Optimal Information-based Classification

by

Baro Hyun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Aerospace Engineering)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor Anouck R. Girard, Co-Chair
Professor Pierre T. Kabamba, Co-Chair
Professor A. Galip Ulsoy
Mariam Abdelhafiz
Associate Professor Mary L. Cummings, MIT

To Jin K. Hyun

ACKNOWLEDGEMENTS

The journey from a junior graduate student to the completion of this dissertation would never have been possible without the dedication of my advisors, Professor Anouck Girard and Pierre Kabamba. I thank Professor Anouck Girard for her straightforwardness and exceptional generosity shown to her students from the first day in Michigan; I thank Professor Kabamba for sharing his enthusiasm in learning and teaching philosophy that have been inspirational. As I look back, our regular meetings were less technically-oriented but more of a collection of delightful anecdotes and educational metaphors, which I learned a great deal from.

I'd like to thank Mariam Faied, for being the "courageous and supportive sister" figure whenever I needed one during our joyful meetings with them.

Next, I'd like to give my gratitude to Professors Galip Ulsoy and Missy Cummings from MIT for being part of my doctoral committee.

A few effusive words to my academic brothers, Christopher Orlowski, Ricardo Bencatel, Johnhenri Richardson and Justin Jackson: it would have been very difficult without sharing the joy and pain with you guys while going through the tunnel of this PhD program. Thanks for being there when I needed help.

I was blessed to work with many talented souls in the ARC lab and owe them many thanks: Zahid Hasan, Calvin Park, Johnny Cheng, Jonathan White (now in D.C.), Weilin Wang, and Andy Klesh (now at JPL).

Special thanks to Sara Spangelo, for making our office a fun and cheerful place along with occasional intellectual discussions, and Chang-Kwon Kang, for teaching

me the basics of tennis and winning against me most of the time.

For the last, but not the least, I would like to give my utmost gratitude to my parents, Nam and Jin, for inspiring me to be where I am, and to my wife, Yukari, for giving me the rest and support whenever I needed, even from Japan.

I'd like to acknowledge that this research was funded by the U.S. Air Force.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

LIST OF APPENDICES

Appendix

ABSTRACT

Classification is the allocation of an object to an existing category among several based on uncertain measurements. Since information is used to quantify uncertainty, it is natural to consider classification and information as complementary subjects. This dissertation touches upon several topics that relate to the problem of classification, such as information, classification, and team classification. Motivated by the U.S. Air Force Intelligence, Surveillance, and Reconnaissance missions, we investigate the aforementioned topics for classifiers that follow two models: classifiers with workload-independent and workload-dependent performance. We adopt workload-independence and dependence as "first-order" models to capture the features of machines and humans, respectively.

We first investigate the relationship between information in the sense of Shannon and classification performance, which is defined as the probability of misclassification. We show that while there is a predominant congruence between them, there are cases when such congruence is violated. We show the phenomenon for both workload-independent and workload-dependent classifiers and investigate the cause of such phenomena analytically.

One way of making classification decisions is by setting a threshold on a measured quantity. For instance, if a measurement falls on one side of the threshold, the object that provided the measurement is classified as one type, otherwise, it is of another type. Exploiting thresholding, we formalize a classifier with dichotomous decisions (i.e., with two options, such as true or false) given a single variable measurement. We further extend the formalization to classifiers with trichotomy (i.e., with three

options, such as true, false or unknown) and with multivariate measurements.

When a team of classifiers is considered, issues on how to exploit redundant numbers of classifiers arise. We analyze these classifiers under different architectures, such as parallel or nested. First, we consider a team of homogeneous (identical) classifiers and provide a fusion-rule, supervisor-based strategy using a parallel architecture. Then, we consider a team of heterogeneous classifiers and provide a strategy using a nested architecture. We show results that confirm that both strategies outperform a single classifier.

CHAPTER I

Introduction

Truth is generally the best
vindication against slander.

Abraham Lincoln

1.1 Background

Classification, also known as categorization, is the matching of an object to one existing category among several on the basis of uncertain measurements. When there are two categories, we speak of *dichotomous* classification, for instance true or false; white or black. When there are three categories, we speak of *trichotomous* classification, for example true, false, or unknown; white, black, or grey. Since the measurements that enable classification contain uncertainty, the classification performance is inherently a random phenomenon. Therefore, probabilistic modeling is the prefered tool to formalize the problem of classification, and the classification performance is quantified by a probabilistic measure. Classification need not be performed by a single entity (a classifier). If a team of classifiers is involved in the act of classifying, it is a *team* classification. The team may consist of classifiers with identical mechanisms and properties (homogeneous team of classifiers) or classifiers with different mechanisms and properties (heterogeneous team of classifiers).

Let us consider classification in military applications [1–3].[1] Unmanned Aerial Vehicles (UAVs) have proved to be an invaluable force multiplier for the Joint Force Commander (JFC) [2]. It is predicted that the UAV market is to more than double over the next decade [4]. UAVs can provide both a persistent and highly capable intelligence, surveillance, and reconnaissance (ISR) platform to troops requiring information [2]; and ISR capability is the number one combatant commander priority for UAVs in the U.S. Army [3].

Due to the technological advance in autonomy, some low-level tasks can be performed by the UAVs themselves. For example, the angle by which the actuators have to move the ailerons on a UAV is completely hidden from the human operator [5]. Despite their effective information-gathering capability and level of autonomy, UAVs still require human operators. Apart from any maintenance or launch and recovery personnel [6], two operators are typically required to operate a UAV: a payload operator and a navigator [5]. However, UAVs in the military are used as sensors rather than as intelligent decision makers.

As a result of rapid advances in electronics and in imaging technologies, UAVs have become the "eyes-in-the-sky" through the use of Electro-Optical (EO) and infrared (IR) sensors, hyperspectral imaging, and LIDAR imaging for instance. However, it has been reported by the Air Force that the volume of sensor data that must be processed from current-generation sensors has become overwhelming, as manpower requirements to deal with these data are burdensome [1]. The analysis of sensor data and the making of some classification decisions are still the role of intelligence personnel. Figure 1.1 illustrates a high-level overview of the U.S. Air Force operation.

[1]This research was supported in part by the United States Air Force grant FA 8650-07-2-3744.

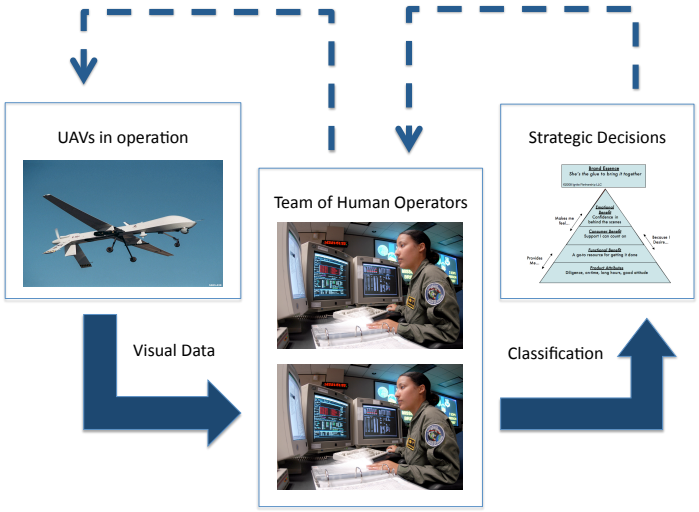


Figure 1.1: An overview of U.S. Air Force ISR operation

1.2 Problem statement

The number of properties one can observe from an object is a determining factor on the success of the observer's classification. Considering the number of properties, or features, an object can possess is a subset of all possible properties an object can hold, it is natural to consider information theory, as we will discuss in more details in later sections. Simply put, information theory is a study of choice and uncertainty given a set of options.

Although information and classification have been both studied with breadth and depth as individual topics, the relationship between them is still unclear. For instance, does increasing or decreasing the amount of information *always* imply anything on the classification performance? Recently, psychologists and neuroscientists have found that there are cases when information is not as informative as we expect, but rather distractive and harmful in terms of making classification decisions [7]. From the standpoint of engineering, information is often used to classify under the assumption that more information gives better classification performance. However, we have little

theoretical knowledge on when such phenomena occur and the reasons behind them.

Furthermore, as pointed out in the previous section, the current state-of-the-art military mission relies on the collaboration between autonomous machines and humans. While the task of information collection is conveniently assigned to autonomous machines, classification decisions are assigned to human operators. Noting that the objective of such missions is making qualitative classification decisions, it is questionable whether the current approach is the best way of utilizing both resources in order to achieve the objective.

The main goal of this dissertation is two-fold: First, to clarify the relationship between information and classification performance, that is, to formally show when information and classification are dependent, or independent, and why. Second, to investigate ways of using a team of multiple classifiers such that the aggregated performance is better than that of a single classifier. For completeness of the dissertation, we revisit the mechanism of a single classifier, provide both analytical and numerical solutions, and further extend the problem to a more generalized setup.

The underlying assumptions of the technical work are as follows:

- The classification performance is solely assessed by the probability of making wrong classification decisions (probability of misclassification). While we are aware that other measures may be as important as the probability of misclassification (time taken to make classification decisions for instance) we believe that it is a proper measure of performance for an initial study. The results presented in the dissertation may vary as the performance measure is modified.

- We define classifier performance as being characterized by the probability of *correct* classification. Thus, improving the classification performance means increasing the probability of correct classification. Throughout the thesis, however, we will often use the notion of probability of *misclassification*, so improving the classification performance indicates *decreasing* the probability of misclassi-

fication since

$$\text{Probability of correct classification} = 1 - \text{Probability of misclassification}.$$

- Our formulation of a classifier is parametric, that is, it requires the knowledge of the distribution which the classifier draws measurements from and makes decisions upon. We assume that the distribution of the measurable property is perfectly known by a prior calibration. Note that the process of calibration is customary in military operations [8–10].

- *Prior* information is defined as the proportion of a type of sub-population among the entire population. We assume that there is a source that provides prior information, such as military intelligence, and that the information is correct.

1.3 Original contributions & anticipated impacts

The original contributions of this dissertation are as follows:

- We show that increasing the amount of information, in the sense of Shannon's, *generally* implies improving classification performance, when classification decisions are made by the maximum likelihood rule and the classification performance is the probability of misclassification. We show the phenomenon for classifiers under two different mechanisms: 1. workload-independent classifier 2. workload-dependent classifier. We demonstrate that, however, increasing the amount of information does *not always* imply improving classification performance, and that this is indeed so for both classifiers with different mechanisms.

- We pose and solve the problem of trichotomous thresholding with a single variable measurement, where the classification decision is based on three options

(true, false, or *unknown*) and determined by two thresholds. Then, we generalize the problem to a multivariate measurement and provide solutions.

- We propose a novel single and team classification model that depends on the individual classifier's confusion matrix and *a priori* information in a static environment. We show that the individual classifiers' decisions in the team can be fused by various logical operators and verify that the single classifier is a special case of the fused model. We show that there are fusion rules that improve the team performance compared to the individual performance.

- We propose a novel classifier architecture that uses a trichotomous classifier with workload-independent performance that turns over the data classified as unknown to a binary classifier with workload-dependent performance. We demonstrate that the novel classifier architecture gives superior classification performance (the probability of misclassification) compared to a single dichotomous classifier. We relate the classifier's performance to the inherent difficulty of the classification task at hand (classifiability), and compare the performances of different classifiers.

The possible implication of these contributions is that one can use the results to design a classification system that does not rely on the quantitative amount of information, but rather on the probability of misclassification. By doing so, a great deal of resources, both machinery and manpower, may be saved since current missions spend much resources on gathering information [1], thus cutting the mission cost. Another implication is that by exploiting the optimal structure of multiple classifiers, the overall classification performance can be significantly improved in comparison with the current operation where it relies on a single classifier.

1.4 Dissertation organization

We begin with a literature survey on several subjects related to the dissertation topic provided in Chap. II. The subjects include information theory and its application, the problem of classification and existing methodologies, and topics related to human factors such as human modeling and human supervisory control.

In the following chapter, we study the relationship between the amount of information and classification performance. We provide some background knowledge, then discuss the relationship for a workload-independent classifier followed by that for a workload-dependent classifier. The main purpose of Chap. III is to clarify when there is dependency, or independence, between information and classification performance, and elucidate why.

In Chap. IV, we study the case of a single classifier by a thresholding scheme. First, we provide a recap on a dichotomous classifier (true or false) given a single variable measurement. Then, we study the solution of a trichotomous classifier (true, false or unknown) given a single variable measurement. Finally, we revisit the solutions of dichotomous and trichotomous classifiers, but further generalize them to the case when a multivariate measurement is given.

The following chapters consider a team of classifiers. In Chap. V, we consider a case with multiple homogeneous (identical) classifiers and provide strategies that give optimal classification decisions. The strategies rely on a notion of a supervisor that uses logical fusion rules. The decision made by the supervisor is optimal in that it minimizes the probability of misclassification of the team. In Chap. VI, we consider a case with multiple heterogeneous classifiers and provide a novel architecture that utilize them. We show that the net performance of the architecture is better than the performance of a single classifier. We also relate the performance to the notion of classifiability, which is a quantification of how inherently difficult a measurement is to classify.

In the last chapter, Chap. VII, we summarize the dissertation, highlight the contributions, and propose future directions. Appendices include a brief review on the Yerkes-Dodson law, proofs of some theorems, an analytical solution to some problems, and a derivation for the classification performance measure.

CHAPTER II

Literature survey

> Science is organized knowledge.
>
> Wisdom is organized life.
>
> Immanuel Kant

This dissertation is at the intersection of several technical areas. Each area is a broad topic by itself, so that an exhaustive review on each topic is beyond the scope of this dissertation. However, we provide an overview on the key results in each area, and the state-of-the-art developments as needed.

We begin with introducing the history of information theory and its applications in Sec. 2.1. In Sec. 2.2, the state-of-the-art techniques in the problem of classification, their applications, and developments in team classification are reviewed. Finally, a literature survey on human operator modeling and human supervisory control is provided in Sec. 2.3.

2.1 Information

2.1.1 Information theory

Information theory deals with the quantification of information and is at the intersection of mathematics, statistics, computer science, physics, neurobiology and

engineering. It was originally developed by Claude Shannon and other engineers at Bell Telephone Laboratories [11] to study fundamental limits on operations such as compressing, reliably storing and communicating data. Information theory now has applications in many areas, including statistical inference, natural language processing, cryptography, network science and other forms of data analysis. It has been a key component in the development of compact discs, mobile phones, cryptography and the Internet.

The history of information theory is generally considered to have started in 1948 with the publication of Claude Shannon's seminal paper, "A Mathematical Theory of Communication" [12]. However, limited information-theoretic ideas appear in the literature prior to 1948.

Harry Nyquist [13] studied factors that affect the "transmission of intelligence", and quantified "intelligence" and the "line speed" at which it can be transmitted by a communication system, giving the relation $W = K \log m$, where W is the speed of transmission of intelligence, m is the number of different voltage levels to choose from at each time step, and K is a constant. In order to quantify continuous signals, such as the spoken language of a person as transmitted over a telephone line, Nyquist discovered a need for discretization of such signals. This observation led to the notion of *sampling* that is now commonly used in the design and analysis of discrete signal processing and discrete-time control systems [14].

A few years later, Ralph Hartley [15] attempted to define information mathematically, making an effort to exclude many personal interpretations of information as it was a loosely defined but easily accessible term. The key ideas behind defining information were the recognition that each word in a message is a selection among a set of possible words and that the selection is sequential; that the more elements in a set of possible words, the more uncertainty in the message; and that knowing the first few words in the message is more informative than knowing the last few. In short, the

more possible words or symbols, the more information each selection carries. Hartley proposed the following formula for the amount of information:

$$H = n \log s, \tag{2.1}$$

where n is the number of symbols transmitted and s is the size of the set of symbols. Another contribution of his was extending the work of Nyquist: Hartley clarified that it is not only the bandwidth of the channel ("line speed") that affects the amount of information but also the time available for transmission.

Shannon [12] posed the central problem of classical information theory, which is the engineering problem of the transmission of information over a noisy channel. The definition of Shannon information, or entropy, usually expressed by the average number of bits needed for storage or communication, plays a central role as a measure of information, choice, and uncertainty. Entropy quantifies the uncertainty involved in predicting the value of a random variable, or, in communications, the expected value of the information contained in a message, usually in units of bits. Equivalently, the Shannon entropy is a measure of the average information content one is missing when one does not know the value of the random variable.

The most fundamental results of the theory are Shannon's source coding theorem, which establishes that, on average, the number of bits needed to represent the result of an uncertain event is given by its entropy; and Shannon's noisy-channel coding theorem, which states that reliable communication is possible over noisy channels provided that the rate of communication is below a certain threshold, called the channel capacity.

The key concepts can be grasped by considering the basis of human communication: language. Intuitively, source coding reflects the notion that common words ("the", "a", "is", ...) should be shorter than less common words ("classification",

"heterogeneous",...) for the sentences to be short. Channel coding reflects the fact that if part of a sentence is not heard or is heard incorrectly due to noise, the listener should still be able to understand the meaning of the message. Building such robustness in communications is done by channel coding. Note that these concepts have nothing to do with the importance or content of the message - the theory is only concerned with the quantity and readability of the message, and not its quality. Shannon's information will be revisited in the next chapter (Chap. III).

While Shannon's information has been the pivotal concept in defining information, it is not the only existing one. Fisher's information describes the information contained in probability distribution functions (pdfs) [16–18], and has been a popular measure in estimation theory. Other notable measures include the Kullback-Leibler entropy (also known as 'cross entropy' or 'relative entropy' [18–20]), and the Rényi information divergence [21]; these are either variations or generalizations of Shannon's information.

2.1.2 Information acquisition

In practice, the acquisition of information by engineering systems is based on data collected through sensors or communication devices. Note that in this thesis, though we are agnostic to how the information is obtained, we focus on models that are more reflective of sensors rather than communication devices, where issues such as communication delays can become important. Commonly used sensor types in aerospace applications include passive (sensors that collect data without emitting any resources, e.g., cameras [22], accelerometers, rate gyros [23]) and active (sensors that collect data by emitting resources, e.g., radars [24], ultrasonic) sensors. Sensors may be isotropic (e.g., ideal antenna [25]), or anisotropic sensors (e.g., most of the sensors). In [26] an acoustic sensor, which has range dependency on spatial and temporal coordinates, is modeled by a Bayesian network. Sensor data generally contain noise,

and this is characterized by the signal-to-noise ratio (SNR) of the sensor.

2.1.3 Information-based trajectories

An information collection system is a mobile agent carrying a receiver that is able to collect information about its environment and store this information for later use. Information collection systems are typically used to accurately obtain, interpret, and utilize knowledge about objects of interest.

Anisotropic sensors within the context of the information collection problem have received much attention to date [27–29]. For instance, fixed camera systems with limited fields of view, direction laser and radar systems are considered in [28, 29], and the references therein. Reference [30] discusses the deployment of such sensors with limited communication but does not address path planning for information collection, nor optimal paths. In previous work of ours, we have posed an optimal control problem for information collection systems where the goal is to maximize the gathered information subject to the vehicle kinematic constraints [31–33]. While proposing a novel information collection model that is based on Shannon's channel capacity equation for isotropic or anisotropic sensors, we provide algorithms that generate optimally informative paths. However, the usefulness of the gathered information was not fully addressed.

Trajectory planning based on information-theoretic measures has been discussed in [34–36]. Missions using information theoretic approaches include area searches with simultaneous localization and mapping (SLAM) [17], decentralized sensor control [37], and optimal sensor placement [38]. In [39], an active sensor management problem for multiple target tracking is formulated using the Rényi information measure, and sensor scheduling algorithms that identify the number and state of moving targets are provided. In [40], a mobile sensing network for the purpose of detecting and capturing mobile targets is provided by posing an optimization problem with a

reward function that represents the improvement in the overall probability of detection. In [8], the problem of path planning for UAVs in the presence of radar-guided surface-to-air missiles is posed as a minimax optimal control problem. While the objective of the dynamic optimization problem is to minimize the maximum of the probability of tracking an aircraft by radar-guided surface-to-air missiles subject to kinematic constraints, the authors present the necessary conditions for optimality and provide extensive numerical solutions under various scenarios. Further work has been completed by solving optimization problems with taking sensor and kinematic constraints into account. For instance, [41] considers an agent seeking shortest paths to a goal with limited field-of-view camera while [42] seeks to minimize the total wheel rotation of a ground vehicle with accounting for the disc kinematics. However, they do not consider implications of information theory within their studies. Reference [43] uses information-theoretic measures, e.g., Shannon's information and Fisher information matrix, to guide the information gathering process for mobile agents in surveillance with strapped-down sensors. In the context of predicting information collection, there have been approaches such as robotic adaptive sampling [44] where expected information gain is used to govern the best sensor paths. There is a large body of work that applies optimal control theory to information gathering problems in terms of belief state formulation or the dual estimation-control problem; notable examples are [28, 45, 46]. Reference [47] considers the development of a target assignment algorithm for information collection and uses heuristic methods.

Information theory is applicable to team-decision problems [48] and artificial intelligence such as in lunar robotic colonies [49].

In our present problem, gathering information has a purpose, which is to make better classification. The widely accepted view on the relationship between information and decision-making, or classification, is that they are congruent, i.e., more information implies a better probability of correct classification. However, to the best

of our knowledge, work that formally addresses whether this is true or not is not in the literature.

2.2 Classification

Classification is a process of allocating items to different sets, and accordingly, it can be equated with categorizing, i.e., the grouping and labeling of items with similar properties together (by sorts). The problem of classification has been tackled in many fields using different names, for example in mathematics (category theory), in computer science (sorting algorithms), in machine learning and pattern recognition (supervised learning, data mining), and in statistics (cluster analysis, inference).

The notion of classification can be traced back to Aristotle in ancient Greece [50]. Indeed, classification in essence is similar to the notion of predicate in classical logic: a predicate is either true or false depending on a discourse variable taken from the universe of discourse [51]. A formal definition of classification will be provided in Chap. III.

2.2.1 Theory of classification

Since the notion of "thinking machine" was first presented by Alan Turing [52, 53] in the 1930s, devising a machine classifier that does not depend on human guides has been one of the holy grails in fields that deal with computation. Efforts building such classifiers have stimulated many sub-fields, such as sensing, data mining, statistical inference, decision making, and learning.

At this point, we introduce a more technical definition of classifier than the one mentioned in the introduction. A classifier is a decider, i.e., a deterministic mapping defined from a set of data into truth values, with the domain of the mapping being a specific realization of a random variable. While both a decider and a classifier are deterministic mappings, the difference between them is that the latter accounts for

the *randomness* of the data being classified.

A good classifier is able to recognize the properties of an entity (pattern recognition, novelty detection), knows how to extract the key information among the properties (data analysis), and is consistent in its performance given the same information (consistency). In [54], the complementary abilities in data analysis of humans and computers are discussed using, as an example, the game of chess. The author also discusses the definition of intelligent data analysis, for example, pattern recognition, and unintelligent data analysis, for example, by over-refinement. The quality of classification is determined by two aspects: how likely the classification is to be correct, and how quickly the result is obtained.

Much attention has recently been paid to the learning aspect of classifiers (thus, "machine learning"). There are largely two types of learning: supervised and unsupervised. Supervised learning requires a set of "training" data such that the classifier can learn the true categories in the dataset, similar to a calibration process, and then the performance of the classifier is examined by a set of "test" data. Thus, supervised learning inherently consists of two-step procedures which are training and testing. On the other hand, unsupervised learning does not require training data, therefore the classification is performed without the knowledge of the true categories and must be learned as it goes. Clustering [55] is one of the common unsupervised learning techniques, among which k-means clustering [56] and Fuzzy c-means clustering [57] are popular ones.

Arguably, the first effort towards a learning classifier were using neural networks. A neural network is a collection of learning agents, known as neurons, interconnected to other agents by communication links. A network attempts to mimic the input-output patterns of data and performs classification based on the learned patterns. Since the discovery of neurons in biology inspired the modeling of such agents by McCulloch and Pitts [58], on which the general neural network form is based, a large

body of research has appeared in the literature in search of a better network for better classification. The Perceptron by Frank Rosenblatt [59] is one of the earliest linear classifiers that is essentially a feed-forward neural network. However, it was only capable of classifying linearly separable data. Later on, "ADAptive LInear Neuron", or ADALIN, was devised by Berney Widrow, and it is more capable than the perceptron [60–62]. The cornerstone of ADALIN was provable results on the performance that the error converges to the least square error asymptotically by the learning rule. A multi-layer version of ADALIN (MADALIN) was later devised as a nonlinear classifier. There have been many more advances in the theory of neural networks; notable ones include back propagation by Werbos [63] and the Cognitron by Fukushima [64]. For a good review on the history of neural networks, see [65].

In statistics, the problem of classification typically appears as a regression problem. For standard regression methodologies, see [66] and the references therein. In [67], the problem of static classification is posed and the solution is investigated analytically. In [68], the author presents a brief review of the available techniques for assessing the accuracy of remotely sensed data and the necessary considerations related to the data such as the classification system, the sampling scheme, the sample size, and spatial autocorrelation.

Thresholding is a particular method of classification where the classification decision is made based on the evaluation with respect to some threshold values. We will revisit the mechanism of thresholding in Chap. IV. The method is ubiquitous in many fields that range from statistics to coding theory, and image processing [69–72]. For a thorough review on the state-of-the-art image thresholding techniques, see [69] and the references therein.

2.2.2 Applications of classification

In robotics, the problem of classification appears as a machine learning problem, such as pattern recognition or novelty detection. Pattern recognition is a study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. In [73], an extensive review survey is presented. In [74], a method for multi-class optical pattern recognition of different perspective views of an object is described.

Novelty detection is the identification of a new or unknown data that a machine learning system is not aware of during training. Novelty detection is similar to the problem of static classification, but it focuses on the novelty of the new signal rather than the similarity to the known signal. For state-of-the-art statistical techniques, see [75] and the references therein.

Recently, much attention has been devoted to the problem of classification with augmenting human operator inputs. In [76, 77], a problem of classifying large datasets in bioinformatics without any *a priori* information is addressed. The authors propose an adaptive active clustering scheme that initially clusters the data sets unsupervised and then allows for adjustment of the classification by the user. Using this approach, the authors show that the misclassification rate can be improved quickly. In [78], a similar problem is addressed, but the focus is on the maximum reduction of the number of learning data involving human interactions. The author proposes a self-controlled exploration/exploitation strategy to select data points based on the combination of representativeness and classifier uncertainty. In [79], a system that collects and analyzes data of the multimodal communication between human-human or human-robot interaction is presented. In [80], the psychology behind some heuristics of human experts' decision-making is discussed.

The problem of classification with human input also relates to aerospace appli-

cations. In the Air Force's Intelligence, Surveillance and Reconnaissance missions, a team of human operators and UAVs are paired to improve the finding of targets. In [81], the authors investigate the use of human operator feedback for target recognition in an ISR scenario where a team of Micro Aerial Vehicles (MAVs) is assigned to fly over a number of objects of interest and the operator must decide whether the object is a target or not. In [82], the authors present decision making strategies under uncertainty and adversary action for the Cooperative Operations in UrbaN TERrain program (COUNTER), where stochastic dynamic programming is employed to optimize the fuel reserves of a UAV.

