from the entire classification system. Mahler [33] 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 'local-

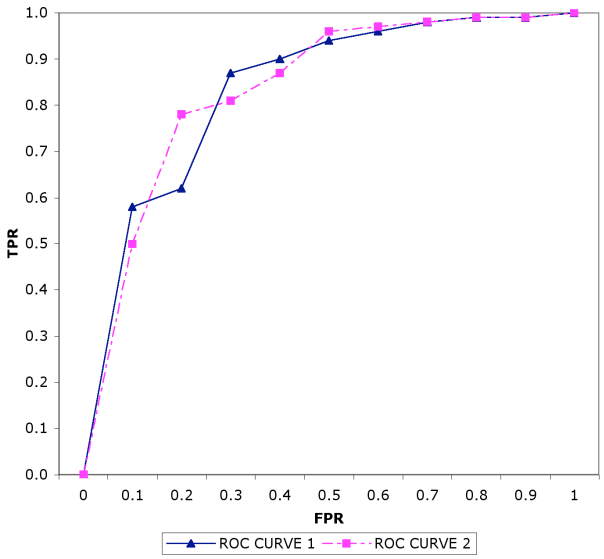


Figure 4.5: ROC Curves of Two Competing Classifier Systems.

ized' in their competence. His preferred approach is to use an information 'metric', the Kullback-Leibler Discrimination functional,

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

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.

V. Conclusions

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. We have shown in this dissertation every fusion system over a finite number of fundamental classification system branches can generate, under test conditions, a corresponding ROC manifold, and under a mild assumption of smoothness of the ROC manifold, a Bayes Optimal Threshold (BOT) can be found for each family of classification systems. Given additional assumptions on the a priori probabilities of a target or non-target, along with given thresholds for the conditional class probabilities, a functional can be generated for each ROC manifold. Any such functional will generate a partial ordering on families of classification systems, on categories of fusion rules, and ultimately on 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 of dimensions higher than 1, as well as to families of classification systems which do not generate ROC manifolds at all. The ROC functional, when populated with the appropriate data choices, will yield a value corresponding the the Bayes Optimal threshold with respect to the classification system family being examined. Another data choice yields the Bayes Cost Threshold, and we have also shown that the Neyman-Pearson threshold of a classification system corresponds to the output of the ROC functional with another fixed data choice (so that it will correspond with the Bayes Optimal Threshold under one particular set of assumptions). Ultimately, a researcher could choose a cost-prior ratio (which seems most reasonable) perturbate it, calculate the mean ROC Functional value, and then 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 (e.g., it would answer the question, “how much change is endured before another classification system is optimal?”) compared with other classification systems. The relationship of the ROC functional to other functionals, including the loss comparison functional, is demonstrated. Finally, there are other functionals to choose, 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.

5.1 Significant Contributions

We believe that significant contributions have been made in this dissertation to the body of knowledge referred to as data or information fusion. The contributions of new and extended applied mathematics were made in the following presentations:

- Rigorous Mathematical descriptions of:
  1. classification systems;
  2. ROC curves, manifolds, spaces.
- Extended and corrected Alsing's ROC convergence theorem [4]. Convergence is shown to occur almost surely as countably infinite random samples are taken from test sample spaces, the sets of which are nested and converging to a true set Ω. The data does not need to be balanced between the classes as assumed by Alsing. We relied upon the writings of Doob [9], Billingsley [5], and Kolmogorov [28] to carefully follow the subtle differences between actual experimental data and its connection to the theory of probability.
- Developed a ROC functional, F_n, which is scalable and can be used without restrictions of continuity, differentiability, convexity, etc., which were necessary to the theory of finding the optimal points on ROC manifolds.
  1. Demonstrated its relation to Bayes Optimal thresholds and Neyman-Pearson thresholds.
  2. Constructed a more robust functional from the ROC functional which may be even more useful than the ROC functional.

- Proved the Min-Max functional is a minimum two norm problem, and can be used without restrictions of continuity, differentiability, convexity, etc., which were necessary to the theory of finding the optimal points on ROC manifolds.
- Demonstrated the pitfalls associated with comparing fusors with fixed branches when doing across fusion, since the label partitions would be different for each classifier family
- Developed a calculus of variations solution to finding optimal elements of ROC manifolds under prior probability and cost constraints for finite classes. This is an extension of known optimizations with two-class problems, which used differential calculus, and is a novel approach which led to discovering a functional that works without the constraints of classifier system families having certain well-behaved properties.
- Developed the Algebra/Category Theory of the fusion of classification systems, including how functors, such as the ROC functional and minimum norm functional, are natural transformations from the categories of fusion rules, and fusors to a partially ordered set. Partial orders arise naturally with an objective function, thereby allowing definitions of fusors to be constructed, as well as defining categories of fusion rules and fusors. This description of data fusion meets the desires of the data fusion communitee as cited in [54].