All the aforementioned work is on the problem of static classification, where the classifier is stationary; there is relatively less work on the problem of kinematic classification, where the classifier can choose where to move. Reference [19] addresses a problem where the goal is to classify one or multiple fixed targets located in an obstacle-populated workspace by planning the path; however, the authors assume that maximizing information leads to the optimal classification performance. We will show this is only a general trend in later chapters.

2.2.3 Team classification

Methods of combining several classifiers in order to enhance the overall classification performance (or shortly, team classification) first appeared in handwriting recognition problems [83–86] and then extended to more general pattern recognition problems such as speech recognition, remote sensing and medical applications [87, 88]. Some of the early methods of team classification are majority voting [85, 86], where each individual classifier has a vote and the category with the majority vote wins, and subset ranking [87], where each individual classifier creates a subset of categories with assigned ranking and the category with the majority ranking wins. There are methods that use fusion rules [89], multiple neural networks [90], expert opinions [91]

and many more. However, since the range of applications that pattern recognition techniques are applicable to is wide, it is difficult to compare the performance of different team classification methods; a method that is applicable in one application may not be applicable in other. Hence the scope of such studies is often limited to a specific type of problem.

Another line of research similar to team classification is the team decision problem. The classical team decision problem is to find the best communication system and the best decision rules, given the expected outcome of decisions, the probabilities of situations, and the cost of communication [92, 93]. In [94], the team decision problem is posed as an optimal control problem. A tutorial on team decision theory and information structures between the teammates is presented in [95], and the references therein. In [96], a novel team model is formulated with a study investigating the role of uncertainty between teammate interaction. It is shown that the optimal level of interaction decreases as the level of uncertainty increases. Motivated by various Air Force missions, in [97], a mathematical definition of collective and collaborative systems is made based on classical team decision theory.

When there are multiple opinions, there are ways to reach a collective decision, such as voting; these methods are termed as *fusion rules*. A study on the trade-off between accuracy and decision time for decision-makers in decentralized settings is presented in [98]. This work considers two fusion rules that combine multiple decision-makers' opinions to make a final collective decision. In [99], it is shown that for a collective decision based on multiple decision makers, the optimal fusion rule is the likelihood-ratio test under the assumption that each individual decision is conditionally independent from any other.

Loosely speaking, one can argue that machine classifiers are good at keeping consistent performance, but they lack in ability to recognize properties and process information when compared to humans. On the other hand, humans are superior to

machine classifiers in recognizing properties and processing information, but they lack consistency in performance. Assuming that these statements correctly characterize the features of machine classifiers and humans to some extent, it is obvious that there are complementary aspects between the two entities. Our work is motivated by recognizing such complementary aspects and seeks collaboration strategies in the task of classification. However, there may be applications where such assumptions do not agree with practice, and in that domain our work should be applied cautiously. In order to clarify the domain, we find a need to understand the characteristics of humans as decision-makers, specifically, how they are modeled and are exploited. The following is the literature review on human modeling and supervisory control.

2.3 Human operator modeling & Human supervisory control

Modeling humans is notoriously difficult, and when done, leads to models that are valid only within a small domain of interest. We divide the section into two subsections, each focusing on: 1. models that try to capture certain human behaviors, 2. supervisory controllers that guide (and restrict) human behaviors assuming that they follow some model.

Research over the past few decades has suggested a number of statistical operator models for certain types of tasks [100–104]. Reference [100] tests the stimuli response of Air Force Cadets in different exhaustion states. In [101], a Bayesian model for optional stopping is developed for the two-hypothesis tasks in which subjects may purchase risk-reducing information before making a decision. Also, a nonparametric model for choice-reaction time is derived. In [102], a speed-accuracy study is presented applying a sequential probability ratio decision procedure on the alternatives for two-choice situations. In [103] the capacity of operators to adapt to changes in a setup that focuses on speed vs. accuracy of the operator responses is examined. Experimental results show that emphasis on speed decreases mean reaction time, but

increases errors. In [104], a linear relationship between the odds of a correct response and reaction time is presented based on the previous works of [101–103] and others. Furthermore, the author suggests the definition of a speed-accuracy operating characteristic analogous to the receiver operating characteristic (ROC) in signal detection theory. In [105], an operator decision model is developed that is based on a binomial distribution, and the validity of the model is shown experimentally in some specific scenarios. In the work of [106], operator behavior is modeled by Hidden Markov Models (HMMs) which allow the inference of higher operator cognitive states from observable operator interactions with a computer interface. The results suggest that the best matching models with experimental data are obtained through unsupervised learning.

Recently, attempts have been made to model human operator behavior in the framework of Discrete Event Systems (DES). In [5, 107], a DES model is used to replicate human-unmanned vehicle behavior for a heterogenous unmanned vehicle system. A validation experiment showed that the model captures the impact of increasing heterogeneity on operator utilization. Reference [108, 109] show ways of modeling human operators using DES.

The Yerkes-Dodson law [110] relates the arousal and performance of a human, specifically that the human performance is poor when the arousal is either too low or too high, but the performance reaches its peak when it is moderately aroused. We will revisit the Yerkes-Dodson law later in Chap. III and Appendix A. In [111], the effect of arousing words on the response time and the skin conductance is studied. Pioneering work by Sheridan [112] suggests a model of the human decision-maker, which is compared to experimental results for human subjects performing a task at a computer-graphics terminal and where the notion of workload is the control variable for supervision. This work suggests a definition of mental workload, which can vary person-by-person, and a notion of expert and novice operators differing by

their productivity.

Analytical and experimental studies on human supervisory control have been conducted. One of the foci in human supervisory control is the modeling of human operators via interaction with a computer interface. A number of human-machine interaction strategies are proposed in [5, 113] for a single operator-multiple heterogeneous vehicles scenario. This single human operator-multiple autonomous vehicles is a new paradigm that inverts the current operator-to-unmanned-vehicles ratio. References [114, 115] study this topic specifically and provide supervisory time strategies (interact, wait, or neglect) for humans. Attention allocation is a study on guiding human's attention in highly-distractive environments. Using active interfaces that provide situation and activity awareness [116], devices such as tactile feedback [117] or auditory feedback [118] are studied as methods for human attention allocation.

Recently, much work has investigated simplified, but essence-capturing, supervisory architectures based on operator models in light of new modeling techniques. Reference [119] developed a supervisory controller that was designed to regulate the operators' cognitive state based on the discretized Yerkes-Dodson law. A team of heterogeneous operators, differing by their task-servicing rate, was modeled using a DES framework, and the validity is shown by an experiment. In [120, 121] a human operator is considered as a state-dependent queueing process where the state is a task arriving at a deterministic rate and optimal control policies are provided such that the queue does not overflow. In [122], a decision support system for sequential visual search tasks is presented while the effectiveness of the system is validated by human-subject experiments. It is shown that the human operator performance improves under the decision support system with automated algorithmic aids.

2.4 Conclusion

In this chapter, we have reviewed several technical areas, covering information theory and its application; classification and team classification; human operator modeling and human supervisory control. In the course of the review, we have discovered several open issues in the literature:

- The relationship between information and classification performance is still unclear, i.e., whether there is an agreement or disagreement between the two measures. Theoretical studies dealing with such issues for classifiers with different mechanisms (workload-independent or workload dependent) have not been found.

- While studies on dichotomous classification have been found in many fields, theoretical studies on trichotomous classification, from mathematical formalization of the problem to numerical and analytical solutions, have not been found. Specifically, a comprehensive study starting from a dichotomous classifier with a single variable measurement to a trichotomous classifier with multivariate measurements, and showing how the former can be generalized to the latter has not been found in the literature.

- Although many fusion-rule based approaches have been proposed, an extensive parametric study in search of the optimal *synergistic* fusion rule, a term that we will define in Chap V, and the sensitivity of such an optimal rule with respect to the environment has not been found in the literature.

- Formalization of a *nested* architecture using heterogeneous classifiers and the validation of its performance have not been found in the literature.

In the following chapter, we will begin with investigating the relationship between information and classification performance. We will begin with introducing some

background knowledge with mathematical definitions, such as Shannon's information and maximum likelihood classification, then provide results for classifiers with workload-independent performance followed by classifiers with workload-dependent performance.

CHAPTER III

On the Independence of Information and Classification Performance

> The pure and simple truth is rarely pure and never simple.
>
> Oscar Wilde

It is believed that collection of sufficient information is the necessary condition for making good classification decisions [81]. Thus, often the objective of a surveillance mission is to gather as much information as possible such that a classification decision can be made with high confidence. Although it is a widely accepted view that more information implies better classification performance, there has been little work on formally proving this. Moreover, to the best of our knowledge, there has been no work on investigating the relationship between information and classification performance for classifiers with different properties.

In this chapter, we study the relationship between the amount of information and the classification performance. We use standard mathematical concepts to investigate the relationship, such as Shannon's information and the maximum likelihood rule, and show that, while there is a predominant congruence between the amount of information and the classification performance, there is also independence between them.

We begin with providing some background knowledge on some concepts, such as the definition of classifiers, probabilistic modeling for our problem formulation, Shannon's information and the maximum likelihood classification. Then, we study the relationship between the amount of information and the probability of misclassification of a classifier whose performance is unaffected by workload, followed by that of a classifier with workload-dependent performance.

3.1 Background

3.1.1 Classifiers

A decider D is a deterministic mapping defined on a set of data into truth values, i.e.,

$$D : \{\text{data}\} \rightarrow \{T, F\}.$$

A classifier C is a decider with the domain of the mapping being a specific realization of a random variable. While both decider and classifier are deterministic mappings, the difference between them is that the latter accounts for the *randomness* of the data being classified.

Processing of the data requires two abilities: recognizing truth out of truth (rate of true positives) and falsehood out of falsehood (rate of true negatives). These abilities are characterized by two independent parameters, σ_T and σ_F, respectively. Note that these parameters are entries in the confusion matrix in signal detection theory [123].

3.1.2 Probabilistic modeling

Collecting information and making classification decisions are generally based on some measurements, and these measurements are typically obtained through imperfect sensors. Since these imperfect sensors introduce uncertainties in the measurement, e.g., sensor noise, the characteristics of the measurements are random. There-

fore, we use probabilistic modeling, rather than deterministic, to investigate the relationship between information and classification performance.

Let X be a discrete random variable that denotes the category of objects of interest that can take two realizations: either T or F.[1] There is a probability associated with the event that X be one of the realizations, given as,

$$P(X = T) = u,\ P(X = F) = 1 - u, \tag{3.1}$$

where $u \in [0,\ 1]$. We denote u as the *prior probability* and it represents the proportion of T objects among the objects of interest.

Let Y be a discrete random variable that denotes an object property that can take two realizations $Y \in \{Y_1, Y_2\}$. Sample Y_1 represents the sensor measuring a property from an F object while Y_2 represents the sensor measuring a property from a T object. For instance, Y_2 can be the profile of a gun from a picture taken from the broadside view of a tank (threat) while Y_1 can be the wheels or the windshield from a picture taken from an automobile (friend).

Note that the number of realizations of Y can be more than two, but we restrict our modeling for simplicity and clarity.

The likelihood of the object property given the object category is modeled by conditional probabilities. For two-option object categories and two-option object properties, the conditional probabilities are given as,

$$\begin{aligned} P(Y = Y_1|X = F) &= \sigma_F, \\ P(Y = Y_2|X = T) &= \sigma_T, \\ P(Y = Y_2|X = F) &= 1 - \sigma_F, \\ P(Y = Y_1|X = T) &= 1 - \sigma_T, \end{aligned} \tag{3.2}$$

[1]Note that "T" and "F" can be interpreted as "True" and "False", respectively, or as "Threat" and "Friend". The subsequent theory does not require choosing an interpretation.

where $\sigma_F, \sigma_T \in [0.5, 1]$ parameterize the conditional probabilities. When $\sigma_{(\cdot)} = 0.5$ the sensor is as bad as a pure guess, while when $\sigma_{(\cdot)} = 1$ the sensor is perfect. Note that the range $\sigma_{(\cdot)} \in [0, 0.5]$ describes the same phenomenon as $\sigma_{(\cdot)} \in [0.5, 1]$, but in a perverse manner.

3.1.3 Shannon's information

Given two random variables (X and Y), Shannon's information [12] describes the uncertainty reduction in one of the random variables (X) by observing another (Y). We begin the definition of information by introducing the notion of entropy.

Definition III.1. Entropy

Entropy is a measure of uncertainty associated with a random variable that can take realizations with assigned probabilities. The entropy of a random variable X is

$$H(X) = -\sum_x P(X = x) \log P(X = x). \tag{3.3}$$

Definition III.2. Conditional entropy

Conditional entropy is a measure of uncertainty associated with a random variable conditioned upon knowledge of another random variable. The conditional entropy of the random variable X conditioned upon the random variable Y is

$$\begin{aligned} H(X|Y) &= \sum_y P(Y = y) H(X|Y = y) \\ &= \sum_y P(Y = y) \sum_x P(X = x|Y = y) \log P(X = x|Y = y). \end{aligned} \tag{3.4}$$

Definition III.3. Shannon's information

Shannon's information is the difference between the entropy and the conditional en-

tropy, i.e.,

$$I(X;Y) = H(X) - H(X|Y). \tag{3.5}$$

For instance, if $I = 0$ then there is no reduction of uncertainty in X by observing Y.

3.1.4 Maximum likelihood classification

The maximum likelihood classification, also known as *likelihood-ratio rule* [99], is a decision rule based on posterior probabilities.

Definition III.4. Bayes' rule

Bayes' rule gives the posterior probability of X given Y. For instance, given $Y = Y_1$ the posterior probability of $X = T$ is

$$P(X = T|Y = Y_1) = \frac{P(Y = Y_1|X = T)P(X = T)}{P(Y = Y_1)}. \tag{3.6}$$

Note that $P(Y = Y_1)$ can be computed by following the theorem of total probability [124].

Definition III.5. Likelihood-ratio rule

Let $O_s \in \{T, F\}$ be a decision variable that follows the likelihood-ratio rule, i.e.,

$$O_s = \begin{cases} T & \text{if } \frac{P(X=T|Y=Y_1)}{P(X=F|Y=Y_1)} > 1 \\ F & \text{if } \frac{P(X=T|Y=Y_1)}{P(X=F|Y=Y_1)} \leq 1. \end{cases} \tag{3.7}$$

Let $f_{Y_0} \in [0, \infty]$ denote the ratio of the posterior probabilities such that,

$$f_1 = f_{Y=Y_1} = \left(\frac{1-\sigma_T}{\sigma_F}\right)\left(\frac{u}{1-u}\right), \tag{3.8a}$$

$$f_2 = f_{Y=Y_2} = \left(\frac{\sigma_T}{1-\sigma_F}\right)\left(\frac{u}{1-u}\right). \tag{3.8b}$$

Let $\delta_{O_{s0}} : \mathcal{R} \rightarrow \{0, 1\}$ be such that

$$\delta_T(f) = \delta_{O_s=T}(f) = \begin{cases} 1 & \text{if } f > 1 \\ 0 & \text{if } f \leq 1, \end{cases} \tag{3.9a}$$

$$\delta_F(f) = \delta_{O_s=F}(f) = \begin{cases} 1 & \text{if } f \leq 1 \\ 0 & \text{if } f > 1. \end{cases} \tag{3.9b}$$

Then, the conditional probabilities of O_s given Y are,

$$P(O_s = T|Y = Y_2) = \delta_T(f_2), \tag{3.10a}$$

$$P(O_s = T|Y = Y_1) = \delta_T(f_1), \tag{3.10b}$$

$$P(O_s = F|Y = Y_2) = \delta_F(f_2), \tag{3.10c}$$

$$P(O_s = F|Y = Y_1) = \delta_F(f_1). \tag{3.10d}$$

3.1.5 Classification performance

The classification performance is defined by the probability of misclassification, and is the sum of the probabilities of two faulty outcomes: false positives and false negatives:

$$P_m = P(O_s = T \wedge X = F) + P(O_s = F \wedge X = T). \tag{3.11}$$

Although we have considered the generic case of equal weights for the two outcomes, there can be different weights associated with the outcomes depending on the strategic objective of the classifier.

Assessing the probability of misclassification yields

$$\begin{aligned} P_m = & P(O_s = T \wedge X = F | Y = Y_2) P(Y = Y_2) \\ & + P(O_s = T \wedge X = F | Y = Y_1) P(Y = Y_1) \\ & + P(O_s = F \wedge X = T | Y = Y_2) P(Y = Y_2) \\ & + P(O_s = F \wedge X = T | Y = Y_1) P(Y = Y_1), \end{aligned} \tag{3.12}$$

by the theorem of total probability. Assuming that the classification is unbiased, we can simplify the expression using conditional independence, i.e., $P(O_s = O_{s0} \wedge X = X_0 | Y = Y_0) = P(O_s = O_{s0} | Y = Y_0) \cdot P(X = X_0 | Y = Y_0)$. This means that given a measurement Y, the classifier's decision O_s does not affect the object category X, and *vice versa*. Substituting, Eq. (3.12) yields:

$$\begin{aligned} P_m = & P(O_s = T | Y = Y_2) P(X = F \wedge Y = Y_2) \\ & + P(O_s = T | Y = Y_1) P(X = F \wedge Y = Y_1) \\ & + P(O_s = F | Y = Y_2) P(X = T \wedge Y = Y_2) \\ & + P(O_s = F | Y = Y_1) P(X = T \wedge Y = Y_1). \end{aligned}$$

Finally,

$$\begin{aligned} P_m = & \delta_T(f_2)(1 - \sigma_F)(1 - u) + \delta_T(f_1)\sigma_F(1 - u) \\ & + \delta_F(f_2)\sigma_T u + \delta_F(f_1)(1 - \sigma_T)u. \end{aligned} \tag{3.13}$$

3.2 Classifiers with workload-independent performance

3.2.1 Summary

Shannon's information and the probability of misclassification can be expressed in terms of σ_T and σ_F (defined in Eq. (3.2)), which are given as

$$\begin{aligned} I(X;Y) = & -u\log u - (1-u)\log(1-u) \\ & + (1-\sigma_T)u\log\left\{\frac{(1-\sigma_T)u}{(1-\sigma_T)u+\sigma_F(1-u)}\right\} \\ & + \sigma_F(1-u)\log\left\{\frac{\sigma_F(1-u)}{(1-\sigma_T)u+\sigma_F(1-u)}\right\} \\ & + \sigma_T u\log\left\{\frac{\sigma_T u}{\sigma_T u+(1-\sigma_F)(1-u)}\right\} \\ & + (1-\sigma_F)(1-u)\log\left\{\frac{(1-\sigma_F)(1-u)}{\sigma_T u+(1-\sigma_F)(1-u)}\right\}, \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} P_m = & \delta_T(f_2)(1-\sigma_F)(1-u) + \delta_T(f_1)\sigma_F(1-u) \\ & + \delta_F(f_2)\sigma_T u + \delta_F(f_1)(1-\sigma_T)u. \end{aligned} \tag{3.15}$$

Note that these two measures are functions of σ_T, σ_F, and u.

3.2.2 Analytical properties

The following theorems are proved in the appendix.

Theorem III.6. *If $\sigma_T \geq 0.5$ and $\sigma_F \geq 0.5$, then Shannon's information $I(X;Y)$ in Eq. (3.14) is a monotonically increasing function of both σ_T and σ_F.*

Theorem III.7. *If $\sigma_T \geq 0.5$ and $\sigma_F \geq 0.5$, then the probability of misclassification P_m in Eq. (3.15) is a monotonically decreasing function of both σ_T and σ_F.*

Corollary III.8. *If $\sigma_T \geq 0.5$ and $\sigma_F \geq 0.5$, and they both undergo increments $\Delta\sigma_T$ and $\Delta\sigma_F$, respectively, with $\Delta\sigma_T \cdot \Delta\sigma_F > 0$, then the corresponding changes in $I(X;Y)$ and P_m, $\Delta I(X;Y)$ and ΔP_m, respectively, satisfy $\Delta I(X;Y) \cdot \Delta P_m < 0$.*

For instance, if σ_T and σ_F are increasing, then Shannon's information $I(X;Y)$ is monotonically increasing while the probability of misclassification P_m is monotonically decreasing. *Vice versa* for decreasing σ_T and σ_F. Corollary III.8 states that there is a predominant congruence between the amount of information and the classification performance.

It is noted, however, that the congruence is because increasing Shannon's information and decreasing the probability of misclassification occur *simultaneously*, when both are caused by increasing σ_T and σ_F, not because increasing information directly implies decreasing probability of misclassification (this would be the logical fallacy of the undistributed middle [51]).

Proposition III.9. *Increasing the amount of information yields an increase of the probability of misclassification only if there is a trade-off between the rates of true positives and true negatives, i.e., $\Delta\sigma_T \cdot \Delta\sigma_F \leq 0$.*

Proof. By contraposition of Corollary III.8. □

Investigating whether the theorem holds for other mechanisms of classification decisions is not within the scope of this study, and it is left as future work.

3.2.3 Numerical examples

Figure 3.1 illustrates the contour plot of information and probability of misclassification in the σ_T-σ_F plane for non-informative prior ($u = 0.5$). Note that the curved lines with boxed levels indicate the information while the straight lines with levels indicate the probability of misclassification.

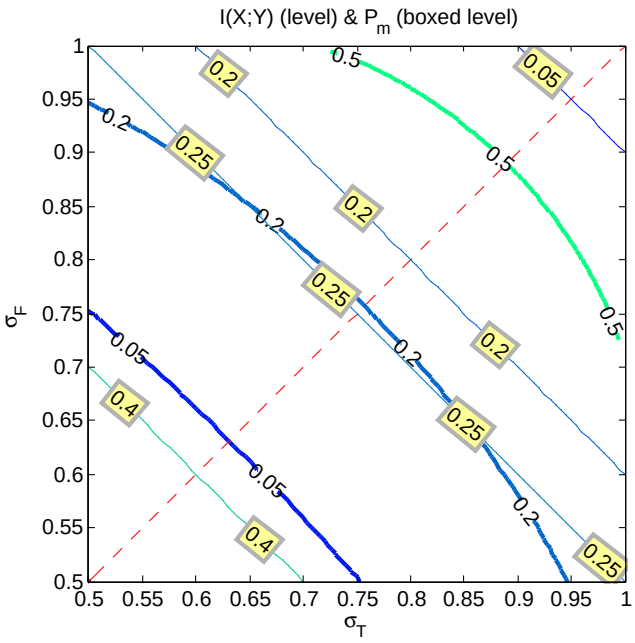


Figure 3.1: Comparison of information and probability of misclassification for $u = 0.5$. The red dashed line is where $\sigma_T = \sigma_F$.

Observation 1. *At $\sigma_T = \sigma_F = 0.5$, the information is zero, and at $\sigma_T = \sigma_F = 1$, it reaches its maximum. From $\sigma_T = \sigma_F = 0.5$ to $\sigma_T = \sigma_F = 1$ following the diagonal line ($\sigma_T = \sigma_F$), the information is a monotonically increasing function. On the other hand, at $\sigma_T = \sigma_F = 0.5$, the probability of misclassification is at its maximum ($P_m = 0.5$) and at $\sigma_T = \sigma_F = 1$ it is zero. From $\sigma_T = \sigma_F = 0.5$ to $\sigma_T = \sigma_F = 1$ following the diagonal line, the probability of misclassification is a monotonically decreasing function.*

Observation 2. *Following the diagonal line ($\sigma_T = \sigma_F$) in Fig. 3.1 along which the gradients of the two measures are collinear, one can observe that there is indeed a general trend showing that increasing the amount of information does imply decreasing the probability of misclassification.*

This observation agrees with the general understanding of the relationship between information and classification performance: One needs more information to improve the classification decision.

Remark III.10. When the prior information is unhelpful ($u = 0.5$), improving σ_T and σ_F equally increases the amount of information collected and decreases the probability of misclassification.

Observation 3. *The information collected by improving σ_T and σ_F is larger as both σ_T and σ_F become perfect. For instance, beginning from σ_T and σ_F that are pure guesses, if σ_F becomes perfect ($\sigma_T = \sigma_F = 0.5 \rightarrow \sigma_T = 0.5$, $\sigma_F = 1$), the information collected is 0.3 ($\Delta I = 0.3$). On the other hand, from a perfect σ_F and only pure guess σ_T, if both σ_T and σ_F become perfect ($\sigma_T = 0.5, \sigma_F = 1 \rightarrow \sigma_T = 1$, $\sigma_F = 1$), the information collected is 0.7 ($\Delta I = 0.7$). Note that the gains in the classification performance for both cases are the same ($\Delta P_m = 0.25$).*

3.2.3.1 A counter example

Figure 3.2 shows a part of the contour region that was shown in Fig. 3.1. Note

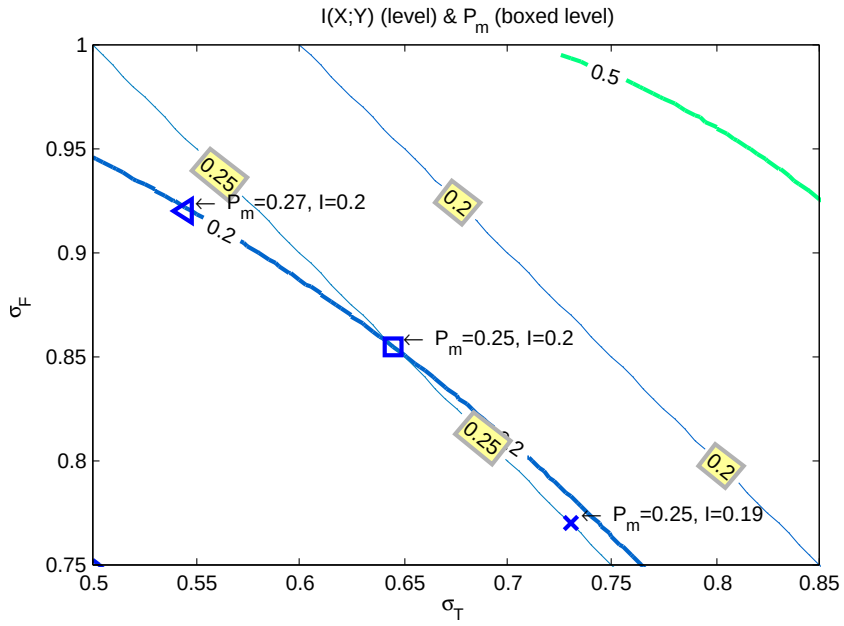


Figure 3.2: Comparison of information and probability of misclassification (zoomed Fig. 3.1)

that there are three markings on the contour (a triangle, a square, and a cross), each indicating a different level of information and probability of misclassification.

Observation 4. *Beginning at the cross and following the probability of misclassification level curve towards the square, one can notice that although the information has increased, the probability of misclassification remains constant ($P_m = 0.25$, $I = 0.19 \rightarrow P_m = 0.25$, $I = 0.2$).*

Observation 5. *Beginning at the square and following the information level curve towards the triangle, one can notice that although the information level is constant, the probability of misclassification has increased ($P_m = 0.25$, $I = 0.2 \rightarrow P_m = 0.27$, $I = 0.2$).*

These are observations that demonstrate that increasing the amount of information collected does *not always* imply improving the classification performance of a workload-independent classifier.

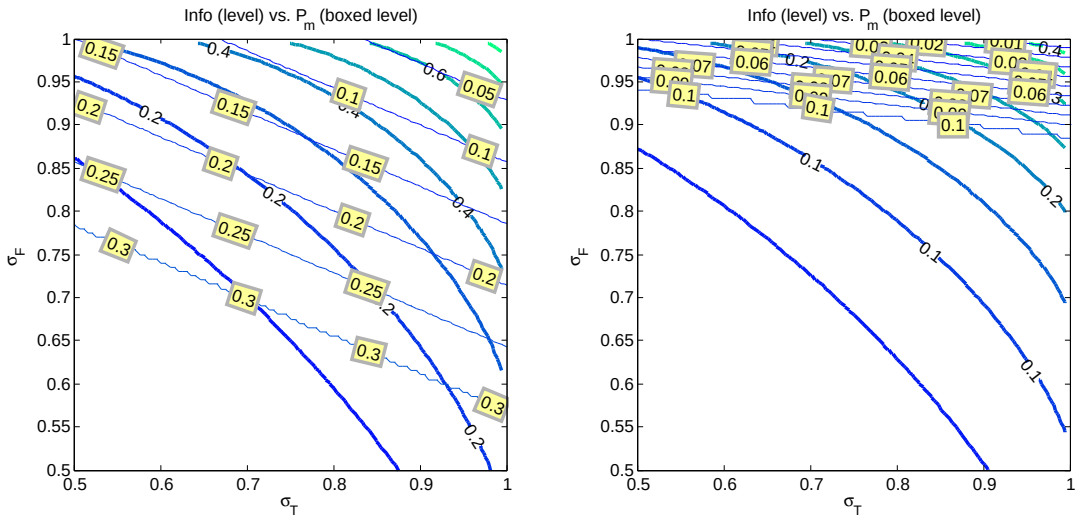
3.2.3.2 Does the prior information matter?

Figure 3.3 shows two contour plots of the information and the probability of misclassification with respect to two different informative priors ($u < 0.5$).

Observation 6. *As u deviates from 0.5, neither the information nor the probability of misclassification are symmetric with respect to the line $\sigma_T = \sigma_F$.*

Note that the prior information determines the contribution of σ_T and σ_F in assessing the probability of misclassification. For instance, for small prior information on the T sub-population (Fig. 3.3 (b)), increasing σ_F has greater contributions on decreasing the probability of misclassification than increasing σ_T.

Also note that even with informative priors, the general trend that increasing information implies improving classification performance is prevalent. However, one



(a) $u = 0.3$ (b) $u = 0.1$

Figure 3.3: The effect of prior information on I and P_m

can also find a counter example that increasing information does not imply improving classification performance.

Observation 7. *As the prior information decreases, the amount of information needed to decrease the probability of misclassification is increased.*

3.2.3.3 Having incorrect prior information

In this section, we investigate the impact of having incorrect prior information on Shannon's information and the probability of misclassification.

Let $u \in [0, 1]$ be the true prior information and $\hat{u} \in [0, 1]$ be the *belief* prior information. The error in prior information is the difference between the true and the belief, i.e.,

$$\Delta u = u - \hat{u}, \tag{3.16}$$

where $\Delta u \in [-1, 1]$. $\Delta u = -1$ indicates that the belief is *extremely aggressive* $(u = 0, \hat{u} = 1)$ while $\Delta u = 1$ indicates that the belief is *extremely conservative* $(u =$

$1, \hat{u} = 0$).

Let I denote Shannon's information associated with the true prior information u and $\hat{I}$ denote Shannon's information associated with the belief prior information $\hat{u}$. Let $\Delta I = I - \hat{I}$ be the error information, i.e., the difference between the true and the belief information.

Similarly, let P_m denote the probability of misclassification associated with the true prior information u and $\hat{P}_m$ denote the probability of misclassification associated with the belief prior information $\hat{u}$. Depending on the range of the prior information, the probability of misclassification in Eq. (3.15) can be expressed in the following form:

$$P_m = \begin{cases} 1-u & \text{if } f_1 > 1 \wedge f_2 > 1, \\ (1-\sigma_T)u + (1-\sigma_F)(1-u) & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ u & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{3.17}$$

Similarly, the probability of misclassification for the belief prior information can be expressed as

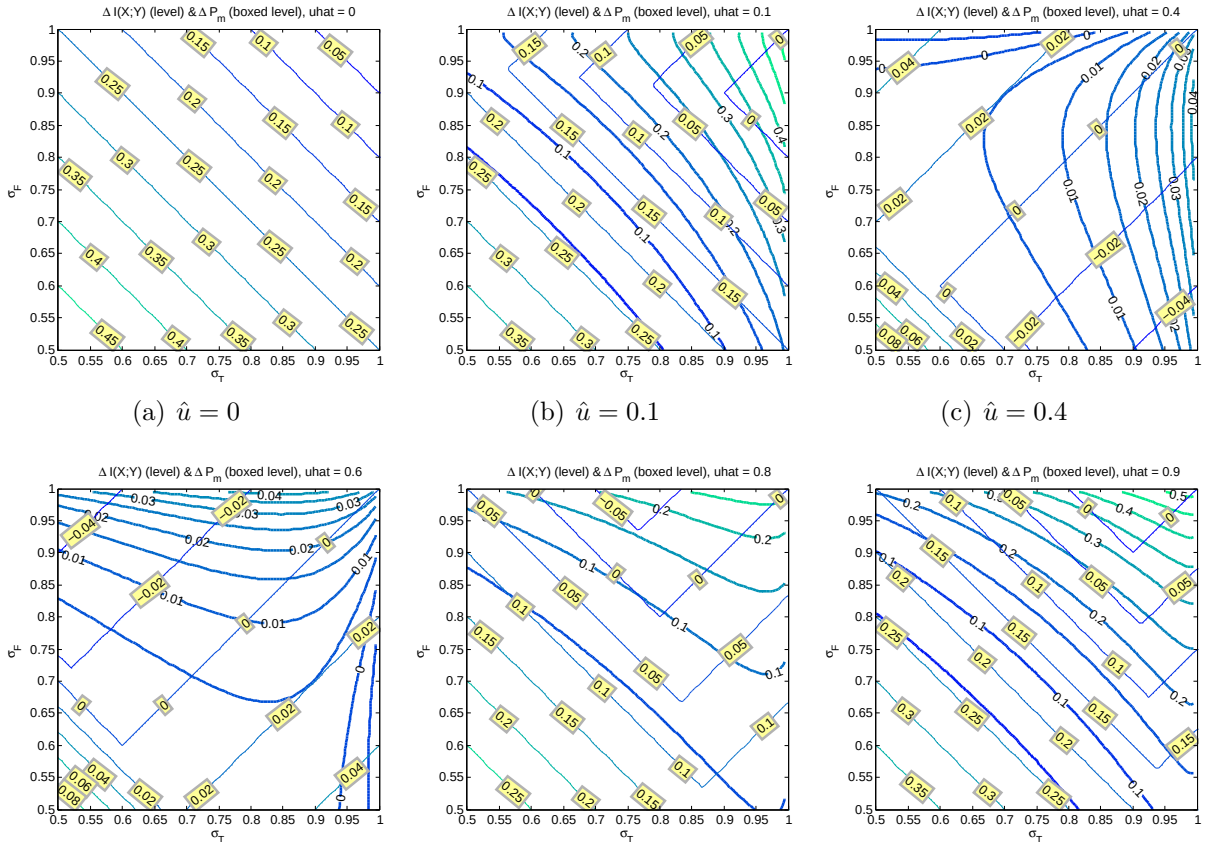
$$\hat{P}_m = \begin{cases} 1-\hat{u} & \text{if } \hat{f}_1 > 1 \wedge \hat{f}_2 > 1, \\ (1-\sigma_T)\hat{u} + (1-\sigma_F)(1-\hat{u}) & \text{if } \hat{f}_1 \leq 1 \wedge \hat{f}_2 > 1, \\ \hat{u} & \text{if } \hat{f}_1 \leq 1 \wedge \hat{f}_2 \leq 1, \end{cases} \tag{3.18}$$

where

$$\hat{f}_1 = \left(\frac{1-\sigma_T}{\sigma_F}\right)\left(\frac{\hat{u}}{1-\hat{u}}\right), \tag{3.19a}$$

$$\hat{f}_2 = \left(\frac{\sigma_T}{1-\sigma_F}\right)\left(\frac{\hat{u}}{1-\hat{u}}\right). \tag{3.19b}$$

Note that u and $\hat{u}$ need not be correlated, i.e., the belief can be completely different from the truth. Therefore, there are nine possible outcomes of the error probability



(a) $\hat{u} = 0$ (b) $\hat{u} = 0.1$ (c) $\hat{u} = 0.4$

(d) $\hat{u} = 0.6$ (e) $\hat{u} = 0.8$ (f) $\hat{u} = 0.9$

Figure 3.4: Error information ΔI and error probability of misclassification ΔP_m with respect to $\hat{u}$ on σ_T-σ_F plane ($u = 0.5$)

of misclassification that depend on the range of u and $\hat{u}$.

Let $\Delta P_m = P_m - \hat{P}_m$ be the error probability of misclassification, i.e., the difference between the true and the belief probability of misclassification. Note that the smaller the error, the closer to the truth. Table 3.1 gives the list of the nine possible outcomes of the error probability of misclassification.

Figure 3.4 illustrates the error information ΔI and the error probability of misclassification ΔP_m with respect to the belief prior information $\hat{u}$. The results shown in Fig. 3.4 can be counterintuitive. For instance, for $\hat{u} = 0.8$ (Fig. 3.4 (e)) and $\sigma_F = 0.6$, the error probability of misclassification ΔP_m is a decreasing function for σ_T increasing from 0.5 to around 0.85, while ΔP_m becomes an increasing function for

Table 3.1: The error probability of misclassification $\Delta P_m = P_m - \hat{P}_m$

		$\hat{P}_m$		
	Condition	$\hat{f}_1 > 1 \wedge \hat{f}_2 > 1$	$\hat{f}_1 \leq 1 \wedge \hat{f}_2 > 1$	$\hat{f}_1 \leq 1 \wedge \hat{f}_2 \leq 1$
P_m	$f_1 > 1 \wedge f_2 > 1$	$\hat{u} - u$	$(\sigma_T - \sigma_F)\hat{u} - u + \sigma_F$	$1 - u - \hat{u}$
	$f_1 \leq 1 \wedge f_2 > 1$	$(\sigma_F - \sigma_T)u - \sigma_F + \hat{u}$	$(\sigma_F - \sigma_T)(u - \hat{u})$	$(\sigma_F - \sigma_T)u - \sigma_F - \hat{u} + 1$
	$f_1 \leq 1 \wedge f_2 \leq 1$	$\hat{u} + u - 1$	$(\sigma_T - \sigma_F)\hat{u} + u + \sigma_F - 1$	$u - \hat{u}$

σ_T increasing from around 0.85 to 1. In other words, when the true and belief prior information do not agree, improving a sensor performance σ_i, $i \in \{T, F\}$ may improve the error caused due to the misbelief ΔP_m only until a certain point, and once beyond that point, improving the sensor performance can worsen the error. Similar counter phenomenon can be found in other cases of $\bar{u}$ as shown in Fig. 3.4. One can use such results as a guide to design sensors when a disagreement between the true and belief prior information is anticipated.

3.3 Classifiers with workload-dependent performance

In this section, we investigate the relationship between the amount of information and the probability of misclassification of a workload-dependent classifier.

The relationship between workload and performance of a classifier is depicted by the Yerkes-Dodson law. It states that, while low or high workload degrades the performance of a classifier, there is a region of workload that yields optimal classification performance. More details can be found in Appendix A.

Let $W \in [0, 1]$ be a workload variable with 0 indicating idle and 1 indicating fully loaded. Recognizing the concavity of the curve in Fig. A.1, we model the Yerkes-Dodson law as a quadratic function of the workload as,

$$\sigma_i = -(4\sigma_i^* - 2)W_i^2 + (4\sigma_i^* - 2)W_i + 0.5,\ i \in \{T, F\}, \tag{3.20}$$

where $\sigma_i^* \in [0.5, 1]$ determines the maximum of σ_i. At $W_i = 0.5$, $\sigma_i = \sigma_i^*$ and at $W_i \neq 0.5$, $\sigma_i < \sigma_i^*$.

It is not known whether when the sensor performance parameters σ_T and σ_F are affected by the workload variable W, they change independently or exhibit any correlations. In this work, we consider the general case, i.e., parameters change independently, so that cases with correlation can be considered as a specific example

of the general results.

Let W_i denote the workload variable corresponding to σ_i with $i \in \{T, F\}$. Given the workload-dependent σ_i with $i \in \{T, F\}$, we numerically investigate the information and probability of misclassification as a function of the workload, where the relevant relationships are summarized in Eq. (3.14)-(3.15).

3.3.1 Numerical examples

Figure 3.5 shows the performance, probability of misclassification, and information as a function of the workload for a peak performance of $\sigma^* = 1$. Similar to the

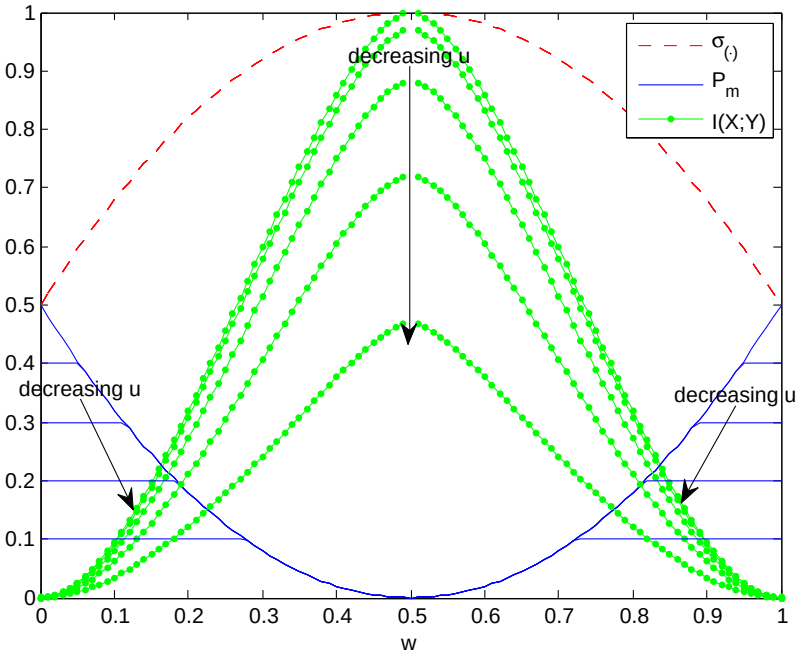


Figure 3.5: Probability of misclassification (P_m) and information ($I(X;Y)$) of a workload-dependent classifier vs. workload variable (W). The prior information (u) is varied from 0.5 to 0.1 with a decrement of 0.1.

Yerkes-Dodson law, Shannon's information for a workload-dependent classifier reaches its maximum at the optimal workload, i.e., $W = 0.5$. Similarly, the probability of misclassification for a workload-dependent classifier reaches its minimum at the

optimal workload.

From Fig. 3.5, it can also be observed that the prior information (u) determines the peak value of the information collected when the workload is optimal ($W = 0.5$) and changes the information collection rate ($\partial I(X;Y)/\partial W$) in $W \neq 0.5$. For the probability of misclassification, the prior information determines the maximum of P_m and the region where the maximum P_m remains constant, however, it does not affect the minimum of P_m. Note that the workload variables where the information reaches its maximum and the probability of misclassification reaches its minimum are identical regardless of the prior information.

Figure 3.6 illustrates the contour plot of information and probability of misclassification in the W_T-W_F plane for non-informative prior ($u = 0.5$), where $W_{(\cdot)}$ indicates the workload variable that determines $\sigma_{(\cdot)}$. Note that the lines with boxed levels indicate the information while the lines with levels indicate the probability of misclassification. Figure 3.7 shows the probability of misclassification as a function of information for different prior information and for $W_T = W_F$, i.e., the workload determines σ_T and σ_F equally. Note that the general trend is that as the amount of information collected increases, the probability of misclassification decreases.

3.3.1.1 A counter example

Figure 3.8 shows a part of the contour region that was shown in Fig. 3.6. Note that there are four markings on the contour (a cross, a square, a circle, and a triangle), each indicating a different level of information and probability of misclassification.

Observation 8. *Beginning at the cross and following the probability of misclassification level curve towards the square, one can notice that although the information has increased, the probability of misclassification remains constant ($P_m = 0.2$, $I = 0.28 \rightarrow P_m = 0.2$, $I = 0.35$).*

Observation 9. *Beginning at the circle and following the information level curve*

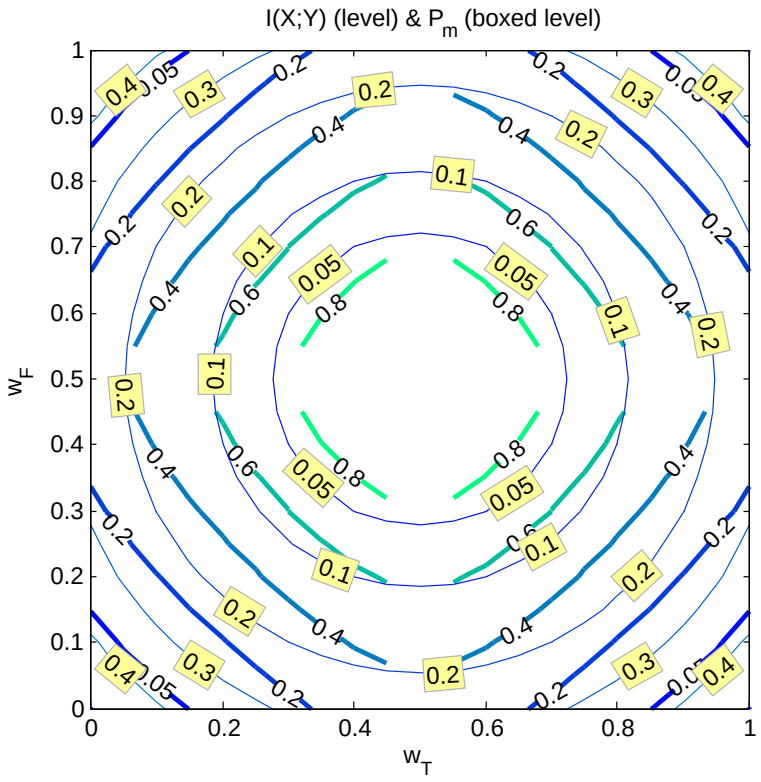


Figure 3.6: Comparison of information (level) and probability of misclassification (boxed level) with respect to workload for $u = 0.5$ and $\sigma^* = 1$

towards the triangle, one can notice that although the information level is constant, the probability of misclassification has increased ($P_m = 0.25$, $I = 0.2 \rightarrow P_m = 0.26$, $I = 0.2$*).*

These are observations that demonstrate that increasing the amount of information collected does *not always* imply improving the classification performance of a workload-dependent classifier.

The mechanism behind the counterintuitive phenomena for a workload-dependent classifier can be thought of similarly as the mechanism behind the counterintuitive phenomena for a workload-independent classifier; While the mechanism for the workload-independent classifier is due to the trade-off between the sensor performances, σ_T and σ_F, we conjecture that the mechanism for the workload-dependent

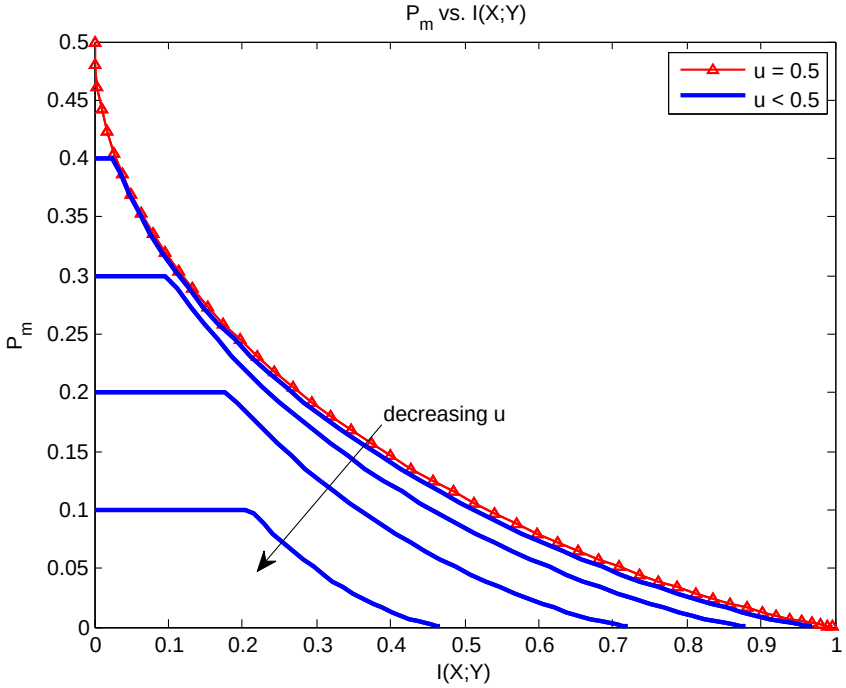


Figure 3.7: Probability of misclassification (P_m) vs. information ($I(X;Y)$). The prior information (u) is varied from 0.5 to 0.1 with decrements of 0.1.

classifier is due to some trade-off between the workload variables, W_T and W_F. Studying the analytical properties and clarifying the mechanism behind the counterintuitive phenomena for the workload-dependent classifier are left as future work.

3.4 Conclusion & future work

In this chapter, we studied the relationship between the amount of information, in the sense of Shannon, and the classification performance, where the classification decision is made by maximum likelihood rule and the performance is assessed by the probability of misclassification. We investigated the relationship for classifiers under two models, one with workload-independent performance and another with workload-dependent performance. We found that increasing the amount of information generally implies improving classification performance for both classifiers. However, we

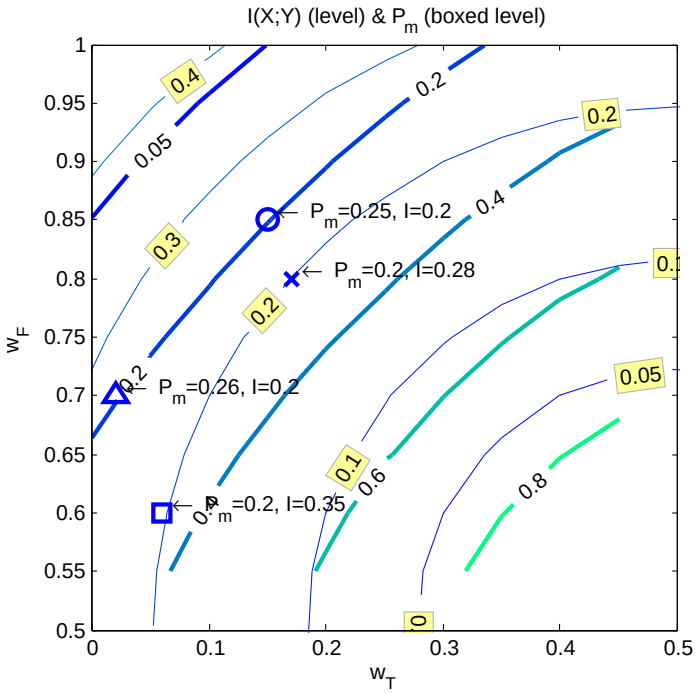


Figure 3.8: Comparison of information (level) and probability of misclassification (boxed level) with respect to workload for $u = 0.5$ and $\sigma^* = 1$ (zoomed Fig. 3.6)

have observed that there are cases when such a general trend is violated for both classifiers, i.e., increasing the amount of information does not always imply improving classification performance.

While it is understandable that increasing the amount of information does not always imply improving classification for classifiers with workload-dependent performance, it is counterintuitive that workload-independent classifiers also exhibit such a phenomenon. We found that the cause of a congruence between the amount of information and the classification performance is simultaneous improvements in the rates of true positives and true negatives of the classifier. On the other hand, we found that the cause of the counterintuitive phenomena is a trade-off between the rates of true positives and true negatives of the classifier.

It is unclear whether such counterintuitive phenomena exist for classifiers with

different decision mechanisms (such as utility-theoretic decision making), or for a multiple classifiers (with parallel or nested structure), or for classifiers using multiple measurements. Moreover, it is obscure how significant such counterintuitive phenomena can be without an engineering application; the significance of the phenomena may vary depending on the application. Investigation of these avenues is left as future work.

CHAPTER IV

A Single Classifier

> Men have an extraordinarily erroneous opinion of their position in nature; and the error is ineradicable.
>
> W. Somerset Maugham

In this chapter, we study the mechanism of a single classifier. As a benchmark, we begin with a classifier with dichotomy (true or false) when a single-variable measurable property is given. Based upon the benchmark example, we generalize to a classifier with trichotomy (true, false or *unknown*) when a single-variable measurable property is provided, and then revisit the two classifiers and generalize the formalization to the multivariate measurement cases.

From Chap. III, we have learned that there are three key parameters in defining a classifier, which are the rates of true positives and true negatives, σ_T and σ_F, and the prior information u. We assume that these parameters are known to us by a prior calibration process. Since the workload-dependency can be considered as a specification of a workload-independent (constant performance with respect to the workload) classifier, we consider only the latter (more general) case in this chapter.

4.1 The problem of thresholding

Assume that a property $w \in \mathcal{R}$ can be measured from each member of a population of objects of interest where the population comprises two disjoint sub-populations, T and F. Each sub-population is characterized by its own distribution of w. Assume that the two distributions are distinct such that if a proper threshold is applied, a classifier can distinguish one sub-population from another based on a measurement of w. Once a threshold is determined, measurement values on one side of the threshold are labeled as originating from a T object while properties on the other side are labeled as originating from an F object.

We consider two types of workload-independent classifiers: 1. one where the classification decision is based on two options (dichotomous), 2. the other where the decision is based on three options (trichotomous). Figure 4.1 illustrates the concept of such classifiers.