5.1.1 Recommendations for Follow-on Research. The work described in this dissertation should be supplemented with the following ideas, which make for future research:

- Find universals in the category of fusors. We suspect that the truth fusor and false fusor along with the arrow induced by the partial order may be universal in some way;

- Allow the categories to propagate arrows, such as an arrow representing time in the event-state category. In this way, stochastic processes may be modeled and explained better;
- There is a need to define what situation/threat refinements are in order to apply this fusion process foundation to the elevated levels of data fusion, as described in the JDL functional model;
- Fnd the common theory behind functionals, such as the ROC functional, and the information measures of effectiveness, such as the Liebler-Kullback cross entropy. Also needed is the full relationship between the ROC functional and the AUC;
- The robustness of the classification systems which are minimum with respect to an objective function needs to be explored further, as well as, examining the possibility that costs are not fixed constants, but rather they are functions of the error axes themselves. Then what is the minimizing argument? Is there a way to find this point on the ROC manifold?
- Develop and seek out applications for which our theory explains and describes the process. Our desire is to build up the examples in order to make the explanations more useful and relevant to those not versed in category theory, but for whom this research would be beneficial.

This short list is not comprehensive, but gives a few good topics both within category theory and linear operator theory to expand the state of our current knowledge.

VI. Vita

Major Steven N. Thorsen was born in 1965 in St. Petersburg, Florida. He graduated from Forest Hill High School, West Palm Beach, Florida in 1983. In 1991, he graduated with a B.A. in mathematics from Florida Atlantic University, Boca Raton, FL, whereupon he took a position as a mathematics instructor at Lake Worth High School, Lake Worth, FL.

Maj Thorsen was accepted into Officer Training School in April of 1993. He was commissioned a Second Lieutenant on July 28, 1993. He performed duties as an Emergency Actions Officer and an Operations Plan Officer from 1993 to 1997 at Seymour Johnson AFB, Goldsboro, North Carolina. While serving there, he studied part-time and received the M.A. in mathematics from East Carlolina University, Greenville, NC.

In 1997, Maj Thorsen switched career fields to become an Operational Analyst, and served at Eglin, AFB, Florida as a Requirements Analyst, an Operations Analyst, and the Executive Officer for the 53rd Electronic Warfare Group. In 1999, he served as an instructor in the Department of Mathematical Sciences at the US Air Force Academy, Colorado Springs, Colorado. He was then selected to pursue a Ph.D. in mathematics, and began studies at the Air Force Institute of Technology in August of 2001 at Wright-Patterson AFB, Dayton, Ohio.

Post graduation, Maj Thorsen will serve as Assistant Professor of Applied Mathematics at the Air Force Institute of Technology.

Maj Thorsen has written one refereed journal article [53] to appear, two refereed publications [49, 52], and three conference proceedings and technical articles [39, 50, 51].

Bibliography

1. Adámek, J., et al. *Abstract and Concrete Categories*. New York: John Wiley and Sons, Inc, 1990.
2. Adams, N. M. and D. J. Hand. "Comparing classifiers when the misallocation costs are uncertain," *Pattern Recognition, 32*:1139–1147 (1999).
3. Adams, N. M. and D. J. Hand. "Improving the Practice of Classifier Performance Assessment," *Neural Computation, XII*:305–311 (2000).
4. Alsing, Stephen G. *The Evaluation of Competing Classifiers*. Ph.D. dissertation, Air Force Institute of Technology, Wright-Patterson AFB OH, March 2000.
5. Billingsley, Patrick. *Probability and Measure* (Third Edition). New York: John Wiley and Sons, 1995.
6. Delic, H. "State-Space Approach to Decentralised Detection," *Electronics Letters, 27*(25):2353–2354 (December 1991).
7. DeLoach, Scott A. and Mieczyslaw M. Kokar. "Category Theory Approach to Fusion of Wavelet-Based Features." *Proceedings of the Second International Conference on Fusion (Fusion 1999)*. 117–124. 1999.
8. Desarathy, Belur V. *Decision Fusion*. IEEE Computer Society Press, 1994.
9. Doob, J. L. *Stochastic Processes*. New York: John Wiley and Sons, Inc., 1953.
10. Duda, Richard O., et al. *Pattern Classification* (Second Edition). New York: John Wiley and Sons, Inc, 2001.
11. Egan, James P. *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.
12. Ferri, C. and J. Hernandez-Orallo and M. A. Salido. *Volume Under the* ROC *Surface for Multi-class Problems. Exact Computation and Evaluation of Approximations*. Technical Report, Dep. Sistemes Informatics i Computacio, Univ. Politecnica de Valencia (Spain), 2003.
13. Gelfand, I. M. and S. V. Fomin. *Calculus of Variations*. Englewood Cliffs: Prentice-Hall, Inc, 1963.
14. Green, David M. and John A. Swets. *Signal Detection Theory and Psychophysics*. New York: John Wiley and Sons, 1966.
15. Hall, David L. and James Llinas. *Handbook of Multisensor Data Fusion*. Boca Raton: CRC Press, 2001.