4.1.1 Dichotomous thresholding

We assume that the distribution of the measurable property w in each sub-population is a Gaussian probability density function (pdf),

$$p_T \sim \mathcal{N}(m_T, s_T^2), \tag{4.1a}$$

$$p_F \sim \mathcal{N}(m_F, s_F^2), \tag{4.1b}$$

where m_i is the mean and s_i, $i \in \{T, F\}$ is the standard deviation of the distribution. For the distinctness of the two distributions, we further assume that $m_T < m_F$ without loss of generality. Let $\tau \in \mathcal{R}$ be the threshold variable. For a classifier that

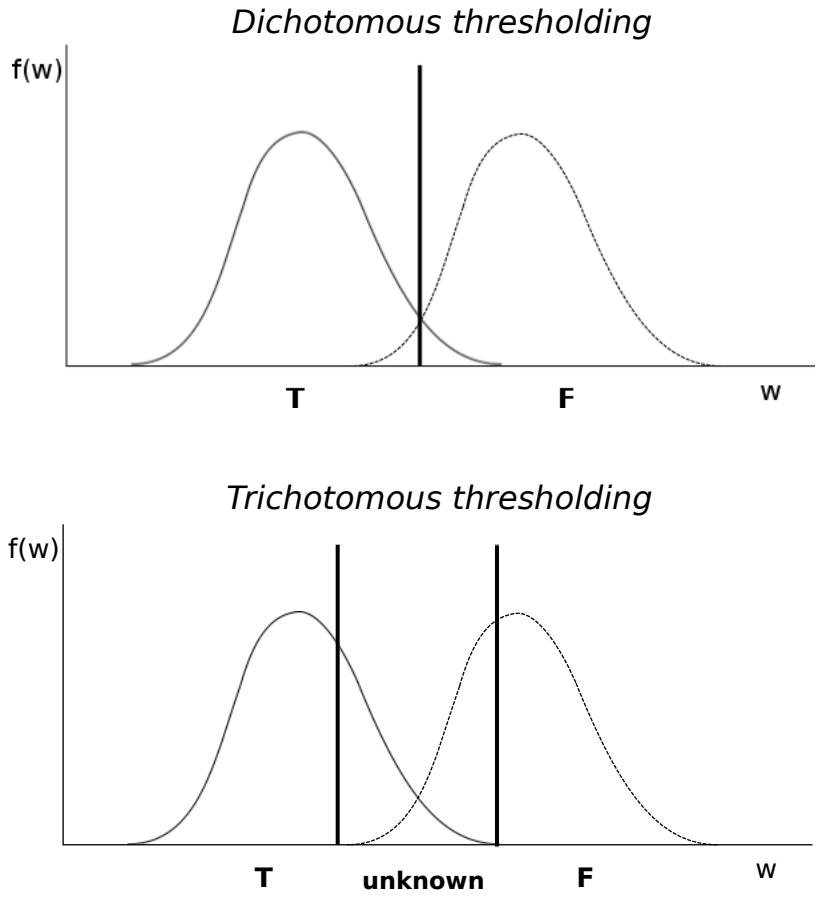


Figure 4.1: Concepts of dichotomous and trichotomous thresholding

uses thresholding, the rates of true positives and negatives are evaluated as:

$$\sigma_T = \int_{-\infty}^{\tau} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{4.2a}$$

$$\sigma_F = \int_{\tau}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{4.2b}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, and $c_i = \sqrt{2s_i^2}$, $i \in \{T, F\}$. The analytical solutions to Eq. (4.2) can be found in Appendix C.

The cost function is the probability of misclassification, i.e.,

$$\begin{aligned} P_m =& \delta_T(f_1)\sigma_F(1-u) + \delta_T(f_2)(1-\sigma_F)(1-u) \\ &+ \delta_F(f_1)(1-\sigma_T)u + \delta_F(f_2)\sigma_T u, \end{aligned} \tag{4.3}$$

where the definitions of u, δ_T, δ_F, f_1, and f_2 can be found in Chap. III. The objective is to determine the optimal threshold that minimizes the probability of misclassification, i.e,

$$\min_{\tau \in \mathcal{R}} P_m.$$

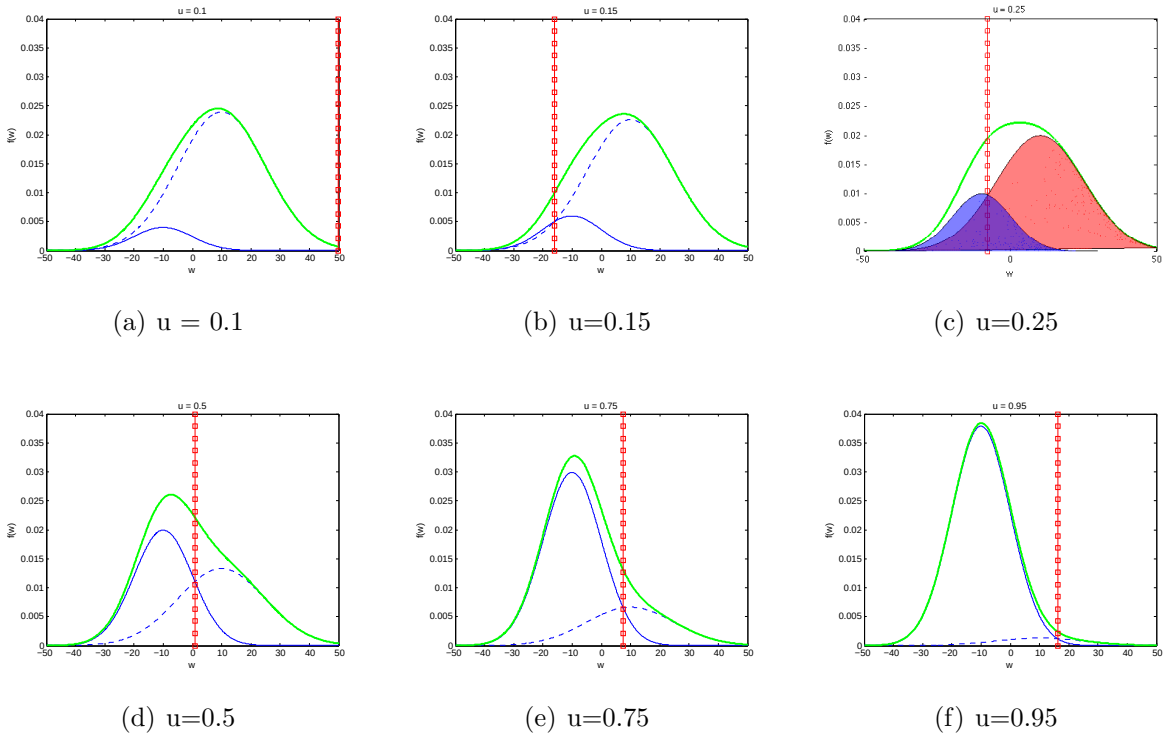
4.1.1.1 Optimal dichotomous thresholding

There are numerical studies on dichotomous classification documented in the literature (see, e.g., [123]). Here, we provide both analytical and numerical results on the subject for completeness of the dissertation.

The solution to the problem of dichotomous classification is stated in the following theorem.

Theorem IV.1. *Assume that a property $w \in \mathcal{R}$ can be measured from a population of objects of interest where the population comprises two disjoint sub-populations, T and F. Each sub-population is characterized by its own distribution of w. Assume that there are prior probabilities that quantify the proportion of T and F objects among the objects of interest. If the distribution of w for each sub-population is Gaussian, then the optimal dichotomous threshold is always at the intersection of the two distributions weighted by their prior probabilities.*

Proof. The key step is to prove that the probability of misclassification in Eq. (4.3), where σ_T and σ_F are defined in Eq. (4.2), is a *differentiable* function of τ. Then the result is obtained by differentiation. The differentiability of P_m with respect to τ is proven as follows.



(a) u = 0.1 (b) u=0.15 (c) u=0.25

(d) u=0.5 (e) u=0.75 (f) u=0.95

Figure 4.2: Thresholding with varying prior information on weighted distribution functions. Blue solid line indicates a distribution with $m_T = -10$, $s_T = 10$ weighted by u, blue dashed line indicates a distribution with $m_F = 10$, $s_F = 15$ weighted by $1 - u$, green thick line indicates the sum of the two distributions weighted by their prior information, and red vertical line indicates the optimal threshold.

Depending on f_1 and f_2, P_m is expressed as the following:

$$P_m = \begin{cases} \sigma_F(1-u) + (1-\sigma_F)(1-u) & \text{if } f_1 > 1 \wedge f_2 > 1, \\ (1-\sigma_T)u + (1-\sigma_F)(1-u) & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ \sigma_F(1-u) + \sigma_T u & \text{if } f_1 > 1 \wedge f_2 \leq 1, \\ (1-\sigma_T)u + \sigma_T u & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{4.4}$$

Since $f_1 > 1 \wedge f_2 \leq 1$ is false for all σ_T, σ_F, and u (see Corollary 2 in Appendix B),

the third condition can be ignored. With some rearrangement, P_m is expressed as

$$P_m = \begin{cases} 1-u & \text{if } f_1 > 1 \wedge f_2 > 1, \\ (1-\sigma_T)u + (1-\sigma_F)(1-u) & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ u & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{4.5}$$

For the first and third conditions in Eq. (4.5), u is a constant with respect to τ. For the second condition, substituting Eq. (4.2) into Eq. (4.5) and noting that

$$1 - \int_{-\infty}^{\tau} p_T dw = \int_{\tau}^{\infty} p_T dw, \tag{4.6a}$$

$$1 - \int_{\tau}^{\infty} p_F dw = \int_{-\infty}^{\tau} p_F dw, \tag{4.6b}$$

the second condition yields,

$$P_m = u \int_{\tau}^{\infty} p_T dw + (1-u) \int_{-\infty}^{\tau} p_F dw, \tag{4.7}$$

which is a differentiable function with respect to τ since it is the sum of two differentiable functions with respect to τ. Therefore, P_m is differentiable with respect to τ. □

Figure 4.2 shows some numerical examples that illustrate the property highlighted in Theorem IV.1.

Remark IV.2. For dichotomous thresholding, P_m is represented by the sum of the "misclassification regions" of the weighted distributions, i.e., Eq. (4.7). By "misclassification regions", we mean the region under the blue dashed line on the left-side of the threshold and the region under the blue solid line on the right-side of the threshold in Fig. 4.2 (c), which appears as purple region.

4.1.2 Trichotomous thresholding

Dichotomous classification corresponds to classical propositional logic where a proposition can either be *true* or *false*. Now, allowing a third status, trichotomous classification corresponds to ternary logic where a proposition can be *unknown* in addition to true or false. The reason we allow the unknown status is that there are cases when dichotomous classifiers are unsatisfactory. For example, the distributions of the sub-populations may not be easily distinguishable. Trichotomous classification can be formalized by extending the notion of dichotomous classification using two thresholds.

Let $\tau_1 \in \mathcal{R}$ and $\tau_2 \in \mathcal{R}$ be the threshold variables such that the cumulative probability distributions are,

$$\sigma_T = \int_{-\infty}^{\tau_1} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{4.8a}$$

$$\sigma_F = \int_{\tau_2}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{4.8b}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, and $c_i = \sqrt{2s_i^2}$ with $i \in \{T, F\}$. The analytical solutions to Eq. 4.8 can be found in Appendix C.

Let us define the range on w between the two thresholds where the classifier is unable to decide as *the region of indecision*, i.e., $[\tau_1, \tau_2]$.

Let P be a pre-specified probability of misclassification that is determined by the mission specification. The objective is to determine the optimal thresholds that minimize the size of the region of indecision, i.e.,

$$\min_{\tau_1, \tau_2} |\tau_2 - \tau_1|,$$

subject to constraints,

$$P_m = P, \tag{4.9a}$$

$$\tau_1 \leq \tau_2. \tag{4.9b}$$

Similarly, the objective of the optimization problem can be formalized as

$$\min_{\tau_1,\tau_2}(\tau_2 - \tau_1)^2,$$

subject to the constraints in Eq. (4.9).

4.1.2.1 Optimal trichotomous thresholding

At minimum τ_1^* and τ_2^*, the problem must satisfy the Karesh-Kuhn-Tucker (K-K-T) conditions [125], i.e.,

$$\frac{\partial}{\partial\tau_1}|\tau_2 - \tau_1| + \lambda_1\frac{\partial}{\partial\tau_1}(P_m - P) + \mu_1\frac{\partial}{\partial\tau_1}(\tau_1 - \tau_2) = 0, \tag{4.10}$$

$$\frac{\partial}{\partial\tau_2}|\tau_2 - \tau_1| + \lambda_1\frac{\partial}{\partial\tau_2}(P_m - P) + \mu_1\frac{\partial}{\partial\tau_2}(\tau_1 - \tau_2) = 0. \tag{4.11}$$

For taking the partial derivative of an absolute value, as present in the first term, generalized gradients, such as Fréchet [126] or Gâteaux [127, 128] derivatives, can be used. However, we defer investigating the analytical properties of a trichotomous classifier to future work.

An easier fix to the smoothing issue of the derivative of $|\tau_2 - \tau_1|$ is by reformulating the cost function to $(\tau_2 - \tau_1)^2$. Here we provide the necessary conditions for such a cost function.

At minimum τ_1^* and τ_2^*, the problem must satisfy the K-K-T conditions, i.e.,

$$\frac{\partial}{\partial \tau_1}(\tau_2 - \tau_1)^2 + \lambda_1 \frac{\partial}{\partial \tau_1}(P_m - P) + \mu_1 \frac{\partial}{\partial \tau_1}(\tau_1 - \tau_2) = 0, \tag{4.12}$$

$$\frac{\partial}{\partial \tau_2}(\tau_2 - \tau_1)^2 + \lambda_1 \frac{\partial}{\partial \tau_2}(P_m - P) + \mu_1 \frac{\partial}{\partial \tau_2}(\tau_1 - \tau_2) = 0, \tag{4.13}$$

where λ_1 and $\mu_1 \geq 0$ are Lagrange multipliers. Reformulating the K-K-T conditions in Eq. (4.12), we get

$$\begin{cases} 2\tau_1^* - 2\tau_2^* + \mu_1 = 0 & \text{if } f_1 > 1 \wedge f_2 > 1, \\ 2\tau_1^* - 2\tau_2^* - \lambda_1 a_T u e^{-(\tau_1^* + b_T)^2/c_T^2} + \mu_1 = 0 & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ 2\tau_1^* - 2\tau_2^* + \mu_1 = 0 & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{4.14}$$

Reformulating the K-K-T conditions in Eq. (4.13), we get

$$\begin{cases} -2\tau_1^* + 2\tau_2^* - \mu_1 = 0 & \text{if } f_1 > 1 \wedge f_2 > 1, \\ -2\tau_1^* + 2\tau_2^* + \lambda_1 a_F (1-u) e^{-(\tau_2^* + b_F)^2/c_F^2} - \mu_1 = 0 & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ -2\tau_1^* + 2\tau_2^* - \mu_1 = 0 & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{4.15}$$

Summing the equations for $f_1 \leq 1 \wedge f_2 > 1$ in Eq. (4.14)-(4.15) yields,

$$-\lambda_1 a_T u e^{-(\tau_1^* + b_T)^2/c_T^2} + \lambda_1 a_F (1-u) e^{-(\tau_2^* + b_F)^2/c_F^2} = 0. \tag{4.16}$$

With some rearrangement, we get

$$\frac{(\tau_1^* + b_T)^2}{c_T^2} - \frac{(\tau_2^* + b_F)^2}{c_F^2} = \log\left\{\frac{a_T}{a_F}\frac{u}{1-u}\right\}, \tag{4.17}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, $c_i = \sqrt{2s_i^2}$ with $i \in \{T, F\}$.

The optimization problem of trichotomous classification is solved numerically by using the MATLAB optimization command, *fmincon*. Figure 4.3 illustrates some

numerical examples of the optimal threshold.

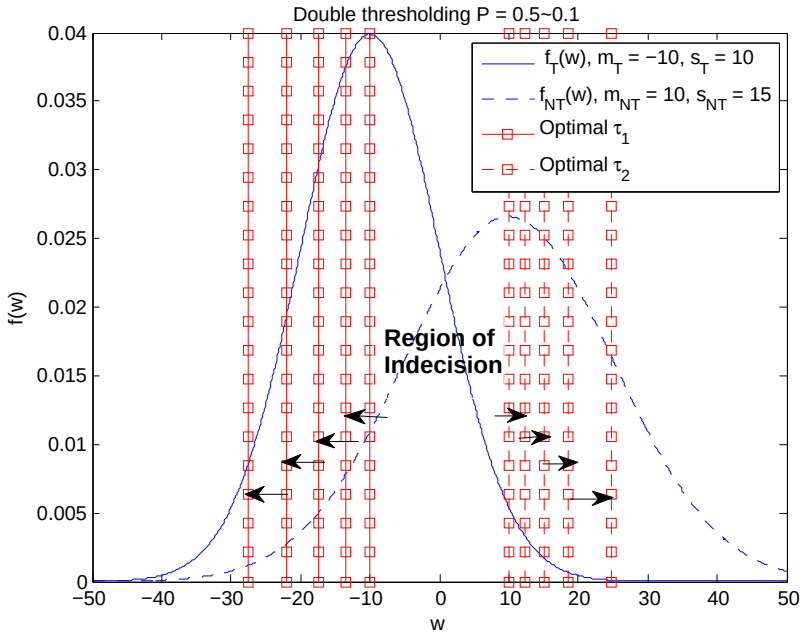


Figure 4.3: An example of trichotomous classification for the mission specification $P = 0.5 \sim 0.1$ with decrements of $\Delta P = 0.1$

One can notice that as the specification for the probability of misclassification is constrained, the region of indecision is widened. If a sampled measurement falls within the region of indecision, trichotomous classifiers are not able to make reliable decisions. Thus, for highly constrained specifications, the sole usage of trichotomous classifiers is not beneficial for the purpose of improving the classification performance. This fact encourages the use of secondary dichotomous classifiers as collaborative teammates. This will be revised in Chap. VI.

4.2 The problem of linear thresholding

In this section, we consider a case for a multivariate measurement and extend the notion of thresholding accordingly. It is important to consider such case since, in practice, it is rare to rely on a single variable measurement while making a classifi-

cation decision. Often, there are multiple sensors that provide several measurements with respect to an object of interest, and the classification decision is made upon multiple sources of measurements.

The problem we face here is similar to what ADALIN [60–62], by B. Widrow, had demonstrated its capability in. Indeed, there are similarities between ADALIN and our approach. For example, in both approaches the decision is determined by a linear combination of some variables, in ADALIN's case a linear combination of inputs and weighting variables. However, the key difference is that while ADALIN requires a set of data that is correctly labeled with its class (by some oracle) *a priori*, our approach does not require such information, but requires the statistical properties of the measurements.

4.2.1 Problem formulation

Let $\mathbf{w} \in \mathcal{R}^n$ be some properties that can be measured from a population of objects of interest where the population comprises two disjoint sub-populations, T and F. Each sub-population is characterized by its own distribution of $\mathbf{w}$. If the distribution of $\mathbf{w}$ for each sub-population is Gaussian, then

$$\mathbf{w_i} \sim \mathcal{N}(\bar{\mathbf{w}}_\mathbf{i}, P_{w_i}),\ i \in \{T, F\}, \tag{4.18}$$

where $\bar{\mathbf{w}}$ is the mean and P_w is the covariance matrix of $\mathbf{w}$.

Let $\mathbf{c} \in \mathcal{R}^n$ be such that it satisfies the constraint:

$$\mathbf{c}^T\mathbf{c} = 1. \tag{4.19}$$

Since $\mathbf{w}$ is a Gaussian random variable, the inner product $\mathbf{c}^T\mathbf{w}$ is also a Gaussian

random variable [129]. Specifically,

$$\mathbf{c}^T\mathbf{w}_i \sim \mathcal{N}(\mathbf{c}^T\bar{\mathbf{w}}_i, \mathbf{c}^T P_{w_i}\mathbf{c}),\ i \in \{T, F\}. \tag{4.20}$$

The key idea is to recognize that $\mathbf{c}^T\mathbf{w}$ is a scalar Gaussian random variable so that the thresholding approach for a single variable (in Sec. 4.1) is still applicable. The difference is that, while the classification decision is made by a single variable threshold in the previous approach, the decision for multivariate measurements is determined by a single variable threshold and a multivariate parameter $\mathbf{c}$.

We use the term *sieving parameter* to denote $\mathbf{c}$, recognizing the filtering role of such a parameter in the problem.

4.2.2 Linear dichotomous thresholding

The distribution of the measurable properties $\mathbf{w} \in \mathcal{R}^n$ of each sub-population is assumed Gaussian and is given as,

$$\mathbf{p}_T \sim \mathcal{N}(\bar{\mathbf{w}}_T, P_{w_T}), \tag{4.21a}$$

$$\mathbf{p}_F \sim \mathcal{N}(\bar{\mathbf{w}}_F, P_{w_F}). \tag{4.21b}$$

Let $w = \mathbf{c}^T\mathbf{w}$ denote the sieved measurement. Consequently, the distributions of the sieved measurement $w \in \mathcal{R}$ of each sub-population are characterized as,

$$\mathbf{c}^T\mathbf{p}_T \sim \mathcal{N}(\mathbf{c}^T\bar{\mathbf{w}}_T, \mathbf{c}^T P_{w_T}\mathbf{c}), \tag{4.22a}$$

$$\mathbf{c}^T\mathbf{p}_F \sim \mathcal{N}(\mathbf{c}^T\bar{\mathbf{w}}_F, \mathbf{c}^T P_{w_F}\mathbf{c}). \tag{4.22b}$$

Let $\tau \in \mathcal{R}$ denote the threshold variable. Then, the cumulative probability distributions are determined as,

$$\sigma_T = \int_{-\infty}^{\tau} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{4.23a}$$

$$\sigma_F = \int_{\tau}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{4.23b}$$

where $a_i = 1/\sqrt{2\pi\left(\mathbf{c}^T P_{w_i}\mathbf{c}\right)}$, $b_i = -\mathbf{c}^T\bar{\mathbf{w}}_i$, $c_i = \sqrt{2\left(\mathbf{c}^T P_{w_i}\mathbf{c}\right)}$ with $i \in \{T, F\}$. The objective is to minimize the probability of misclassification P_m by choosing the threshold variable τ and the sieving parameter $\mathbf{c}$ simultaneously, i.e.,

$$\min_{\tau,\mathbf{c}} P_m(\tau, \mathbf{c}),$$

subject to an equality constraint,

$$\mathbf{c}^T\mathbf{c} = 1.$$

4.2.2.1 Optimal linear dichotomous thresholding

Let τ^* and $\mathbf{c}^*$ be the optimal threshold variable and the optimal sieving parameter obtained by solving the optimization problem. Then, the following condition holds for the optimal linear dichotomous threshold, i.e.,

$$\left(\mathbf{c}^*\right)^T \mathbf{w} = \tau^*. \tag{4.24}$$

For the optimal sieving parameter $\mathbf{c}^*$, the distribution of sieved measurement $w^* = \left(\mathbf{c}^*\right)^T \mathbf{w}$ of each sub-population is characterized by,

$$\left(\mathbf{c}^*\right)^T \mathbf{p}_T \sim \mathcal{N}(\left(\mathbf{c}^*\right)^T \bar{\mathbf{w}}_T, \left(\mathbf{c}^*\right)^T P_{w_T}\mathbf{c}^*), \tag{4.25a}$$

$$\left(\mathbf{c}^*\right)^T \mathbf{p}_F \sim \mathcal{N}(\left(\mathbf{c}^*\right)^T \bar{\mathbf{w}}_F, \left(\mathbf{c}^*\right)^T P_{w_F}\mathbf{c}^*). \tag{4.25b}$$

At minimum τ^*, the problem must satisfy the K-K-T conditions, i.e.,

$$\frac{\partial}{\partial \tau} P_m + \lambda_1 \frac{\partial}{\partial \tau}(\mathbf{c}^T\mathbf{c} - 1) = 0, \tag{4.26}$$

$$\frac{\partial}{\partial \mathbf{c}} P_m + \lambda_1 \frac{\partial}{\partial \mathbf{c}}(\mathbf{c}^T\mathbf{c} - 1) = 0. \tag{4.27}$$

By the K-K-T conditions in Eq. (4.26), we get

$$-a_T u e^{-(\tau^* + b_T)^2/c_T^2} + a_F(1-u)e^{-(\tau^* + b_F)^2/c_F^2} = 0. \tag{4.28}$$

Using the chain-rule, Eq. (4.27) yields,

$$\frac{\partial P_m}{\partial w}\frac{\partial w}{\partial \mathbf{c}} + \lambda_1 \frac{\partial}{\partial \mathbf{c}}(\mathbf{c}^T\mathbf{c} - 1) = 0. \tag{4.29}$$

By the K-K-T conditions in Eq. (4.27), we get

$$\begin{bmatrix} 2\lambda_1 c_1 \\ 2\lambda_1 c_2 \end{bmatrix} = 0_{2\times 1}. \tag{4.30}$$

We use the MATLAB numerical solver, *fmincon*. Figures 4.4 and 4.5 show a numerical example of optimal linear dichotomous thresholding for $\mathbf{w} \in \mathcal{R}^2$. Figure 4.4 shows the distribution of $\mathbf{w}$ for each sub-population and the optimal threshold variable and sieving parameter. Figure 4.5 shows the distribution of $w^* = \mathbf{c}^*\mathbf{w}$ for each sub-population and the optimal threshold variable. Note that once the multivariate property $\mathbf{w}$ is sieved by $\mathbf{c}^*$, the optimal threshold τ^* is at the intersection between the two distributions weighted by its prior probability. The result agrees with the single variable property case shown in Sec. 4.2.2.

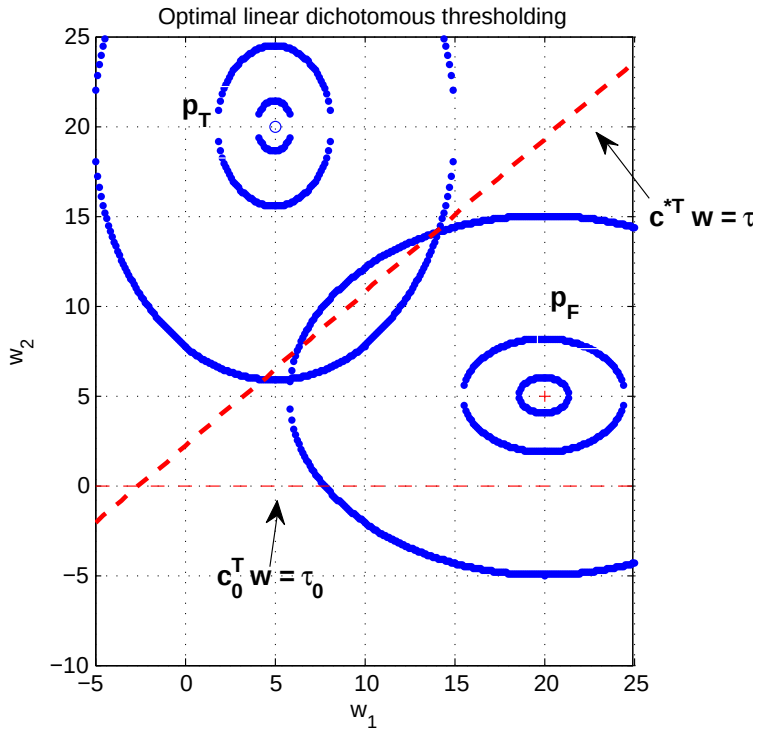


Figure 4.4: Optimal linear dichotomous thresholding for $\bar{\mathbf{w}}_T = [5,\ 20]$, $\bar{\mathbf{w}}_F = [20,\ 5]$, $P_{w_T} = \mathrm{diag}(10, 5)$, $P_{w_F} = \mathrm{diag}(5, 10)$, $\mathbf{c} = [0,\ 1]$, $\tau_0 = 0$ $(u = 0.5)$

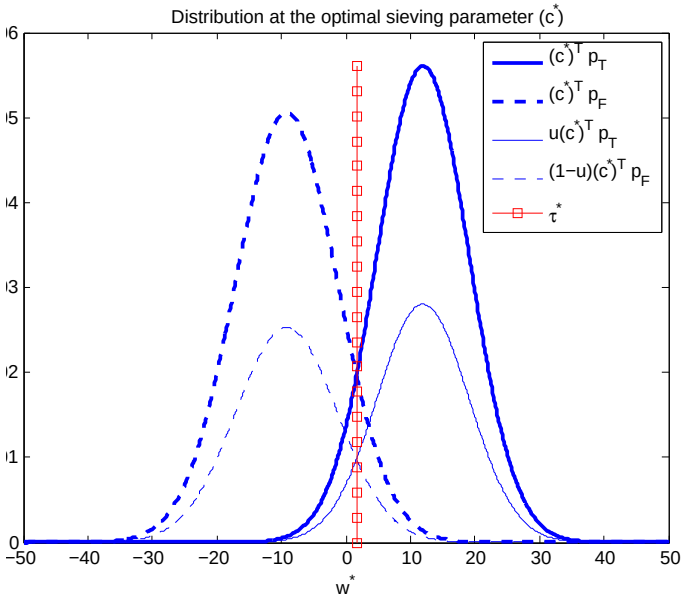


Figure 4.5: Distribution of w^* $(u = 0.5)$

4.2.3 Linear trichotomous thresholding

We extend the notion of linear thresholding to trichotomous classification. Let τ_1 and τ_2 be the threshold variables. Let the distribution of the measurable properties $\mathbf{w} \in \mathcal{R}^n$ of each sub-population be defined as in Eq. (4.21) and the distributions of the sieved measurement $w \in \mathcal{R}$ of each sub-population be defined as in Eq. (4.22). Let τ_1 and τ_2 be the threshold variables. Then, the cumulative probability distributions are determined by,

$$\sigma_T = \int_{-\infty}^{\tau_1} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{4.31a}$$

$$\sigma_F = \int_{\tau_2}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{4.31b}$$

where $a_i = 1/\sqrt{2\pi\left(\mathbf{c}^T P_{w_i} \mathbf{c}\right)}$, $b_i = -\mathbf{c}^T \bar{\mathbf{w}}_i$, $c_i = \sqrt{2\left(\mathbf{c}^T P_{w_i} \mathbf{c}\right)}$ with $i \in \{T, F\}$.