16. Hand, David J. and Robert J. Till. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning*, *45*:171–186 (2001).

17. Hanley, James A. and Barbara J. McNeil. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, *143*:29–36 (April 1982).

18. Haspert, J. K. *Optimum ID Sensor Fusion for Multiple Target Types*. Technical Report IDA Document D-2451, Institute for Defense Analyses, March 2000.

19. Healy, M.J. "Colimits in Memory: Category Theory and Neural Systems." *IEEE Transactions on Neural Networks*. Proceedings of International Joint Conference on Neural Networks (IJCNN '99). 492–496. July 1999.

20. Healy, M.J. "Category Theory Applied to Neural Modeling and Graphical Representations." *IEEE Transactions on Neural Networks*. INNS-ENNS Proceedings of International Joint Conference on Neural Networks (IJCNN '00). 35–40. July 2000.

21. Healy, M.J. and T.P. Caudell and Yunhai Xiao. "From Categorical Semantics to Neural Network Design." *IEEE Transactions on Neural Networks*. Proceedings of International Joint Conference on Neural Networks (IJCNN '03). 1981–1986. 2003.

22. Healy, M.J. and T.P. Caudell. "A categorical semantic analysis of ART architectures." *IEEE Transactions on Neural Networks*. Proceedings of International Joint Conference on Neural Networks (IJCNN '01). 38–43. July 2001.

23. Hoballah, Imad Y. and Pramod K. Varshney. "Distributed Bayesian Signal Detection," *IEEE Transactions on Information Theory*, 995–1000 (September 1989).

24. Hussain, Awais M. "Multisensor Distributed Sequential Detection," *IEEE Transactions on Aerospace and Electronic Systems*, 698–708 (July 1994).

25. Kokar, Mieczyslaw M. and Jerzy A. Tomasik and Jerzy Weyman. "A Formal Approach to Information Fusion." *Proceedings of the Second International Conference on Information Fusion (Fusion'99)*. 133–140. July 1999.

26. Kokar, Mieczyslaw M. and Jerzy A. Tomasik and Jerzy Weyman. "Data vs. Decision Fusion in the Category Theory Framework." *Proceedings of the Fourth International Conference on Fusion (Fusion 2001)*. 2001.

27. Kokar, Mieczyslaw M. and Zbigniew Korona. "A formal approach to the design of feature-based multi-sensor recognition systems," *Information Fusion*, 77–89 (June 2001).

28. Kolmogorov, A. N. *Foundations of the Theory of Probability*. New York: Chelsea Publishing Company, 1956.

29. Lawvere, F. William and Stephen H. Schanuel. *Conceptual Mathematics, A First Introduction to Categories*. Cambridge, UK: Cambridge University Press, 1991.

30. Ludeman, Lonnie C. *Random Processes Filtering, Estimation, and Detection*. Hoboken: John Wiley and Sons, Inc., 2003.

31. Luenberger, David G. *Optimization by Vector Space Methods*. Wiley Professional Paperback Series, New York: John Wiley and Sons, Inc, 1969.

32. MacLane, Saunders. *Categories for the Working Mathematician* (Second Edition). New York: Springer, 1978.

33. Mahler, Ronald P.S. *An Introduction to Multisource-Multitarget Statistics and its Applications*. MN: Lockheed Martin, 2000.

34. Van Trees, Harry L. *Detection, Estimation, and Modulation Theory*. New York: John Wiley and Sons, Inc., 2001.

35. McClarty, Colin. *Elementary Categories, Elementary Toposes*. New York: Oxford University Press, 1992.

36. Metz, Charles E. "Basic Principles of ROC Analysis," *Seminars in Nuclear Medicine*, *8*(4):283–298 (October 1978).

37. Mossman, Douglas. "Three-way ROCs," *Medical Decision Making*, *19*(1):78–89 (Jan-Mar 1999).

38. Munkres, James R. *Topology A First Course*. Englewood Cliffs: Prentice-Hall, Inc., 1975.

39. Oxley, Mark E. and Steven N. Thorsen. "Fusion and integration: What's the difference?." *Proceedings of the Seventh International Conference on Information Fusion*, edited by Per Svensson and Johan Schubert. 429–434. International Society of Information Fusion, Jun 2004.