Let P be a pre-specified probability of misclassification. The objective is to minimize the region of indecision, i.e., $[\tau_1, \tau_2]$, by choosing the threshold variables and the sieving parameter,

$$\min_{\tau_1, \tau_2, \mathbf{c}} |\tau_2 - \tau_1|,$$

subject to constraints,

$$P_m = P, \tag{4.32a}$$

$$\mathbf{c}^T\mathbf{c} = 1, \tag{4.32b}$$

$$\tau_1 \leq \tau_2, \tag{4.32c}$$

$$\sigma_T \geq 0.5, \tag{4.32d}$$

$$\sigma_F \geq 0.5, \tag{4.32e}$$

$$\sigma_T \leq 1, \tag{4.32f}$$

$$\sigma_F \leq 1. \tag{4.32g}$$

Similarly, the objective of the optimization problem can be formalized as

$$\min_{\tau_1,\tau_2,\mathbf{c}} (\tau_2 - \tau_1)^2,$$

subject to the constraints in Eq. (4.32).

4.2.3.1 Optimal linear trichotomous thresholding

At minimum τ_1^*, τ_2^* and $\mathbf{c}^*$, the problem must satisfy the K-K-T conditions, i.e.,

$$\begin{aligned}
&\frac{\partial}{\partial \tau_1}|\tau_2 - \tau_1| + \lambda_1 \frac{\partial}{\partial \tau_1}(P_m - P) + \lambda_2 \frac{\partial}{\partial \tau_1}(\mathbf{c}^T\mathbf{c} - 1) \\
&+ \mu_1 \frac{\partial}{\partial \tau_1}(\tau_1 - \tau_2) + \mu_2 \frac{\partial}{\partial \tau_1}(0.5 - \sigma_T) + \mu_3 \frac{\partial}{\partial \tau_1}(0.5 - \sigma_F) \\
&+ \mu_4 \frac{\partial}{\partial \tau_1}(\sigma_T - 1) + \mu_5 \frac{\partial}{\partial \tau_1}(\sigma_F - 1) = 0, \qquad (4.33) \\
&\frac{\partial}{\partial \tau_2}|\tau_2 - \tau_1| + \lambda_1 \frac{\partial}{\partial \tau_2}(P_m - P) + \lambda_2 \frac{\partial}{\partial \tau_2}(\mathbf{c}^T\mathbf{c} - 1) \\
&+ \mu_1 \frac{\partial}{\partial \tau_2}(\tau_1 - \tau_2) + \mu_2 \frac{\partial}{\partial \tau_2}(0.5 - \sigma_T) + \mu_3 \frac{\partial}{\partial \tau_2}(0.5 - \sigma_F) \\
&+ \mu_4 \frac{\partial}{\partial \tau_2}(\sigma_T - 1) + \mu_5 \frac{\partial}{\partial \tau_2}(\sigma_F - 1) = 0, \qquad (4.34) \\
&\frac{\partial}{\partial \mathbf{c}}|\tau_2 - \tau_1| + \lambda_1 \frac{\partial}{\partial \mathbf{c}}(P_m - P) + \lambda_2 \frac{\partial}{\partial \mathbf{c}}(\mathbf{c}^T\mathbf{c} - 1) \\
&+ \mu_1 \frac{\partial}{\partial \mathbf{c}}(\tau_1 - \tau_2) + \mu_2 \frac{\partial}{\partial \mathbf{c}}(0.5 - \sigma_T) + \mu_3 \frac{\partial}{\partial \mathbf{c}}(0.5 - \sigma_F) \\
&+ \mu_4 \frac{\partial}{\partial \mathbf{c}}(\sigma_T - 1) + \mu_5 \frac{\partial}{\partial \mathbf{c}}(\sigma_F - 1) = 0. \qquad (4.35)
\end{aligned}$$

As pointed out previously, the partial derivative of an absolute value can be performed by generalized gradients, such as Fréchet [126] or Gâteaux [127, 128] derivatives. However, we defer investigating the analytical properties of a linear trichotomous classifier to future work.

We use the MATLAB numerical solver, *fmincon*. Figures 4.6 and 4.7 illustrate a numerical example of optimal linear trichotomous thresholding for $\mathbf{w} \in \mathcal{R}^2$. Figure 4.6 shows the distribution of $\mathbf{w}$ for each sub-population and the optimal threshold variable and sieving parameter. Figure 4.7 shows the distribution of $w^* = \mathbf{c}^*\mathbf{w}$ for each sub-population and the optimal threshold variable.

Note that with *fmincon* the solution is very sensitive to the initial conditions. Either the solver finds a solution near the initial condition or an infeasible solution.

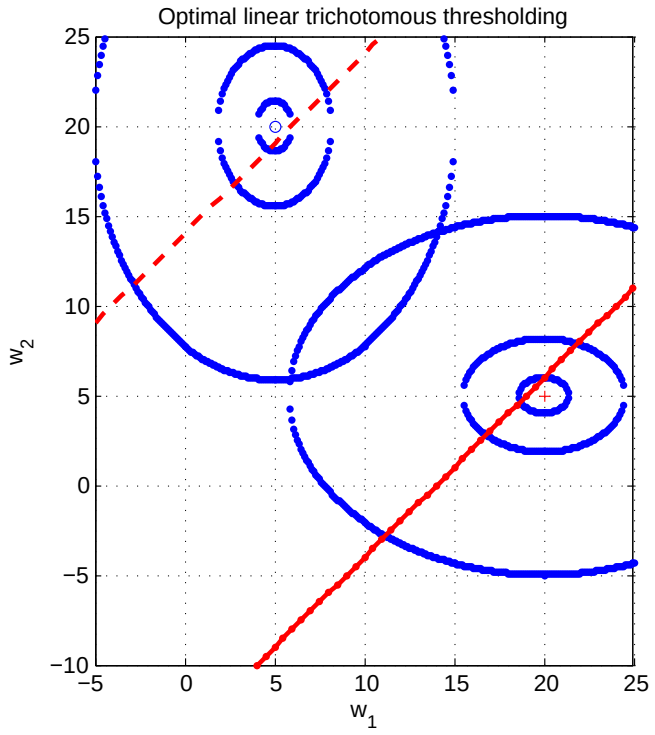


Figure 4.6: Optimal linear trichotomous thresholding for $\bar{\mathbf{w}}_T = [5,\ 20]$, $\bar{\mathbf{w}}_F = [20,\ 5]$, $P_{w_T} = \mathrm{diag}(10, 5)$, $P_{w_F} = \mathrm{diag}(5, 10)$, $\mathbf{c}_0 = [\frac{1}{\sqrt{2}},\ -\frac{1}{\sqrt{2}}]$, $\tau_0 = [-10,\ 10]$

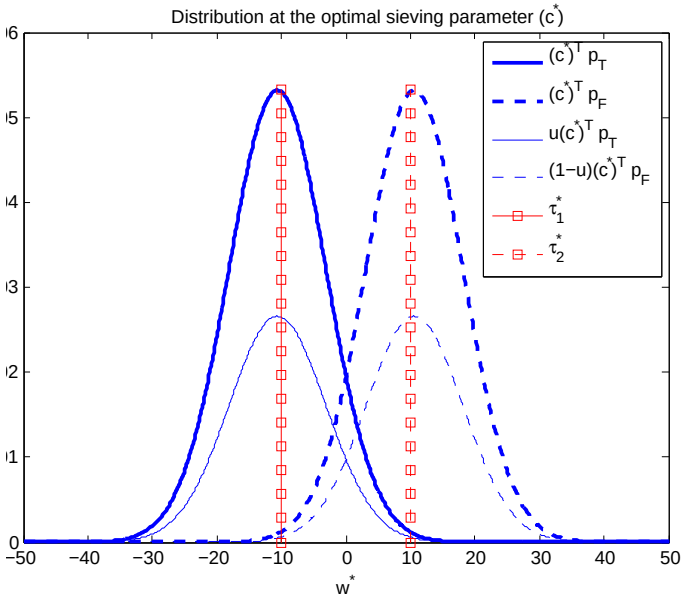


Figure 4.7: Distribution of w^*

4.3 Conclusion & future work

In this chapter, we have studied the operation of a single classifier when the classification decision is based on either dichotomy or trichotomy. We have investigated cases when the measurement is either a single variable property or multivariate. We have formalized a thresholding methodology that provides optimal classification decisions by minimizing the probability of misclassification.

There are several directions that we think deserve further investigation. One is in investigating the relationship between the specification variable P in trichotomic thresholding and the minimal probability of misclassification P_m analytically. Although we have found numerically that there may be a monotonic relationship (either increasing or decreasing) between the two parameters, we have found counterexamples and at this point we are unclear whether the examples are due to the problem itself, or the lack of sophistication of the optimization technique.

Another is extending the work of linear thresholding in Sec. 4.2 to incorporate *learning*. This aspect will be revisited in Chap. VII.

CHAPTER V

A Team of Homogeneous Classifiers

> It is not best that we should all think alike; it is a difference of opinion that makes horse races.
>
> Mark Twain

The purpose of this chapter is to assess the performance of a team of classifiers based on the performance of the individual classifiers in the team, prior information, and fusion rules that combine the individual classifiers' decisions. We define a fusion rule to be *synergistic* if, under this rule, the performance of a homogeneous team of classifiers (i.e., a team consisting of two classifiers with identical properties) is better than the performance of each classifier operating alone. We show that, while some fusion rules are synergistic, others are not. We also show that, depending on the prior information about the objects to be classified, some fusion rules are preferable to others because of synergistic effects.

The number of classifiers that we consider in this study is strictly even. For odd numbers of classifiers, a common strategy to reach a collective decision is by *voting*. A problem formulation and solutions for such case can be found in [98]. In this chapter, we consider a two-classifier team scenario as a benchmark problem. Discussion on extending our framework to even numbers of classifiers is provided in the conclusion.

5.1 Problem formulation

A supervisor is an entity that makes a collective decision based on the opinions of the team members. A consensus happens when all the team members have the same opinion while a dissensus happens when two of the team members differ in their opinions. When there are multiple opinions, a supervisor uses a fusion rule, e.g., voting, so that the team reaches a collective decision. Thus, a problem is to determine the fusion rule ($F.R.$) that minimizes the probability of team misclassification, i.e.,

$$\min_{F.R.} P_m.$$

subject to the individual performance of each team member. The search space size for $F.R.$s is dependent on the number of members in the team. For a two-member team, there are 16 fusion rules, as shown in Table 5.1. For each entry in the truth table, there can be two outcomes, T or F, which implies that there are 16 possible ways of fusing the outcomes of the classifiers.

Table 5.1: Truth table for two-classifier team

		Classifier A	
		T	F
Classifier B	T	$\{T, F\}$	$\{T, F\}$
	F	$\{T, F\}$	$\{T, F\}$

5.1.1 Performance of a single classifier

Let X be a discrete random variable which denotes the status of an unidentified object such that $X \in \{T, F\}$. Let A be a discrete random variable which denotes the classifier (or operator) decision such that $A \in \{T, F\}$. The conditional probabilities

of a classifier making a decision given a certain object status are,

$$P(A = T|X = F) = 1 - \sigma_F, \tag{5.1a}$$

$$P(A = F|X = F) = \sigma_F, \tag{5.1b}$$

$$P(A = F|X = T) = 1 - \sigma_T, \tag{5.1c}$$

$$P(A = T|X = T) = \sigma_T. \tag{5.1d}$$

Equation (5.1) provides the confusion matrix of a classifier expressed in conditional probabilities.

Let $P(X = T) = u$ and $P(X = F) = 1 - u$ where $u \in [0, 1]$ denotes the prior information of target population. The probability of misclassification is the sum of probabilities of two faulty outcomes: false positive and false negative:

$$P_m = P(A = T \wedge X = F) + P(A = F \wedge X = T). \tag{5.2}$$

Using the product rule yields

$$\begin{aligned} P_m &= (1 - \sigma_F)(1 - u) + (1 - \sigma_T)u \\ &= (1 - \sigma_F) + u(\sigma_F - \sigma_T). \end{aligned} \tag{5.3}$$

Figure 5.1 illustrates the probability of misclassification as a function of the prior information.

As predicted by Eq. (5.3), the model tells us that for equal true positive and negative rates (*e.g.*, $\sigma_T = \sigma_F = 0.7$), the probability of misclassification is insensitive with respect to the prior information. For larger true negative rates than false positive rates (*e.g.*, $\sigma_T = 0.7$, $\sigma_F = 0.8$), increasing the target population linearly increases the probability of misclassification, and *vice versa*.

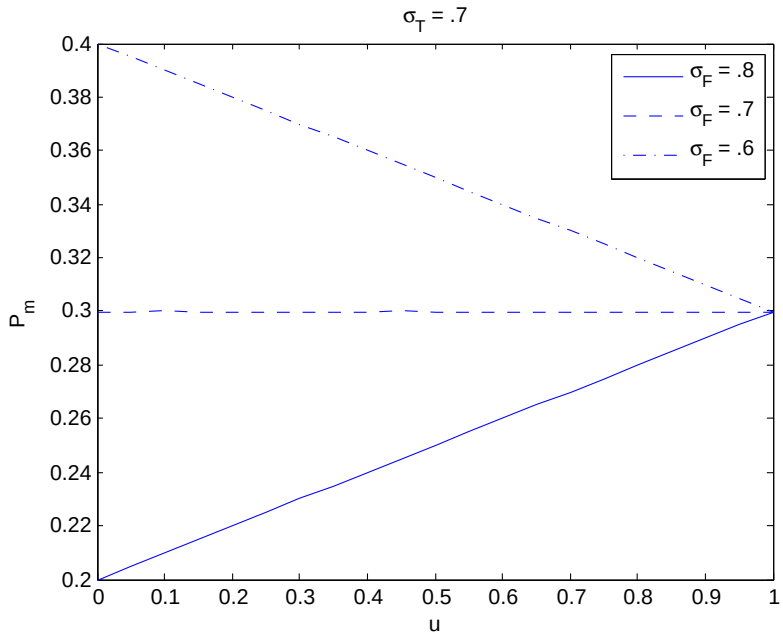


Figure 5.1: Single classifier performance with respect to varying true positive rates (σ_F)

5.1.1.1 Weighted performance

Each term in the performance measure can be weighted as,

$$P_m = \omega_F(1-\sigma_F)(1-u) + \omega_T(1-\sigma_T)u, \tag{5.4}$$

where ω_F, $\omega_T \in \mathcal{R}$ denote the weighting parameters.

These weighting parameters can be determined based on some external information. For instance, in military operations, the weighting parameters are determined by a policy maker who assesses the potential outcomes of certain instances. Also, the weighting parameters can be exploited for a team of heterogenous classifiers. For example, a misclassification by a novice classifier may be weighted less than that of an expert. In this dissertation, we assume that $\omega_F = \omega_T = 1$.

5.1.2 Supervisory decisions

Consider a supervisor which is an entity that makes the final decision on the unidentified object property based on the classifiers' suggestions. The supervisory decision is formulated by comparing the posterior probabilities of two hypotheses.

By Bayes' rule, the posterior probability of $X = X_0$ conditioned on $A = A_0$ is

$$P(X = X_0|A = A_0) = \frac{P(A = A_0|X = X_0)P(X = X_0)}{P(A = A_0)}. \tag{5.5}$$

The posterior probabilities of the four possible outcomes are summarized in Table 5.2.

Table 5.2: Summary of the posterior probabilities for a single classifier

X_0	A_0	$P(X = X_0\|A = A_0)$
T	T	$\frac{\sigma_T u}{\sigma_T u+(1-\sigma_F)(1-u)}$
F	T	$\frac{(1-\sigma_F)(1-u)}{\sigma_T u+(1-\sigma_F)(1-u)}$
T	F	$\frac{(1-\sigma_T)u}{(1-\sigma_T)u+\sigma_F(1-u)}$
F	F	$\frac{\sigma_F(1-u)}{(1-\sigma_T)u+\sigma_F(1-u)}$

For instance, if the classifier decides that $A = T$, then the supervisor compares the posterior probability of $P(X = T|A = T)$ and $P(X = F|A = T)$ from the table, then chooses the most likely hypothesis of X.

The supervisory decision rule by maximum likelihood classification is

$$O_s = \begin{cases} T & \text{if } \frac{P(X=T|A=A_0)}{P(X=F|A=A_0)} > 1, \\ F & \text{if } \frac{P(X=T|A=A_0)}{P(X=F|A=A_0)} \leq 1. \end{cases} \tag{5.6}$$

Let $f_{A_0} \in [0, \infty)$ denote the ratio of the posterior probabilities such that,

$$f_T = f_{A=T} = \frac{\sigma_T u}{(1-\sigma_F)(1-u)}, \tag{5.7a}$$

$$f_F = f_{A=F} = \frac{(1-\sigma_T)u}{\sigma_F(1-u)}. \tag{5.7b}$$

Let $\delta_{O_{s0}} : \mathcal{R} \to \{0, 1\}$ such that

$$\delta_T(f) = \delta_{O_s=T}(f) = \begin{cases} 1 & \text{if } f > 1, \\ 0 & \text{if } f \leq 1, \end{cases} \tag{5.8a}$$

$$\delta_F(f) = \delta_{O_s=F}(f) = \begin{cases} 1 & \text{if } f \leq 1, \\ 0 & \text{if } f > 1. \end{cases} \tag{5.8b}$$

Then, the conditional probabilities of the supervisor decision given an operator decision are,

$$P(O_s = T|A = T) = \delta_T(f_T), \tag{5.9a}$$

$$P(O_s = T|A = F) = \delta_T(f_F), \tag{5.9b}$$

$$P(O_s = F|A = T) = \delta_F(f_T), \tag{5.9c}$$

$$P(O_s = F|A = F) = \delta_F(f_F). \tag{5.9d}$$

Based on these probabilities, we assess the performance of the supervisory decision (a supervisor) O_s .

5.1.2.1 Performance of supervisory decisions

Assessing the probability of misclassification yields

$$\begin{aligned} P_{ms} &= P(O_s = T \wedge X = F) + P(O_s = F \wedge X = T) \\ &= P(O_s = T \wedge X = F | A = T)P(A = T) \\ &+ P(O_s = T \wedge X = F | A = F)P(A = F) \\ &+ P(O_s = F \wedge X = T | A = T)P(A = T) \\ &+ P(O_s = F \wedge X = T | A = F)P(A = F), \end{aligned} \tag{5.10}$$

by the theorem of total probability. Assuming that the supervisory decision is unbiased, we can relax the expression by conditional independence, i.e., $P(O_s = O_{s0} \wedge X = X_0 | A = A_0) = P(O_s = O_{s0} | A = A_0) \cdot P(X = X_0 | A = A_0)$. This means that given a classifier A's decision, the supervisory decision O_s and the object status X do not influence each other. Substituting Eq. (5.9) yields,

$$\begin{aligned} P_{ms} &= \delta_T(f_T)(1 - \sigma_F)(1 - u) + \delta_T(f_F)\sigma_F(1 - u) \\ &+ \delta_F(f_T)\sigma_T u + \delta_F(f_F)(1 - \sigma_T)u. \end{aligned} \tag{5.11}$$

Figure 5.2 shows the performance measures comparison for a classifier with $\sigma_T = \sigma_F = 0.7$. It is noted that there is a region in u where the supervisor performs better than the unsupervised classifier. Also the overall performance of the supervisor is no worse than that of the unsupervised classifier regardless of u.

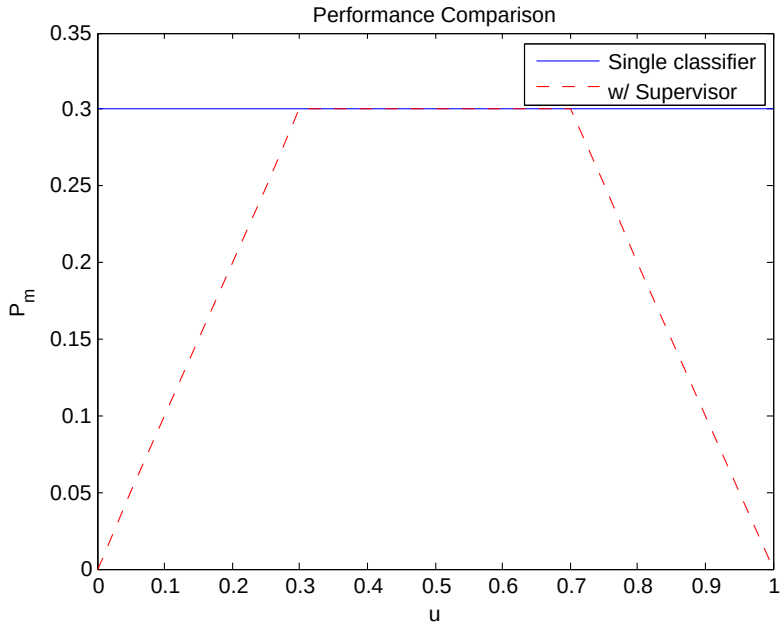


Figure 5.2: Performance measure for a single classifier with and without supervisor ($\sigma_T = \sigma_F = 0.7$)

5.2 Synergistic fusion rules

5.2.1 Performance of a two-classifier team

In a similar way to the single classifier case, here we define the performance of a two-classifier team.

Let $\mathcal{O}$ identify a member of a team of two classifiers, where $\mathcal{O} \in \{A, B\}$ with A and B each representing an individual classifier decision such that $A,\ B \in \{T, F\}$. The confusion matrix in conditional probability form for each individual classifier is

defined as,

$$P(\mathcal{O} = T|X = F) = 1 - \sigma_{F_{\mathcal{O}}}, \tag{5.12a}$$

$$P(\mathcal{O} = F|X = F) = \sigma_{F_{\mathcal{O}}}, \tag{5.12b}$$

$$P(\mathcal{O} = F|X = T) = 1 - \sigma_{T_{\mathcal{O}}}, \tag{5.12c}$$

$$P(\mathcal{O} = T|X = T) = \sigma_{T_{\mathcal{O}}}, \tag{5.12d}$$

$$\mathcal{O} = \{T, F\}. \tag{5.12e}$$

The false positive and false negative rates for each individual classifier are defined as in Eq. (5.1).

5.2.2 Fusion rules

Unlike the case of a single classifier, there can be many ways of assessing the probability of misclassification for a team. By assessing, we mean fusing the classification outcomes of the individual classifiers according to some logical rules.

Table 5.3: A list of fusion rules (F.R.) for a team of two classifiers (A, B). T and F denote the truth values.

A	B	Fusion rule No.															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	T	T	F	T	T	T	F	T	T	F	F	T	F	T	F	F	F
T	F	T	T	F	T	T	F	F	T	T	T	F	F	F	T	F	F
F	T	T	T	T	F	T	T	F	F	F	T	T	F	F	F	T	F
F	F	T	T	T	T	F	T	T	F	T	F	F	T	F	F	F	F
Note		(T)		$(\Rightarrow)$	$(\Leftarrow)$	$(\vee)$		$(\Leftrightarrow)$						$(\wedge)$			(F)

Table 5.8 shows the truth table of 16 possible logical fusion rules for a two-classifier team. We consider the four basic logical operators (conjunction, disjunction, implication, and biconditional) out of the 16 possible fusion rules as the candidate fusion rules as an initial investigation of the approach.

The following shows the formulation of the performance measure for each fusion rule. We assume that the decisions of A and B are conditionally independent given X.

5.2.2.1 Conjunction $(A \wedge B)$

Table 5.4 shows the truth table for the conjunction rule. Given the truth table,

Table 5.4: Truth table for conjunction rule

A	B	$A \wedge B$
T	T	T
T	F	F
F	T	F
F	F	F

the probability of misclassification is given as,

$$\begin{aligned} P_m = & P(A = B = T \wedge X = F) \\ & + P(A = B = F \wedge X = T) \\ & + P(A = T \wedge B = F \wedge X = T) \\ & + P(A = F \wedge B = T \wedge X = T). \end{aligned} \tag{5.13}$$

Using the conditional independence assumption, i.e., $P(A = T \wedge B = T|X = F) = P(A = T|X = F) \cdot P(B = T|X = F)$ yields,

$$\begin{aligned} P_m = & P(A = T|X = F)P(B = T|X = F)P(X = F) \\ & + P(A = F|X = T)P(B = F|X = T)P(X = T) \\ & + P(A = T|X = T)P(B = F|X = T)P(X = T) \\ & + P(A = F|X = T)P(B = T|X = T)P(X = T). \end{aligned} \tag{5.14}$$

Substituting Eq. (5.12) yields,

$$P_m = (1-\sigma_{F_A})(1-\sigma_{F_B})(1-u) + ((1-\sigma_{T_A}) + (1-\sigma_{T_B}) - (1-\sigma_{T_A})(1-\sigma_{T_B}))u$$
$$= (1-\sigma_{F_A})(1-\sigma_{F_B}) + u((1-\sigma_{T_A}) + (1-\sigma_{T_B}) - (1-\sigma_{T_A})(1-\sigma_{T_B}) - (1-\sigma_{F_A})(1-\sigma_{F_B})). \tag{5.15}$$

5.2.2.2 Disjunction $(A \vee B)$

Table 5.5 shows the truth table for the disjunction rule. Given the truth table,

Table 5.5: Truth table for disjunction rule

A	B	$A \wedge B$
T	T	T
T	F	T
F	T	T
F	F	F

the probability of misclassification is given as,

$$\begin{aligned}
P_m &= P(A = B = T \wedge X = F) + P(A = B = F \wedge X = T) \\
&\quad + P(A = T \wedge B = F \wedge X = F) \\
&\quad + P(A = F \wedge B = T \wedge X = F) \\
&= P(A = T|X = F)P(B = T|X = F)P(X = F) \\
&\quad + P(A = F|X = T)P(B = F|X = T)P(X = T) \\
&\quad + P(A = T|X = F)P(B = F|X = F)P(X = F) \\
&\quad + P(A = F|X = F)P(B = T|X = F)P(X = F) \\
&= ((1-\sigma_{F_A}) + (1-\sigma_{F_B}) - (1-\sigma_{F_A})(1-\sigma_{F_B}))(1-u) + (1-\sigma_{T_A})(1-\sigma_{T_B})u \\
&= (1-\sigma_{F_A}) + (1-\sigma_{F_B}) - (1-\sigma_{F_A})(1-\sigma_{F_B}) \\
&\quad + u((1-\sigma_{T_A})(1-\sigma_{T_B}) - (1-\sigma_{F_A}) - (1-\sigma_{F_B}) + (1-\sigma_{F_A})(1-\sigma_{F_B})).
\end{aligned} \tag{5.16}$$

5.2.2.3 Implication

Table 5.6 shows the truth table for the implication rule ($A \Rightarrow B$).

Table 5.6: Truth table for implicational rule

A	B	$A \wedge B$
T	T	T
T	F	F
F	T	T
F	F	T

Given the truth table, the probability of misclassification is given as,

$$\begin{aligned}
P_m &= P(A = B = T \wedge X = F) + P(A = B = F \wedge X = F) \\
&\quad + P(A = T \wedge B = F \wedge X = T) \\
&\quad + P(A = F \wedge B = T \wedge X = F) \\
&\quad = (1 - (1 - \sigma_{F_A}) + (1 - \sigma_{F_A})(1 - \sigma_{F_B}))(1 - u) + ((1 - \sigma_{T_B}) - (1 - \sigma_{T_A})(1 - \sigma_{T_B}))u \\
&\quad = 1 - (1 - \sigma_{F_A}) + (1 - \sigma_{F_A})(1 - \sigma_{F_B}) \\
&\quad + u((1 - \sigma_{T_B}) - (1 - \sigma_{T_A})(1 - \sigma_{T_B}) - 1 + (1 - \sigma_{F_A}) - (1 - \sigma_{F_A})(1 - \sigma_{F_B})).
\end{aligned} \tag{5.17}$$

The probability of misclassification for $B \Rightarrow A$ is obtained by switching A and B in Eq. (5.17), given as

$$\begin{aligned}
P_m &= 1 - (1 - \sigma_{F_B}) + (1 - \sigma_{F_A})(1 - \sigma_{F_B}) \\
&\quad + u((1 - \sigma_{T_A}) - (1 - \sigma_{T_A})(1 - \sigma_{T_B}) - 1 + (1 - \sigma_{F_B}) - (1 - \sigma_{F_A})(1 - \sigma_{F_B})).
\end{aligned} \tag{5.18}$$

5.2.2.4 Biconditional ($A \Leftrightarrow B$)

Table 5.7 shows the truth table for the biconditional rule. Given the truth table,

Table 5.7: Truth table for biconditional rule

A	B	$A \wedge B$
T	T	T
T	F	F
F	T	F
F	F	T

the probability of misclassification is given as,

$$\begin{aligned} P_m &= P(A = B = T \wedge X = F) + P(A = B = F \wedge X = F) \\ &+ P(A = T \wedge B = F \wedge X = T) \\ &+ P(A = F \wedge B = T \wedge X = T) \\ &= (1 - (1 - \sigma_{F_A}) - (1 - \sigma_{F_B}) + 2(1 - \sigma_{F_A})(1 - \sigma_{F_B}))(1 - u) \\ &+ ((1 - \sigma_{T_A}) + (1 - \sigma_{T_B}) - 2(1 - \sigma_{T_A})(1 - \sigma_{T_B}))u \\ &= 1 - (1 - \sigma_{F_A}) - (1 - \sigma_{F_B}) + 2(1 - \sigma_{F_A})(1 - \sigma_{F_B}) \\ &+ u((1 - \sigma_{T_A}) + (1 - \sigma_{T_B}) - 2(1 - \sigma_{T_A})(1 - \sigma_{T_B}) - 1 \\ &+ (1 - \sigma_{F_A}) + (1 - \sigma_{F_B}) - 2(1 - \sigma_{F_A})(1 - \sigma_{F_B})). \end{aligned} \tag{5.19}$$

5.2.3 Aggregated team performance

The team performance under such fusion rules can be expressed in the following aggregated form (let subscript $\mathcal{T}$ denote "Team"):

$$P_m = (1 - \sigma_F)_{\mathcal{T}}(1 - u) + (1 - \sigma_T)_{\mathcal{T}} u. \tag{5.20}$$

Table 5.8 - 5.9 summarize the aggregated false positive and false negative rates.

Definition V.1. Homogeneous Team

A homogeneous team is a team such that all of the team members have the same true

Table 5.8: Aggregated false positive rates

Fusion Rule	$(1-\sigma_F)_{\mathcal{T}}$
$A \wedge B$	$(1-\sigma_{F_A})(1-\sigma_{F_B})$
$A \vee B$	$(1-\sigma_{F_A})+(1-\sigma_{F_B})-(1-\sigma_{F_A})(1-\sigma_{F_B})$
$A \Rightarrow B$	$1-(1-\sigma_{F_A})+(1-\sigma_{F_A})(1-\sigma_{F_B})$
$B \Rightarrow A$	$1-(1-\sigma_{F_B})+(1-\sigma_{F_A})(1-\sigma_{F_B})$
$A \Leftrightarrow B$	$1-(1-\sigma_{F_A})-(1-\sigma_{F_B})+2(1-\sigma_{F_A})(1-\sigma_{F_B})$

Table 5.9: Aggregated false negative rates

Fusion Rule	$(1-\sigma_T)_{\mathcal{T}}$
$A \wedge B$	$(1-\sigma_{T_A})+(1-\sigma_{T_B})-(1-\sigma_{T_A})(1-\sigma_{T_B})$
$A \vee B$	$(1-\sigma_{T_A})(1-\sigma_{T_B})$
$A \Rightarrow B$	$(1-\sigma_{T_B})-(1-\sigma_{T_A})(1-\sigma_{T_B})$
$B \Rightarrow A$	$(1-\sigma_{T_A})-(1-\sigma_{T_A})(1-\sigma_{T_B})$
$A \Leftrightarrow B$	$(1-\sigma_{T_A})+(1-\sigma_{T_B})-2(1-\sigma_{T_A})(1-\sigma_{T_B})$

positive and true negative rates, i.e.,

$$\sigma_{T_O} = \bar{\sigma}_T,\ \sigma_{F_O} = \bar{\sigma}_F\ , O \in \{A, B\},$$

where $\bar{\sigma}_i \in [0.5,\ 1]$, $i \in \{T, F\}$. If a team is not homogeneous, then it is heterogeneous.

Figure 5.3 illustrates the probability of misclassification P_m of different fusion rules with respect to the prior information u for a homogeneous team. Figure 5.4 shows the probability of misclassification of different fusion rules with respect to the prior information for a heterogeneous team.

Compared to a single classifier case, two-homogeneous-classifier teams with conjunction, disjunction, and implication fusion rules are synergistic depending on u. On the other hand, the biconditional fusion rule is non-synergistic for all u. Similarly for a heterogenous team, conjunction, disjunction, and implication fusion rules are synergistic for some u, but the biconditional rule is non-synergistic for all u.

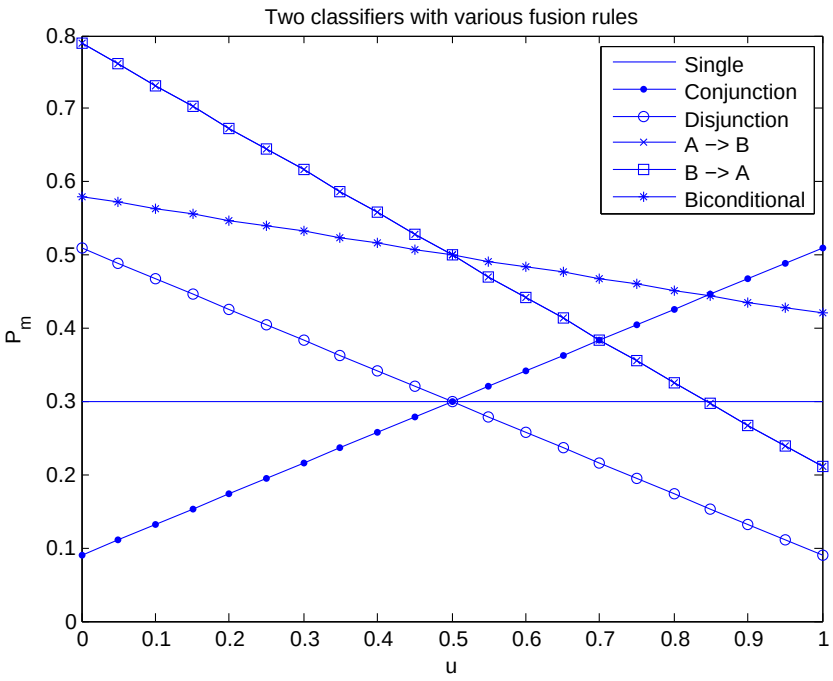
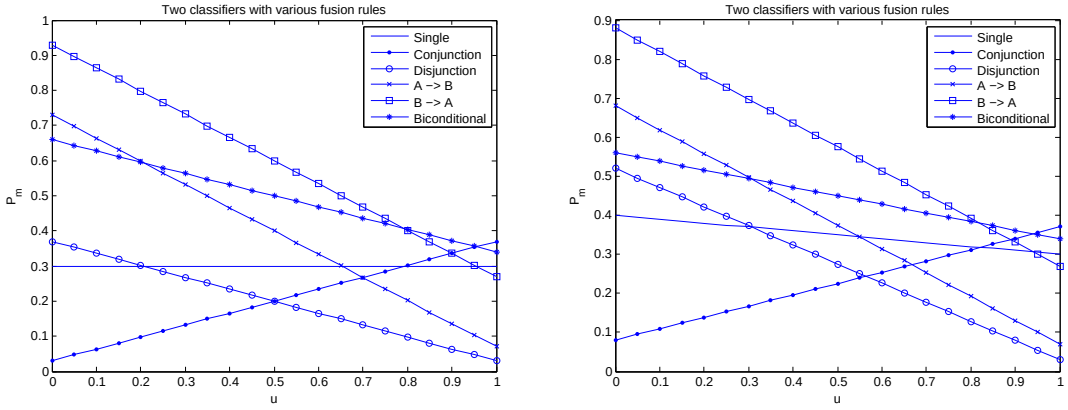


Figure 5.3: A homogeneous team performance $\sigma_F = 0.7$, $\sigma_T = 0.7$



(a) $\sigma_{F_A} = \sigma_{T_A} = 0.7$, $\sigma_{F_B} = \sigma_{T_B} = 0.9$

(b) More frequent false alarms $\sigma_{F_A} = 0.6$, $\sigma_{T_A} = 0.7$, $\sigma_{F_B} = 0.8$, $\sigma_{T_B} = 0.9$

Figure 5.4: Heterogeneous team performance

5.2.4 Supervisory decision for classifier team

We formulate the supervisor decisions based on the individual classifier decisions.

By Bayes rule,

$$P(X = X_0|A = A_0 \wedge B = B_0) = \frac{P(A = A_0 \wedge B = B_0|X = X_0)P(X = X_0)}{P(A = A_0 \wedge B = B_0)}.$$

The posterior probabilities of the eight possible outcomes are summarized in Table 5.10.

Table 5.10: Summary of the posterior probabilities for a classifier team

X_0	A_0	B_0	$P(X = X_0\|A = A_0 \wedge B = B_0)$
T	T	T	$\frac{(1-(1-\sigma_{T_A}))(1-(1-\sigma_{T_B}))u}{(1-(1-\sigma_{T_A}))(1-(1-\sigma_{T_B}))u+(1-\sigma_{F_A})(1-\sigma_{F_B})(1-u)}$
F	T	T	$\frac{(1-\sigma_{F_A})(1-\sigma_{F_B})(1-u)}{(1-(1-\sigma_{T_A}))(1-(1-\sigma_{T_B}))u+(1-\sigma_{F_A})(1-\sigma_{F_B})(1-u)}$
T	T	F	$\frac{(1-(1-\sigma_{T_A}))(1-\sigma_{T_B})u}{(1-(1-\sigma_{T_A}))(1-\sigma_{T_B})u+(1-\sigma_{F_A})(1-(1-\sigma_{F_B}))(1-u)}$
F	T	F	$\frac{(1-\sigma_{F_A})(1-(1-\sigma_{F_B}))(1-u)}{(1-(1-\sigma_{T_A}))(1-\sigma_{T_B})u+(1-\sigma_{F_A})(1-(1-\sigma_{F_B}))(1-u)}$
T	F	T	$\frac{(1-\sigma_{T_A})(1-(1-\sigma_{T_B}))u}{(1-\sigma_{T_A})(1-(1-\sigma_{T_B}))u+(1-(1-\sigma_{F_A}))(1-\sigma_{F_B})(1-u)}$
F	F	T	$\frac{(1-(1-\sigma_{F_A}))(1-\sigma_{F_B})(1-u)}{(1-\sigma_{T_A})(1-(1-\sigma_{T_B}))u+(1-(1-\sigma_{F_A}))(1-\sigma_{F_B})(1-u)}$
T	F	F	$\frac{(1-\sigma_{T_A})(1-\sigma_{T_B})u}{(1-\sigma_{T_A})(1-\sigma_{T_B})u+(1-(1-\sigma_{F_A}))(1-(1-\sigma_{F_B}))(1-u)}$
F	F	F	$\frac{(1-(1-\sigma_{F_A}))(1-(1-\sigma_{F_B}))(1-u)}{(1-\sigma_{T_A})(1-\sigma_{T_B})u+(1-(1-\sigma_{F_A}))(1-(1-\sigma_{F_B}))(1-u)}$

The supervisory decision rule by maximum likelihood classification is

$$O_s = \begin{cases} T & \text{if } \frac{P(X=T|A=A_0 \wedge B=B_0)}{P(X=F|A=A_0 \wedge B=B_0)} > 1, \\ F & \text{if } \frac{P(X=T|A=A_0 \wedge B=B_0)}{P(X=F|A=A_0 \wedge B=B_0)} \leq 1. \end{cases} \tag{5.21}$$

Let $f_{A_0,B_0} \in [0\ \infty)$ denote the ratio of the posterior probabilities such that,

$$f_{T,T} = f_{A=T,B=T} = \frac{(1-(1-\sigma_{T_A}))(1-(1-\sigma_{T_B}))u}{(1-\sigma_{F_A})(1-\sigma_{F_B})(1-u)}, \tag{5.22a}$$

$$f_{T,F} = f_{A=T,B=F} = \frac{(1-(1-\sigma_{T_A}))(1-\sigma_{T_B})u}{(1-\sigma_{F_A})(1-(1-\sigma_{F_B}))(1-u)}, \tag{5.22b}$$

$$f_{F,T} = f_{A=F,B=T} = \frac{(1-(1-\sigma_{T_B}))(1-\sigma_{T_A})u}{(1-\sigma_{F_B})(1-(1-\sigma_{F_A}))(1-u)}, \tag{5.22c}$$

$$f_{F,F} = f_{A=F,B=F} = \frac{(1-\sigma_{T_A})(1-\sigma_{T_B})u}{(1-(1-\sigma_{F_A}))(1-(1-\sigma_{F_B}))(1-u)}. \tag{5.22d}$$

Using Eq. (D.8), we can define the conditional probabilities of a supervisor decision given the team decisions. Table 5.11 summarizes the probabilities.

Table 5.11: Summary of the conditional probabilities for a supervisor decision given the team decisions

O_s	A_0	B_0	$P(O_s = O_{s0} \mid A = A_0 \wedge B = B_0)$
T	T	T	$\delta_T(f_{T,T})$
F	T	T	$\delta_F(f_{T,T})$
T	T	F	$\delta_T(f_{T,F})$
F	T	F	$\delta_F(f_{T,F})$
T	F	T	$\delta_T(f_{F,T})$
F	F	T	$\delta_F(f_{F,T})$
T	F	F	$\delta_T(f_{F,F})$
F	F	F	$\delta_F(f_{F,F})$

5.2.4.1 Performance of supervisory decisions

We assess the performance measure of a supervised team with respect to the fusion rules. Note that there are sequences of decisions involved until the final supervisory decision is made: Classifiers A and B each make an independent classification decision, then the two decisions are fused by some fusion rule, and the final supervisory decision is made based on the fused decision. In other words, this architecture uses fusion by logical operator and Bayes inference to reach the final supervisory decision.

Let $O_f \in \{T, F\}$ denote the fusion rule decision. By the product rule, the proba-

bility of misclassification for a two-classifier team with a supervisor is

$$P_{ms} = P(O_s = T \wedge X = F) + P(O_s = F \wedge X = T) \tag{5.23}$$

$$\begin{aligned} &= P(O_s = T \wedge X = F | O_f = T) P(O_f = T) \\ &+ P(O_s = T \wedge X = F | O_f = F) P(O_f = F) \\ &+ P(O_s = F \wedge X = T | O_f = T) P(O_f = T) \\ &+ P(O_s = F \wedge X = T | O_f = F) P(O_f = F). \end{aligned} \tag{5.24}$$

Assuming that O_s and X are conditionally independent given O_f, we get

$$\begin{aligned} P_{ms} &= P(O_s = T | O_f = T) P(X = F \wedge O_f = T) \\ &+ P(O_s = T | O_f = F) P(X = F \wedge O_f = F) \\ &+ P(O_s = F | O_f = T) P(X = T \wedge O_f = T) \\ &+ P(O_s = F | O_f = F) P(X = T \wedge O_f = F). \end{aligned} \tag{5.25}$$

Here, we provide an outline for assessing Eq. (5.25) for the biconditional fusion rule as an example. For biconditional rule, the outcome of the fusion rule O_f is equivalent to the followings:

$$O_f = T \Leftrightarrow (A = T \wedge B = T) \vee (A = F \wedge B = F), \tag{5.26}$$

$$O_f = F \Leftrightarrow (A = T \wedge B = F) \vee (A = F \wedge B = T). \tag{5.27}$$

For mutually exclusive events a and b, it can be shown that

$$P(O_s | a \vee b) = \frac{P(O_s|a)P(a) + P(O_s|b)P(b)}{P(a) + P(b)}, \tag{5.28}$$

and

$$P(X \wedge (a \vee b)) = P(X \wedge a) + P(X \wedge b). \tag{5.29}$$

Using Eq. (5.26)-(5.29), together with Tables 5.10 and 5.11, we can evaluate the probability of misclassification in Eq. (5.25).

Figure 5.5 compares the performance of fusion rules for a homogeneous team with $\sigma_{F_{\mathcal{O}}} = \sigma_{T_{\mathcal{O}}} = 0.5$, $\mathcal{O} \in \{A, B\}$.

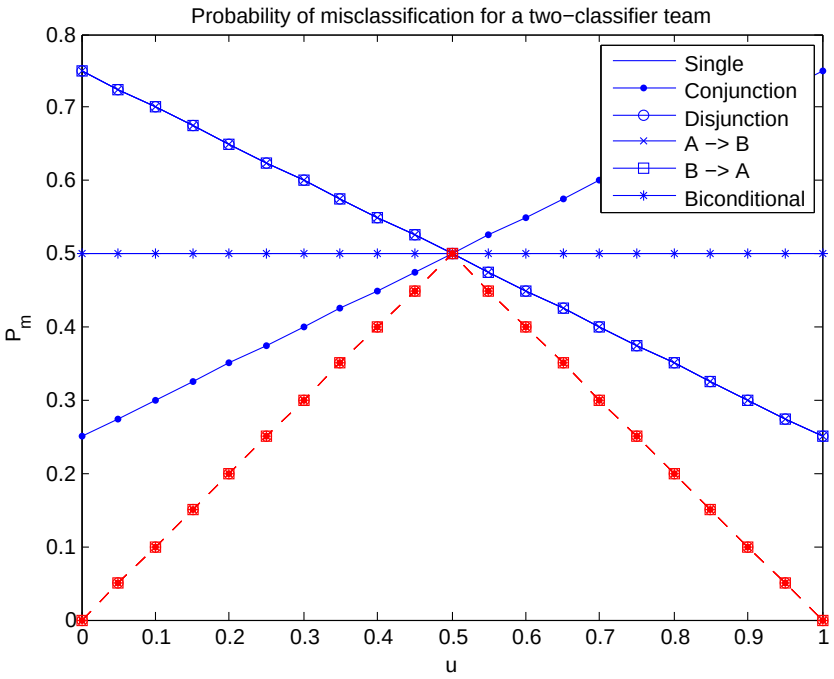


Figure 5.5: A comparison of a homogeneous team performance with and without supervisor. Blue solid line and red dashed line indicate unsupervised and supervised, respectively. ($\sigma_F = 0.5$, $\sigma_T = 0.5$)

The probability of misclassification for the supervised team decision is always less than or equal to the probability of misclassification for the unsupervised team decision regardless of the fusion rules and the prior information u. Also, the probability of misclassification for the supervised team decision is always less than or equal to the probability of misclassification for the single classifier decision. Therefore, the fusion

rules with supervisory control are always synergistic compared to the case for a single classifier without the supervisory control.

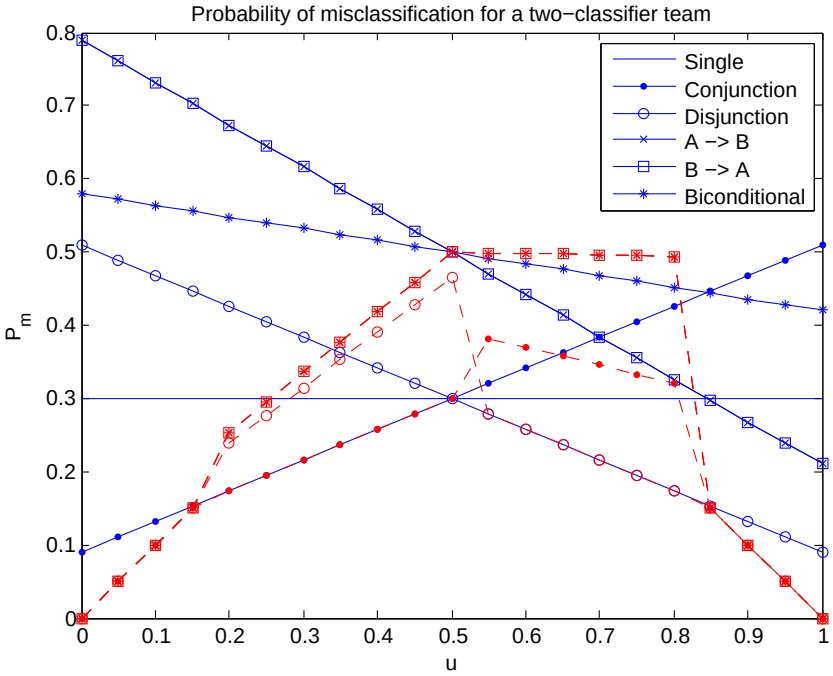


Figure 5.6: A comparison of a homogeneous team performance with and without supervisor ($\sigma_F = 0.7$, $\sigma_T = 0.7$)

However, there are supervised fusion rules that perform worse than unsupervised fusion rules in some region of u. Figure 5.6 shows supervised team performance for $\sigma_{F_\mathcal{O}} = \sigma_{T_\mathcal{O}} = 0.7$, $\mathcal{O} \in \{A, B\}$. For instance, the supervised team with conjunction rule does worse than the unsupervised team with disjunction rule for $u \in [0.55, 0.8]$. The implication of these results is that there exists a performance-optimal fusion rule for a classifier team that varies with the prior information u.

We further investigate the rest of the fusion rules. Figure 5.7 illustrates the optimal fusion rule among the 16 rules and the corresponding minimal probability of misclassification P_m for $\sigma_{F_\mathcal{O}} = \sigma_{T_\mathcal{O}} = 0.5$, $\mathcal{O} \in \{A, B\}$. Note that the ordinate of the plot above denotes the fusion rule number as labeled in Table 5.8. The result indicates that when both classifiers are as good as pure guesses, the optimal fusion rule is to

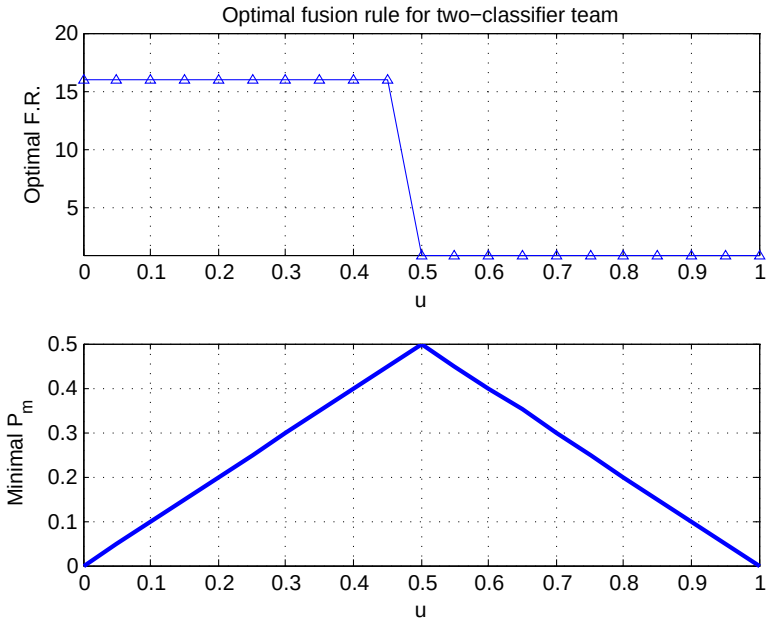


Figure 5.7: The optimal fusion rule for a two-classifier team and the corresponding minimal probability of misclassification ($\sigma_F = 0.5$, $\sigma_T = 0.5$)

completely ignore their opinions and classify either as F (F.R. No. 16) when the prior information u is smaller than 0.5 or as T (F.R. No. 1) when the prior information u is greater than or equal to 0.5. The minimal probability of misclassification P_m reaches its maximum when the prior information is uninformative ($u = 0.5$).

Figure 5.8 illustrates the optimal fusion rule among the 16 rules and the corresponding minimal probability of misclassification P_m for $\sigma_{F_\mathcal{O}} = \sigma_{T_\mathcal{O}} = 0.7$, $\mathcal{O} \in \{A, B\}$. As the classifiers become more reliable than pure guesses, the optimal fusion rule exploits the classifiers appropriately according to the level of prior information at present. For instance, when the prior information is either small ($u < 0.2$) or large ($u > 0.8$), the optimal fusion rule is to ignore the classifiers' opinions and classify either as F or T. However, when the prior information is in mid-range ($0.2 \leq u \leq 0.8$), the optimal rule is either the conjunction (No. 13) or the disjunction rule (No. 5).