40. Patton, D. D. "Introduction to Medical Decision Making," *Seminars in Nuclear Medicine*, *8*(4):273–278 (October 1978).

41. Provost, Foster J. and Tom Fawcett. "Robust Classification Systems for Imprecise Environments." *AAAI/IAAI*. 706–713. 1998.

42. Provost, Foster J. and Tom Fawcett. "Robust Classification for Imprecise Environments," *Machine Learning*, *42*(3):203–231 (2001).

43. Royden, H. L. *Real Analysis* (Third Edition). Englewood Cliffs: Prentice-Hall, Inc., 1988.

44. Rudin, Walter. *Real and Complex Analysis* (Third Edition). Boston: WCB/McGraw-Hill, 1987.

45. Scharf, Louis L. *Statistical Signal Processing*. Englewood Cliffs: Prentice-Hall, Inc., 2002.

46. Schubert, Christine and Mark E. Oxley and Kenneth W. Bauer. "A Comparison of ROC Curves for Label-Fused Within and Across Classifier Systems." *Proceedings of*

the Ninth International Conference on Information Fusion. International Society of Information Fusion, 2005.

47. Susanka, Sarah. *The Not So Big House*. Newtown, CT: The Taunton Press, 1998.

48. Swets, John A. "Measuring the Accuracy of Diagnostic Systems," *Science*, *240*:1285–1293 (June 1988).

49. Thorsen, Steven N. *Boundedness of a Class of Linear Operators Constructed from Conditional Expectation Operators*. Masters Thesis, East Carolina University, Greenville NC, March 1997.

50. Thorsen, Steven N. and Mark E. Oxley. "Describing Data Fusion using Category Theory." *Proceedings of the Sixth International Conference on Information Fusion*. 1202–1206. July 2003.

51. Thorsen, Steven N. and Mark E. Oxley. "Comparing Fusors within a Category of Fusors." *Proceedings of the Seventh International Conference of Information Fusion*. 435–441. Stockholm, Sweden: ISIF, June 2004.

52. Thorsen, Steven N. and Mark E. Oxley. "Multisensor Fusion Description Using Category Theory." *IEEE Aerospace Conference*. 2016–2021. 6-13 March 2004.

53. Thorsen, Steven N. and Mark E. Oxley. "A Description of Competing Fusion Systems," *Information Fusion Journal* (2005). (to appear).

54. Wald, Lucien. "Some Terms of Reference in Data Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, *37*(3):1190–1193 (May 1999).

55. Waltz, E. and J. Llinas. *Multisensor Data Fusion*. Norwood: Artech House, 1990.

56. Xu, Lei and Adam Krzyák and Y. Ching Suen. "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition." *IEEE Transactions on Systems, Man, and Cybernetics*. 418–435. May/June 1992.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT DATE**	**3. DATES COVERED** *(From - To)*
15-00-2005	Doctoral Dissertation	May 2003 - Nov 2005

4. TITLE AND SUBTITLE
THE APPLICATION OF CATEGORY THEORY AND ANALYSIS OF RECEIVER OPERATING CHARACTERISTICS TO INFORMATION FUSION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Thorsen, Steven N.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 641, WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/DS/ENC/05-02

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Devert Wicker, Senior Engineer
Air Force Research Laboratory, Sensors Directorate
Bldg 620, Rm C2S69, 2241 Avionics Circle
Wright-Patterson AFB, OH 45433-7321
(937)-255-1115 ext 4155

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Multisensor data fusion is presented in a rigorous mathematical format, with definitions consistent with the desires of the data fusion community. A model of event-state fusion is developed and described . Definitions of fusion rules and fusors are introduced, along with functor categories, of which they are objects. Defining fusors and competing fusion rules involves the use of an objective function of the researchers choice. One such objective function, a functional on families of classification systems, and in particular receiver operating characteristics (ROCs), is introduced. Its use as an objective function is demonstrated in that the argument which minimizes it (a particular ROC), corresponds to the Bayes Optimal threshold, given certain assumptions, within a family of classification systems. This constraint is extended to ROC manifolds in higher dimensions. Under different data assumptions, the minimizing argument of the ROC functional is shown to be the point of a ROC manifold corresponding to the Neyman-Pearson criteria. A second functional is shown to determine the min-max threshold. A more robust functional is developed.

15. SUBJECT TERMS
Category theory, receiver operating characteristic, classification system, calculus of variations, ROC

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Mark E. Oxley, AFIT/ENC
U	U	U	UU	109	**19b. TELEPONE NUMBER (*Include area code*)** (937) 255-7776 ext 4515