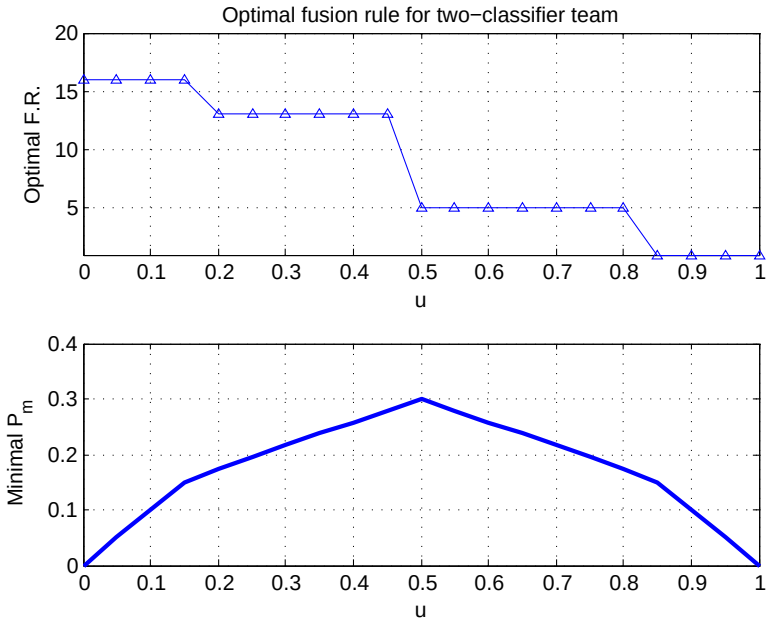


Figure 5.8: The optimal fusion rule for a two-classifier team and the corresponding minimal probability of misclassification ($\sigma_F = 0.7$, $\sigma_T = 0.7$)

5.3 Conclusion & future work

In this chapter, we studied the performance of a classifier team under several fusion rules. It was shown numerically that the supervised decisions for a single classifier are no worse than the unsupervised decisions regardless of the prior information. Moreover, we showed that there are synergistic fusion rules for unsupervised and supervised team decisions compared to a single classifier. The study showed that depending on the level of prior information, there is a performance-optimal fusion rule for the team.

Our framework provides a mechanism for using numbers of classifiers in a team: Divide the classifiers into two subgroups, and divide the classifiers within the subgroups into two, and so on, until the number of the smallest group members becomes two. Then, we can apply our framework for two-classifier teams sequentially starting from the smallest subgroups.

One of the implications of this work is the use of multiple fusion rules in uncertain situations when prior information is not completely known. For instance, we are provided by the intelligence with a set of possible prior information values, as opposed to a single number, with probabilities associated to each plausible prior information. One way to overcome the situation and make the best classification decision based on our results is as follows: Since there is an optimal fusion rule for a specific prior information, a fusion of a set of optimal fusion rules corresponding to the set of possible prior information, where each of the optimal fusion rule is weighted by the associated probability, can be considered. As future work, we propose to investigate this "super-fusion" approach and examine whether the approach is reliable when the prior information is uncertain.

CHAPTER VI

A Team of Heterogeneous Classifiers

> However beautiful the strategy, you should occasionally look at the results.
>
> Winston Churchill

We consider a team composed of a workload-independent, trichotomous classifier and a workload-dependent, dichotomous classifier (*mixed-initiative* team). The team is structured in a *nested* architecture, that is: first the primary, workload-independent, trichotomous, classifier examines the classification task, and if the task is classified as unknown, the secondary, workload-dependent, dichotomous, classifier is called upon.

We demonstrate that having two classifiers, a trichotomous classifier (true, false, or *unknown*) with workload-independent performance that turns over the data classified as unknown to a dichotomous classifier (true or false) with workload-dependent performance, gives superior classification performance (lower probability of misclassification) compared to a single dichotomous classifier.

6.1 Mixed-initiative nested thresholding

We consider a mixed-initiative nested classification where two heterogeneous classifiers are composed in a nested architecture. Figure 6.1 shows the concept. We

assume the following:

i. The workload-independent classifier and the workload-dependent classifier examine the task independently.

ii. The workload of the secondary classifier is determined by the probability of being called upon by the primary classifier.

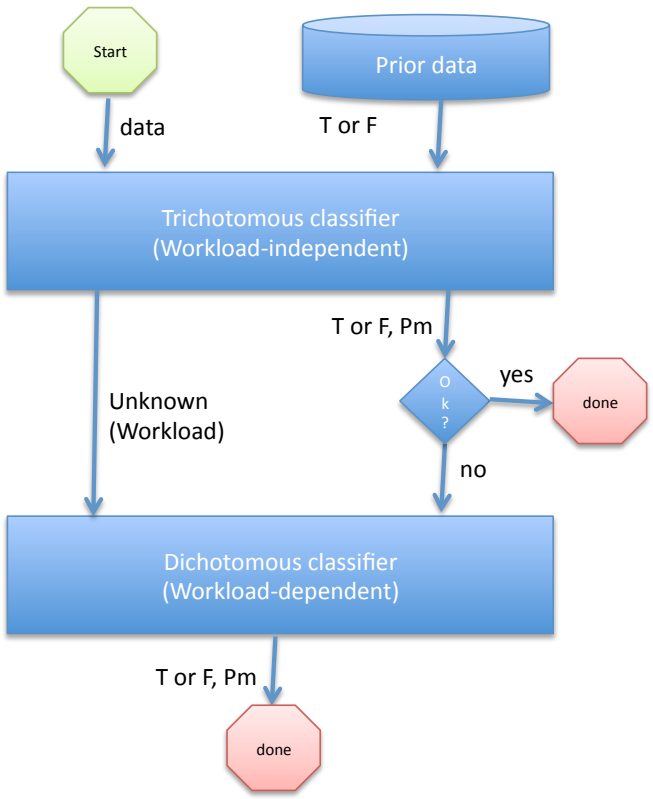


Figure 6.1: Concept of mixed-initiative nested classification

6.1.1 Problem formulation

6.1.1.1 Workload-independent trichotomous classifier

Let τ_1 and τ_2 be the threshold variables. Then, the cumulative probability distributions for Gaussian distributions are

$$\sigma_{T_1} = \int_{-\infty}^{\tau_1} a_T e^{-(w+b_T)^2/c_T^2} dw \tag{6.1a}$$

$$\sigma_{F_1} = \int_{\tau_2}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw \tag{6.1b}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, and $c_i = \sqrt{2s_i^2}$ with $i \in \{T, F\}$.

The region of indecision, i.e., $[\tau_1, \tau_2]$, of the primary trichotomous classifier determines the workload applied to the secondary classifier. We define a workload variable, $W \in [0, 1]$, with 0 indicating idle and 1 indicating fully loaded. Let $f_i(w) = a_i e^{-(w+b_i)^2/c_i^2}$ with $i \in \{T, F\}$, then the workload variable is defined as

$$W = \int_{\tau_1}^{\tau_2} u f_T(w) + (1-u) f_F(w) dw. \tag{6.2}$$

Note that the range of W is [0, 1] for any τ_1 and τ_2. We assume that the workload variable is normalized such that the maximum value is unity when the workload-dependent classifier is fully loaded.

6.1.1.2 Workload-dependent dichotomous classifier

The classification performance of a human operator is modeled as follows. Recognizing the concavity of the curve, we model the Yerkes-Dodson law (Fig. A.1 in the

appendix) as a quadratic function of the workload as,

$$\sigma_{T_2} = -(4\sigma_T^* - 2)W^2 + (4\sigma_T^* - 2)W + 0.5, \tag{6.3}$$

$$\sigma_{F_2} = -(4\sigma_F^* - 2)W^2 + (4\sigma_F^* - 2)W + 0.5, \tag{6.4}$$

where $\sigma_{(\cdot)}^* \in [0.5, 1]$ determines the maximum of $\sigma_{(\cdot)}$.

6.1.1.3 Probability of misclassification for two classifiers

The probability of misclassification is a sum of contributions of two faulty outcomes: false positives and false negatives. By the theorem of total probability, the probability of misclassification includes all possible cases of misclassification by the two classifiers. To maintain the flow, the derivation can be found in Appendix D. The probability of misclassification for two classifiers is

$$P_m^2 = \bar{\sigma}_1^T R_2 \bar{\sigma}_2, \tag{6.5}$$

where

$$\bar{\sigma}_i = \begin{bmatrix} \sigma_{F_i} & 1 - \sigma_{F_i} & 1 - \sigma_{T_i} & \sigma_{T_i} \end{bmatrix}^T, \; i = 1, 2$$

$$R_2 = \begin{bmatrix} \delta_T(f_{1,1})(1-u) & \delta_T(f_{1,2})(1-u) & 0 & 0 \\ \delta_T(f_{2,1})(1-u) & \delta_T(f_{2,2})(1-u) & 0 & 0 \\ 0 & 0 & \delta_F(f_{1,1})u & \delta_F(f_{1,2})u \\ 0 & 0 & \delta_F(f_{2,1})u & \delta_F(f_{2,2})u \end{bmatrix},$$

with

$$
\begin{aligned}
f_{1,1} &= \left(\frac{1-\sigma_{T_1}}{\sigma_{F_1}}\right)\left(\frac{1-\sigma_{T_2}}{\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right), \\
f_{1,2} &= \left(\frac{1-\sigma_{T_1}}{\sigma_{F_1}}\right)\left(\frac{\sigma_{T_2}}{1-\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right), \\
f_{2,1} &= \left(\frac{\sigma_{T_1}}{1-\sigma_{F_1}}\right)\left(\frac{1-\sigma_{T_2}}{\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right), \\
f_{2,2} &= \left(\frac{\sigma_{T_1}}{1-\sigma_{F_1}}\right)\left(\frac{\sigma_{T_2}}{1-\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right).
\end{aligned}
$$

The global objective of the nested team architecture is to minimize the probability of misclassification by choosing the threshold variables for the primary trichotomous classifier, i.e.,

$$\min_{\tau_1,\tau_2} P_m^2,$$

subject to inequality constraints,

$$\tau_1 \leq \tau_2 \tag{6.6a}$$

$$\sigma_{T_1} \geq 0.5, \tag{6.6b}$$

$$\sigma_{F_1} \geq 0.5, \tag{6.6c}$$

$$\sigma_{T_1} \leq 1, \tag{6.6d}$$

$$\sigma_{F_1} \leq 1. \tag{6.6e}$$

This formalism allows the two classifiers to have the same goal, although the mechanism behind how each classifier functions is different. Also, due to the inequality constraints, the formulation does not allow the trichotomous classifier to experience perverse behavior, i.e., $\sigma_{(\cdot)} \in [0,\, 0.5]$.

6.1.2 Classifiability

The fundamental difficulty of a classification task is determined by the nature of the distributions that are to be classified. Given two undistinguishable distributions, e.g., two Gaussian distributions with identical mean and variance, it is impossible to make the classifier's performance better than a pure guess because the task itself is *unclassifiable*. Recognizing this, we use the term *classifiability* to quantify the fundamental difficulty of the classification task at hand.

Definition VI.1. Classifiability

Classifiability is quantified as the reciprocal of the minimal probability of misclassification performed by a dichotomous classifier on a logarithmic scale, i.e.,

$$\text{Classifiability} = \log \frac{1}{P_m^{1^*}}, \tag{6.7}$$

where $P_m^{1^*}$ denotes the minimal probability of misclassification of a dichotomous classifier. Note that the measure is the *information content* defined by Shannon [12].

Figure 6.2 illustrates an example of two Gaussian distributions, each representing the distribution of either the T or F sub-population. Figure 6.3 illustrates the classifiability as a function of the distance between the means of the distributions. Note that as the distance between the means increases, the classifiability of the task increases. On the other hand, if the distance between the means is zero ($|m_T - m_F| = 0$), the classifiability reaches its minimum, $\log\left(\frac{1}{0.5}\right)$.

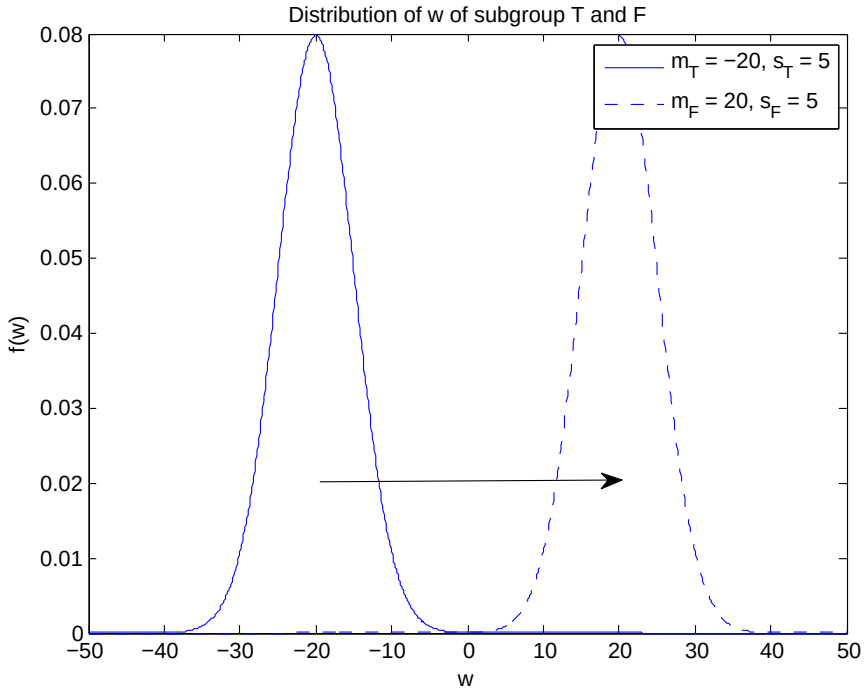


Figure 6.2: Two Gaussian distributions p_T (moving mean) and p_F with an equal variance

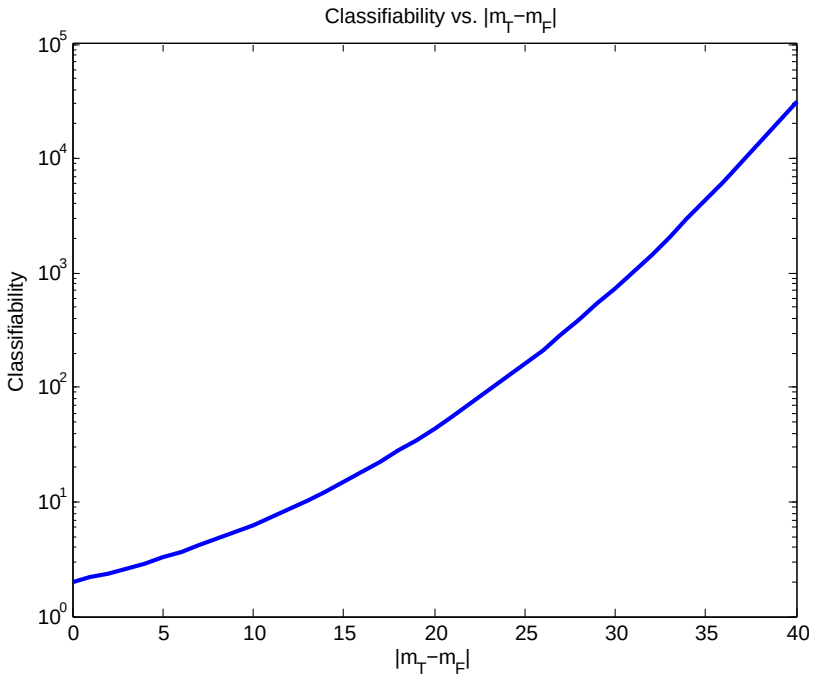


Figure 6.3: Classifiability of a dichotomous classifier for two Gaussian distributions with moving means

6.1.3 Optimal mixed-initiative nested thresholding

At minimum τ_1^* and τ_2^*, the problem must satisfy the K-K-T conditions, i.e.,

$$\frac{\partial}{\partial\tau_1}P_m^2 + \mu_1\frac{\partial}{\partial\tau_1}(\tau_1-\tau_2) + \mu_2\frac{\partial}{\partial\tau_1}(0.5-\sigma_{T_1}) + \mu_3\frac{\partial}{\partial\tau_1}(0.5-\sigma_{F_1}) + \mu_4\frac{\partial}{\partial\tau_1}(\sigma_{T_1}-1) + \mu_5\frac{\partial}{\partial\tau_1}(\sigma_{F_1}-1) = 0, \tag{6.8}$$

$$\frac{\partial}{\partial\tau_2}P_m^2 + \mu_1\frac{\partial}{\partial\tau_2}(\tau_1-\tau_2) + \mu_2\frac{\partial}{\partial\tau_2}(0.5-\sigma_{T_1}) + \mu_3\frac{\partial}{\partial\tau_2}(0.5-\sigma_{F_1}) + \mu_4\frac{\partial}{\partial\tau_2}(\sigma_{T_1}-1) + \mu_5\frac{\partial}{\partial\tau_2}(\sigma_{F_1}-1) = 0. \tag{6.9}$$

Investigating the analytical properties of the solution for mixed-initiative nested classifiers is left as future work.

We solve the optimization problem by using the MATLAB *fmincon* command. Figure 6.4 illustrates the performance comparison between the dichotomous classifier and the mixed-initiative nested classifiers with different initializations of the threshold variables shown on a logarithmic scale. Note that the search space has multiple local minima, so that depending on the initial conditions the performance of the mixed-initiative nested classifiers can be different. It is clear, however, that while the performance of the nested classifiers is sensitive to the initialization of the threshold variables, it is no worse than the dichotomous classifier regardless of the initialization as shown in Fig 6.4. Note that the performances of both dichotomous and mixed-initiative classifiers are linearly decreasing functions (on a logarithmic scale) with respect to the classifiability.

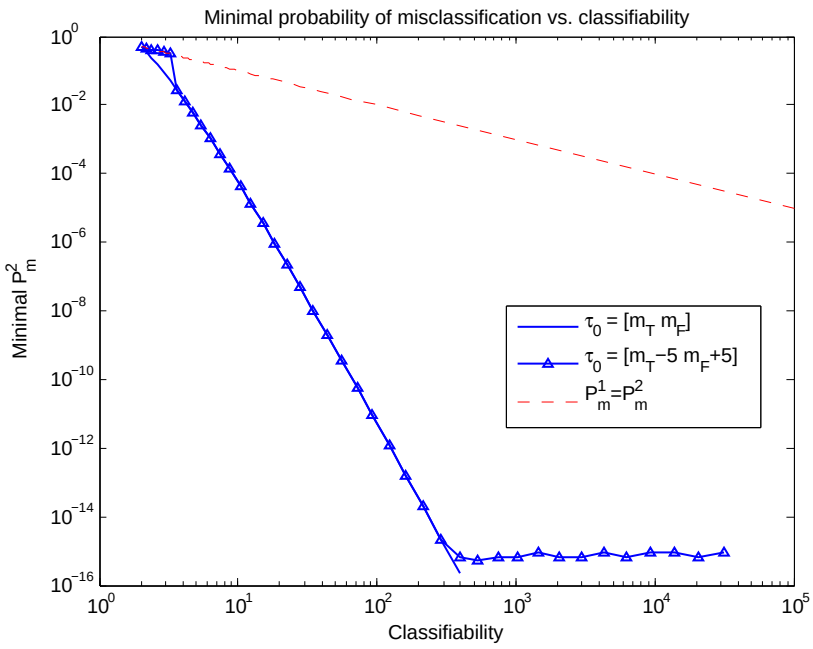


Figure 6.4: Comparison of dichotomous and mixed-initiative thresholding performance

6.2 Linear mixed-initiative nested thresholding

So far, we have studied a case when a scalar measurable quantity w is provided. In this section, we generalize the work by introducing multi-dimensional measurable quantities, such that the decision variable for thresholding is no longer a choice of a scalar value, but a choice of multi-dimensional variables.

6.2.1 Problem formulation

Consider a multivariate property $\mathbf{w} \in \mathcal{R}^n$ that can be measured from a population of objects of interest where the population comprises two disjoint sub-populations, T and F, and each sub-population is characterized by its own distribution of $\mathbf{w}$. Let $\mathbf{c}$ denote a sieving parameter that satisfies the constraint:

$$\mathbf{c}^T\mathbf{c} = 1. \tag{6.10}$$

Let $w = \mathbf{c}^T\mathbf{w}$ denote the sieved measurement and τ_1 and τ_2 be the threshold variables. If the distribution of the measurable property for each sub-population is Gaussian, the cumulative probability distributions are determined as,

$$\sigma_{T_1} = \int_{-\infty}^{\tau_1} a_T e^{-(w+b_T)^2/c_T^2} dw \tag{6.11a}$$

$$\sigma_{F_1} = \int_{\tau_2}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw \tag{6.11b}$$

where $a_i = 1/\sqrt{2\pi\left(\mathbf{c}^T P_{w_i}\mathbf{c}\right)}$, $b_i = -\mathbf{c}^T\bar{\mathbf{w}}_i$, $c_i = \sqrt{2\left(\mathbf{c}^T P_{w_i}\mathbf{c}\right)}$ with $i \in \{T, F\}$. For more background on this formulation, see Sec. 4.2

The region of indecision, i.e., $[\tau_1, \tau_2]$, determines the workload applied to the human operator. We define a workload variable, $W \in [0, 1]$, with 0 indicating idle and 1 indicating fully loaded. Let $f_i(w) = a_i e^{-(w+b_i)^2/c_i^2}$ with $i \in \{T, F\}$, then the workload variable is defined as

$$W = \int_{\tau_1}^{\tau_2} u f_T(w) + (1-u) f_F(w) dw. \tag{6.12}$$

The classification performance of a human operator is modeled as follows. Recognizing the convexity of the curve, we model the Yerkes-Dodson law as a quadratic function of the workload as,

$$\sigma_{T_2} = -(4\sigma_T^* - 2)W^2 + (4\sigma_T^* - 2)W + 0.5, \tag{6.13}$$

$$\sigma_{F_2} = -(4\sigma_F^* - 2)W^2 + (4\sigma_F^* - 2)W + 0.5 \tag{6.14}$$

where $\sigma_{(\cdot)}^* \in [0.5, 1]$ determines the maximum of $\sigma_{(\cdot)}$.

The global objective of the nested team architecture is to minimize the probability

of misclassification by choosing the threshold variables for the machine classifier, i.e.,

$$\min_{\tau_1,\tau_2,\mathbf{c}} P_m^2,$$

subject to constraints,

$$\mathbf{c}^T\mathbf{c} = 1, \tag{6.15a}$$

$$\tau_1 \leq \tau_2, \tag{6.15b}$$

$$\sigma_{T_1} \geq 0.5, \tag{6.15c}$$

$$\sigma_{F_1} \geq 0.5, \tag{6.15d}$$

$$\sigma_{T_1} \leq 1, \tag{6.15e}$$

$$\sigma_{F_1} \leq 1. \tag{6.15f}$$

Note that P_m^2 is defined in Eq. (D.10).

6.2.2 Optimal linear mixed-initiative nested thresholding

At minimum τ_1^*, τ_2^* and $\mathbf{c}^*$, the problem must satisfy the K-K-T conditions, i.e.,

$$\begin{aligned}
&\frac{\partial}{\partial\tau_1}P_m^2 + \mu_1\frac{\partial}{\partial\tau_1}(\tau_1-\tau_2) + \mu_2\frac{\partial}{\partial\tau_1}(0.5-\sigma_{T_1}) + \mu_3\frac{\partial}{\partial\tau_1}(0.5-\sigma_{F_1}) \\
&\qquad + \mu_4\frac{\partial}{\partial\tau_1}(\sigma_{T_1}-1) + \mu_5\frac{\partial}{\partial\tau_1}(\sigma_{F_1}-1) = 0, \qquad (6.16)\\
&\frac{\partial}{\partial\tau_2}P_m^2 + \mu_1\frac{\partial}{\partial\tau_2}(\tau_1-\tau_2) + \mu_2\frac{\partial}{\partial\tau_2}(0.5-\sigma_{T_1}) + \mu_3\frac{\partial}{\partial\tau_2}(0.5-\sigma_{F_1}) \\
&\qquad + \mu_4\frac{\partial}{\partial\tau_2}(\sigma_{T_1}-1) + \mu_5\frac{\partial}{\partial\tau_2}(\sigma_{F_1}-1) = 0, \qquad (6.17)\\
&\frac{\partial}{\partial\mathbf{c}}P_m^2 + \lambda_1\frac{\partial}{\partial\mathbf{c}}(\mathbf{c}^T\mathbf{c}-1) + \mu_1\frac{\partial}{\partial\mathbf{c}}(\tau_1-\tau_2) + \mu_2\frac{\partial}{\partial\mathbf{c}}(0.5-\sigma_{T_1}) \\
&\qquad + \mu_3\frac{\partial}{\partial\mathbf{c}}(0.5-\sigma_{F_1}) + \mu_4\frac{\partial}{\partial\mathbf{c}}(\sigma_{T_1}-1) + \mu_5\frac{\partial}{\partial\mathbf{c}}(\sigma_{F_1}-1) = 0. \qquad (6.18)
\end{aligned}$$

Figure 6.5 and 6.6 illustrate a numerical example of optimal linear mixed-initiative

nested thresholding.

While the MATLAB optimization routine '*fmincon*' is suitable to solve the problem since it can handle nonlinear constraints, often the solution either converges to one of the many local minima or to infeasible solutions. Optimization routines that exploit randomness, such as Simulated Annealing [130] or Genetic Algorithms [131], can be used to find the global minimum.

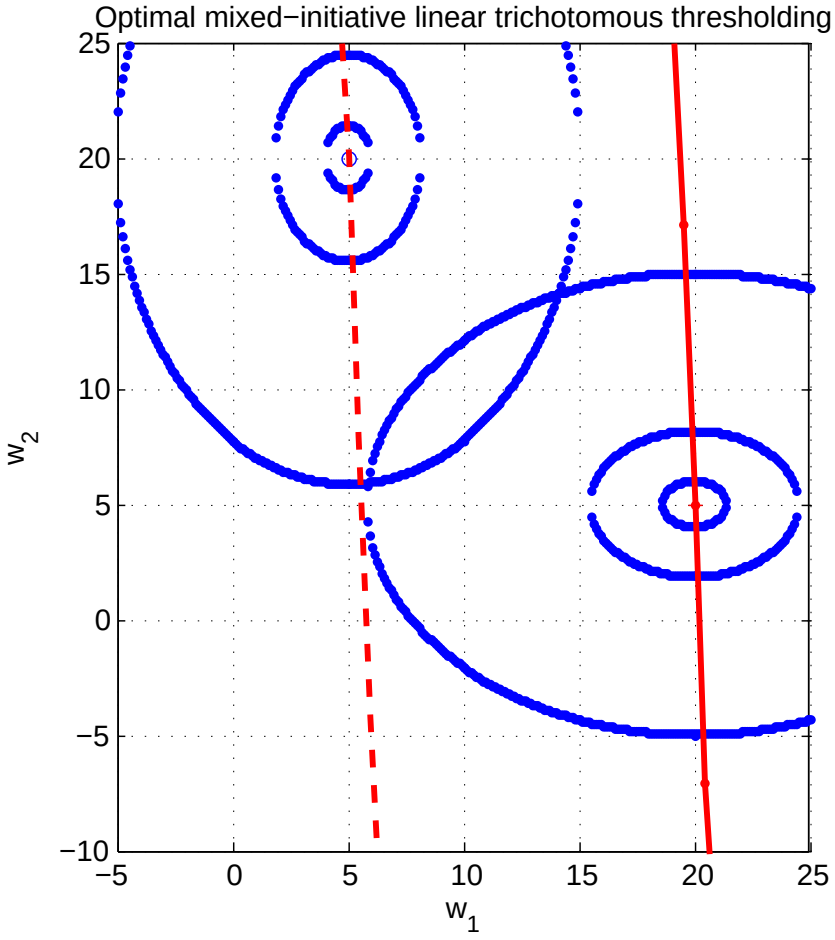


Figure 6.5: Optimal mixed-initiative linear trichotomous thresholding for $\bar{\mathbf{w}}_T = [5,\ 20]$, $\bar{\mathbf{w}}_F = [20,\ 5]$, $P_{w_T} = \mathrm{diag}(10, 5)$, $P_{w_F} = \mathrm{diag}(5, 10)$, $\mathbf{c}_0 = [0.5,\ 0.5]$, $\tau_0 = [-20,\ 20]$. The optimum is at $\mathbf{c}^* = [0.991,\ 0.0412]$, $\tau^* = [5.82,\ 20.19]$, $P_m^* = 7.17 \cdot 10^{-10}$.

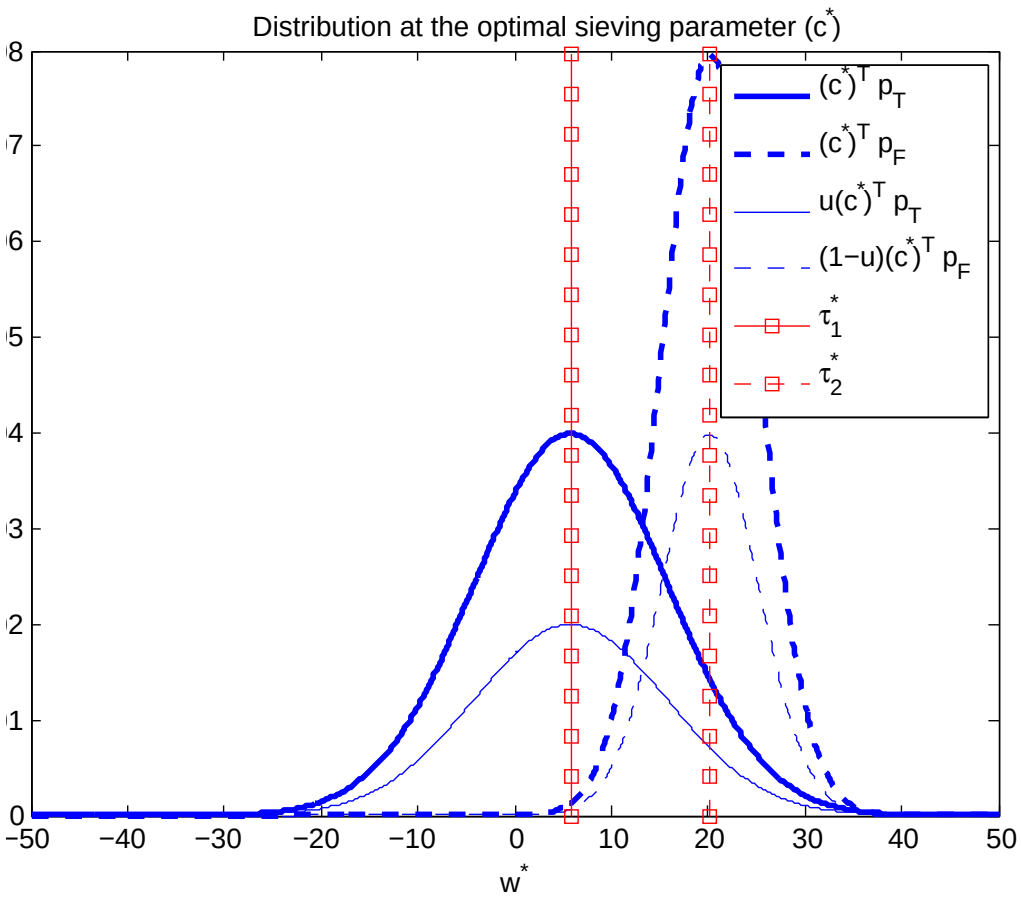


Figure 6.6: Distribution of $w^* = \mathbf{c}^*\mathbf{w}$

6.3 Mixed-initiative nested thresholding for n team members

Previously, we considered mixed-initiative nested thresholding with two classifiers: a primary trichotomous classifier and a secondary dichotomous classifier. In this section, we consider the same regime, but extend the number of classifiers involved and study the properties of such an architecture.

6.3.1 Problem formulation

Consider a nested classification architecture where the ratio of the number of workload-dependent classifiers to the number of workload-independent classifiers is a design variable. Figure 6.7 illustrates the concept. In one case, a workload-

independent classifier serves as a primary classifier and distributes any unclassifiable tasks to secondary workload-dependent classifiers (left-hand side figure), while in another case, multiple primary classifiers process the incoming tasks and, when there are any unclassifiable tasks, deliver them to a secondary workload-dependent classifier (right-hand side figure).

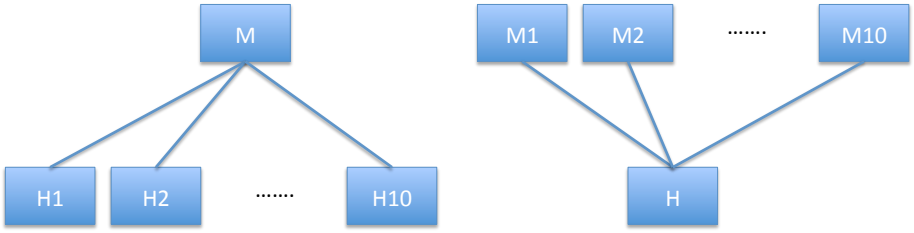


Figure 6.7: Mixed-initiative nested thresholding with more than two team members. (M denotes a workload-independent classifier and H denotes a workload-dependent classifier)

Note that the left-hand side theme in Fig. 6.7 is somewhat analogous to the current state-of-the-art mission operation in the U.S Air Force with unmanned vehicles: an unmanned aerial vehicle with several human operators involved in the operation and management of the vehicle. On the other hand, the right-hand side theme is similar to what the current operation is going towards: a number of unmanned aerial vehicles with a single human operator supervising.

There are several assumptions made for this study. First, when there are multiple classifiers, either primary or secondary, we assume that they are identical. Second, each classifier in the team works independently, that is, when a classification task is given to two classifiers, they examine the task solely according to their abilities without affecting each other.

Note that in this study we consider either the ratio of a single primary classifier to multiple secondary classifiers ($1 : n$), or the ratio of multiple primary classifiers to a secondary classifier ($n : 1$). Doing so, the analysis is simpler than considering

the general case of $m : n$ where allocating tasks becomes an issue. For our case, we assume that if a task is passed from a primary classifier to multiple secondary classifiers, it is uniformly distributed to each secondary classifier.

Let $n \in \{\frac{1}{m}, \frac{1}{m-1}, \cdots, \frac{1}{2}, 1, 2, \cdots, m\}$ denote the ratio of the number of workload-dependent classifiers to the number of workload-independent classifiers in the system with $m \in \mathcal{N}$. For example, $n = 0.1$ means that there is a single workload-dependent classifier and 10 workload-independent classifiers.

Once the primary layer of workload-independent classifiers receives measurements, the unclassifiable measurements are sent to the secondary layer of workload-dependent classifiers. We quantify all the unclassifiable measurements from a classifier in the primary layer as the *total workload*. Let $W \in [0, \infty)$ denote the total workload applied to the secondary layer of workload-dependent classifiers due to a workload-independent classifier in the primary layer. The total workload W is then uniformly distributed to the secondary layer classifiers such that each classifier has its own individual workload. Let

$$W_n = \frac{W}{n} \tag{6.19}$$

denote the individual workload applied to each workload-dependent classifier given the ratio n.

We consider the total workload W as a design variable. Note that although there can be more than two classifiers in the architecture, using the probability of misclassification for two classifiers as the cost function is still valid. This is because of the particular one-to-one setup between the classifiers in the primary and secondary layers, and the assumptions that the classifiers are identical and work independently.

The objective of the problem is to minimize the probability of misclassification by

choosing the ratio number n, i.e.,

$$\min_{n} P_m^2(W, n),$$

subject to inequality constraints,

$$\tau_1 \leq \tau_2 \tag{6.20a}$$

$$\sigma_{T_1} \geq 0.5, \tag{6.20b}$$

$$\sigma_{F_1} \geq 0.5, \tag{6.20c}$$

$$\sigma_{T_1} \leq 1, \tag{6.20d}$$

$$\sigma_{F_1} \leq 1. \tag{6.20e}$$

In words, the goal is to find the optimal ratio of the number of workload-dependent classifiers to the number of workload-independent classifiers to use in the architecture such that the probability of misclassification is minimized.

6.3.2 Optimal mixed-initiative nested thresholding for n members

Let us choose the following parameters as the independent variables: the mean and variance of the distribution of each sub-population $(m_{(\cdot)}, s^2_{(\cdot)})$, the prior probability (u), and the total workload (W). Once the independent variables are determined, we solve a double minimization problem with two sets of minimizers: the threshold variables (τ_1, τ_2) and the ratio number (n). Figure 6.8 shows the formal procedure of solving the problem.

Note that here the minimal probability of misclassification $P_m^*(W = \bar{W}, n^*)$, or simply P_m^*, is the best performance that can be achieved given the ratio number n. The ratio number n^* is optimal if the achievable probability of misclassification by n^* is minimal compared to that of other possible configurations in n.

Start;
Determine m_T, m_F, s_T^2, s_F^2;
for $W = [0,\ W_{max}]$ **do**
  for $n = [\frac{1}{10}, \frac{1}{9}, \cdots, 1, 2, \cdots, 10]$ **do**
    Solve $P_m^*(W = \bar{W}, n = \bar{n}) = \min_{\tau_1,\tau_2} P_m(W = \bar{W}, n = \bar{n})$
  end
  Solve $P_m^*(W = \bar{W}, n^*) = \min_n P_m^*(W = \bar{W}, n)$
end
Stop;

Figure 6.8: Algorithm for determining the optimal ratio n^* and the corresponding minimal probability of misclassification P_m^*

Figure 6.9, 6.10, and 6.11 show the optimal ratio n^*, the minimal probability of misclassification P_m^*, and the individual workload W_n, respectively, as a function of the total workload W.

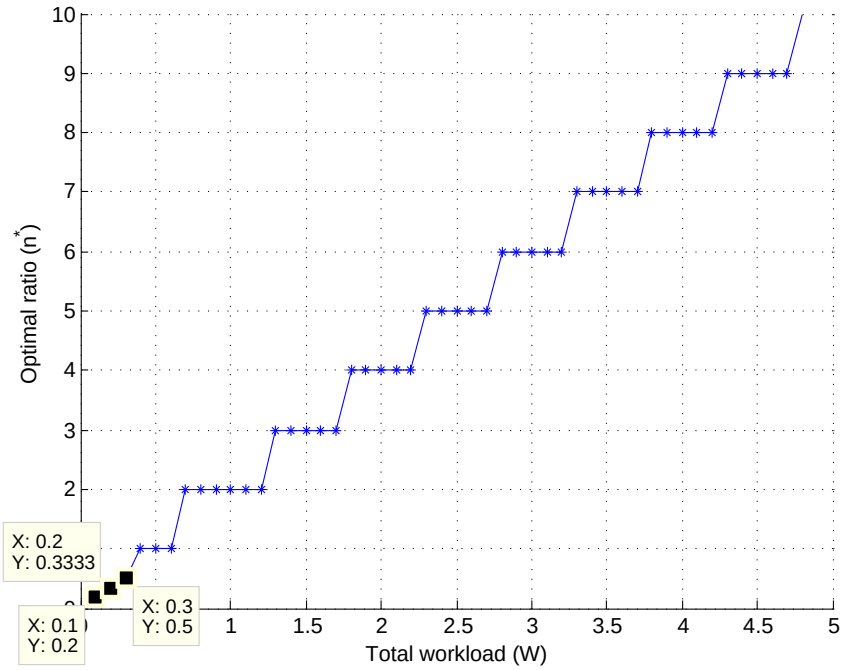


Figure 6.9: Optimal ratio as a function of the total workload ($u = 0.5, \sigma^* = 1$)

The optimal ratio with respect to the total workload in Fig. 6.9 shows that as the total workload increases, the optimal configuration is to increase the number of

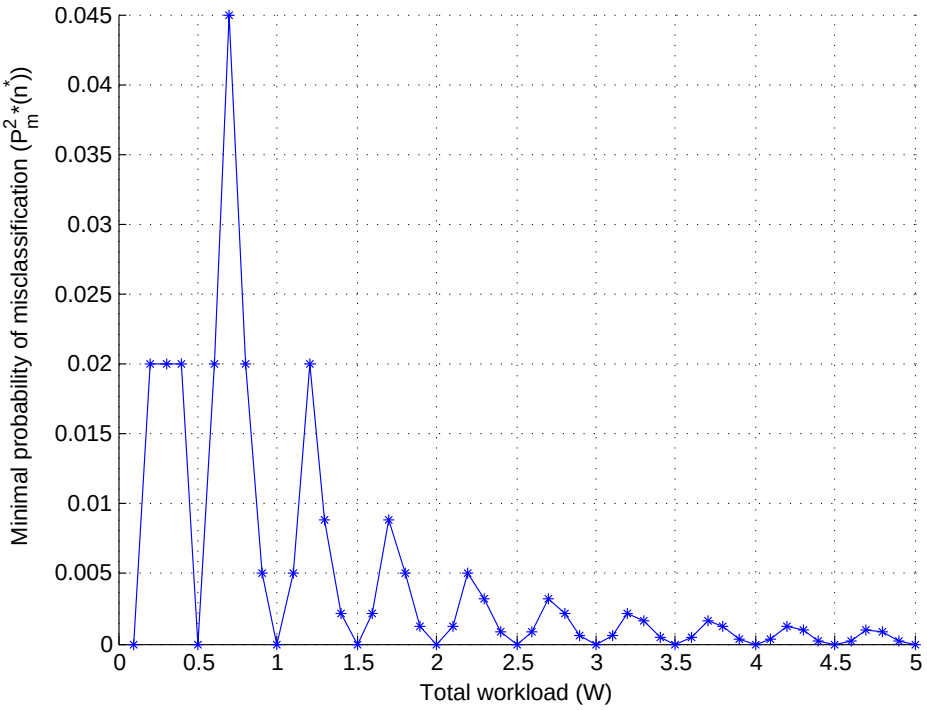


Figure 6.10: Minimal probability of misclassification as a function of the total workload ($u = 0.5, \sigma^* = 1$)

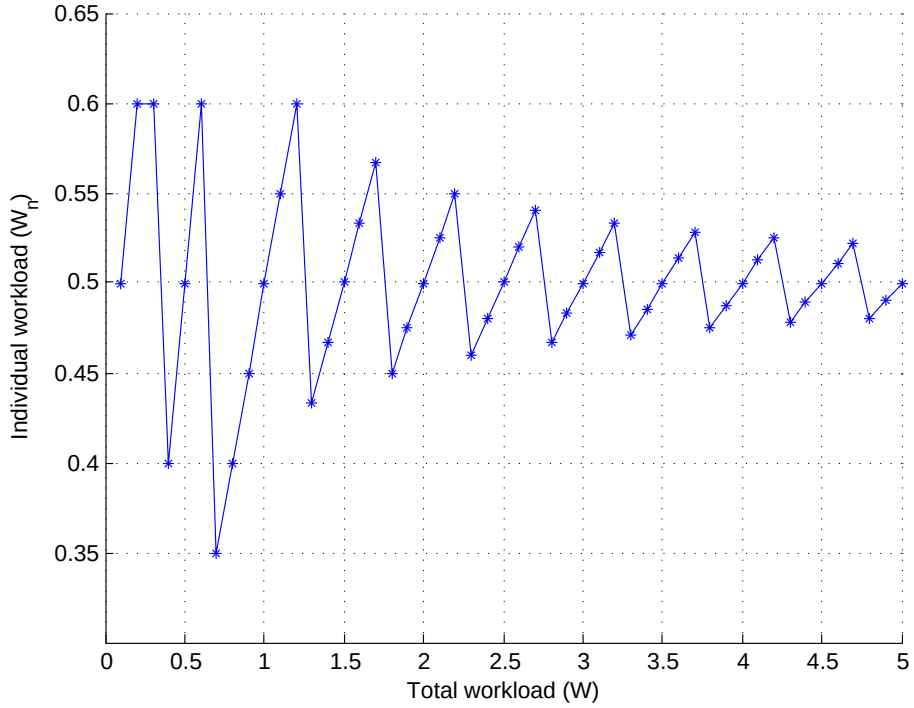


Figure 6.11: Individual workload as a function of the total workload ($u = 0.5, \sigma^* = 1$)

workload-dependent classifiers so that the individual workload is not overwhelming. This can be confirmed by the result of Fig. 6.11 in which the individual workload asymptotically reaches to 0.5, the workload that gives the optimal performance. The minimal probability of misclassification is asymptotically reaching to zero as the total workload increases, as shown in Fig. 6.10. Figure 6.12 is the subplot version of the previous three figures for comparison purpose.

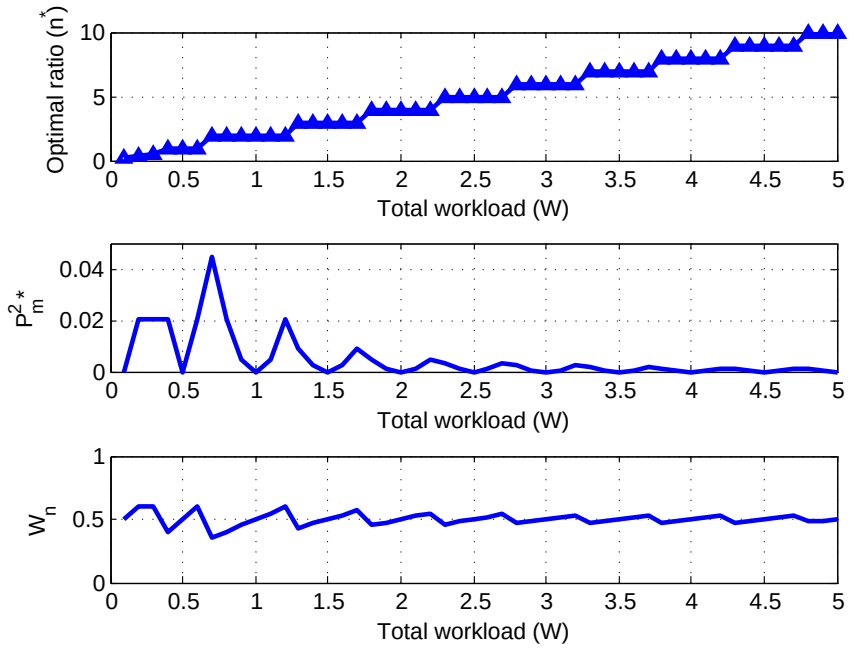


Figure 6.12: Optimal ratio, minimal probability of misclassification, individual workload as a function of the total workload ($u = 0.5, m_T = -20, m_F = 20, s_{(\cdot)} = 5, \sigma^* = 1$)

6.3.2.1 Sensitivity with respect to σ^*

In this section, we conduct a numerical sensitivity analysis of the results shown previously with respect to the maximum performance parameter σ^*. Recall that $\sigma^* \in [0.5,\ 1]$ determines the maximum of σ_i, $i \in \{T, F\}$. Since the individual classifier performance can hardly be perfect ($\sigma^* = 1$) in practice, it is reasonable to consider sensitivity analysis and examine how the optimal solutions change.

Figure 6.13, 6.14, and 6.15 show the optimal ratio n^*, the minimal probability of misclassification P_m^*, and the individual workload W_n, respectively, as a function of the total workload W. Note that the results show different cases of σ^* and the performance of the primary workload-independent classifier is arbitrarily fixed to 0.7, i.e., $\sigma_{T_1} = \sigma_{F_1} = 0.7$.

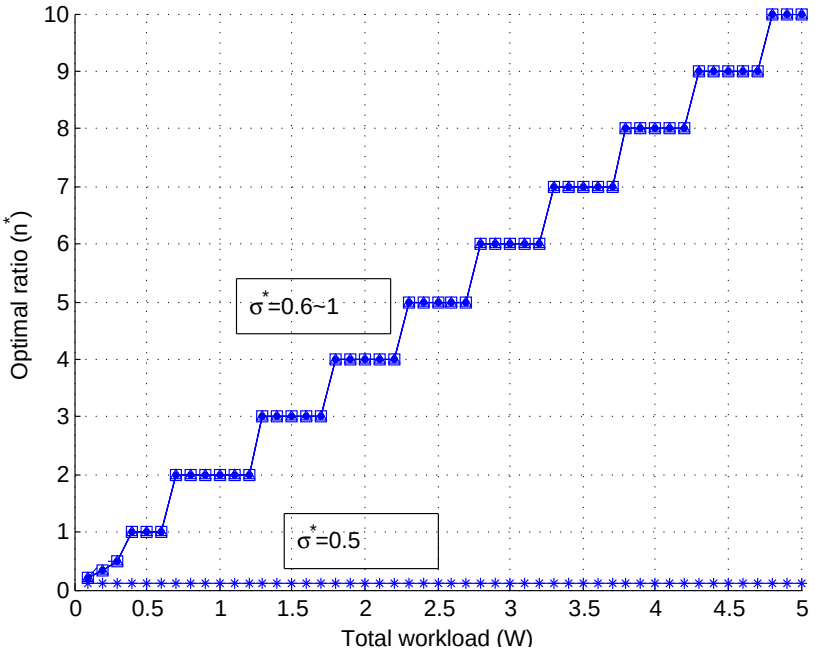


Figure 6.13: Optimal ratio as a function of the total workload ($u = 0.5, \sigma_{T_1} = \sigma_{F_1} = 0.7$)

A notable observation in Fig. 6.13 is that when the workload-dependent classifier is as bad as a pure guess ($\sigma^* = 0.5$), the optimal configuration is to maximize the number of the primary workload-independent classifiers while minimizing the number of workload-dependent classifiers ($n^* = \frac{1}{10}$). As a consequence, the individual workload for pure guessing is a linearly increasing function with respect to the total workload, as shown in Fig. 6.15. For the minimal probability of misclassification, as shown in in Fig. 6.14, we observe that the measure asymptotically reaches a value determined by the workload-dependent classifier performance, specifically $1 - \sigma^*$.

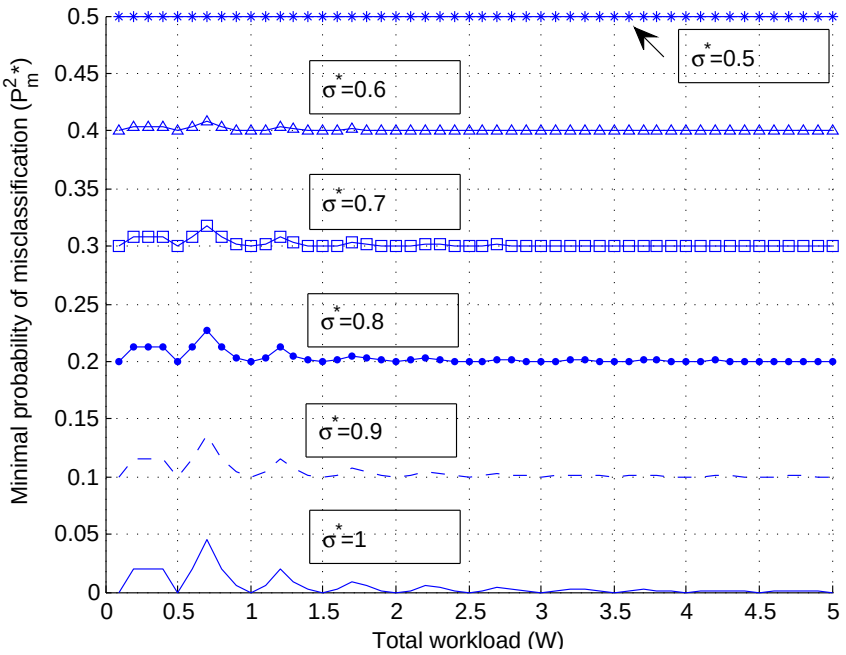


Figure 6.14: Minimal probability of misclassification as a function of the total workload ($u = 0.5, \sigma_{T_1} = \sigma_{F_1} = 0.7$)

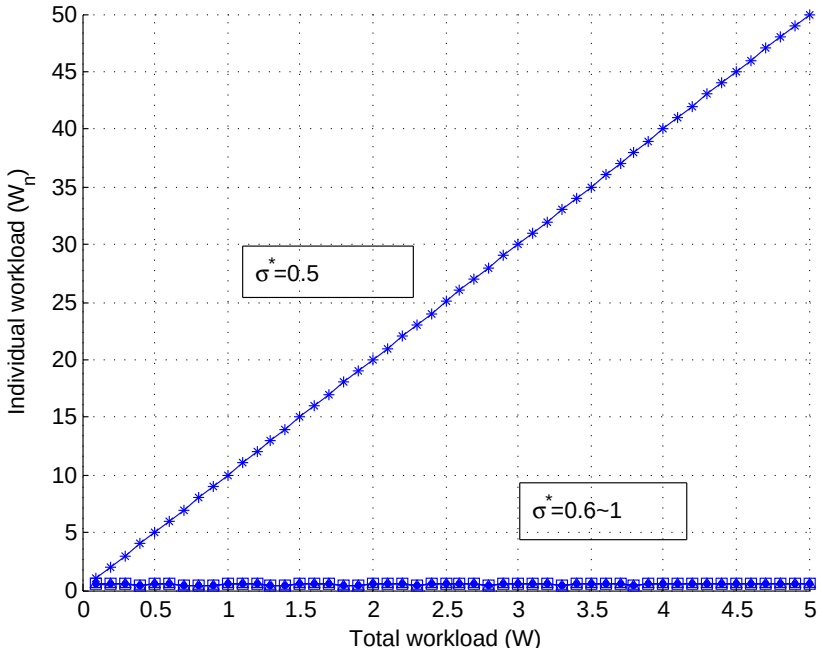


Figure 6.15: Individual workload as a function of the total workload ($u = 0.5, \sigma_{T_1} = \sigma_{F_1} = 0.7$)

6.4 Conclusion & future work

In this chapter, we have proposed a novel classifier architecture that uses a trichotomous classifier with workload-independent performance that turns over the data classified as unknown to a binary classifier with workload-dependent performance. We demonstrate that the novel classifier architecture gives superior classification performance (the probability of misclassification) compared to a single dichotomous classifier, relate the classifier's performance to the inherent difficulty of the classification task at hand (classifiability), and compare the performance of different classifiers.

As future work, identifying the important parameters in the problem of (linear) mixed-initiative nested thresholding will be addressed. This includes assessing the impact of the correctness of the prior information u on the optimal solution and recasting the problem with a dimensionless number so that solutions under various parameters can be described qualitatively by such number. Moreover, we will consider nested thresholding architecture with tertiary layers; the secondary classifier is trichotomous such that any unclassifiable tasks by the secondary classifier activates the tertiary mobile sensor that collects more data. Finally, as one of the assumptions that we made is that the knowledge of the distributions of w is provided by calibration, future work will address the case when the distributions are partially given.

CHAPTER VII

Epilogue

> I run in a void. Or maybe I should put it the other way: I run in order to *acquire* a void.
>
> Haruki Murakami

7.1 Summary

In this dissertation, we have studied classification along with information and team classification. In Chap. I, we provided the background that motivated the problem, the problem statement, and a list of original contributions of the dissertation. In Chap. II, we reviewed the background literature related to this work, specifically on the mathematical formulation of information, the problem of classification, and human factors. Prior to studying the mechanism of a classifier, we investigated the relationship between information and classification performance in Chap. III. We found that while there is a predominant congruence between information and classification performance, there is also independence between them, which was shown for both workload-independent and workload-dependent classifiers. In Chap. IV, we posed the problem of classification by thresholding and gave both analytical and numerical analyses of the solutions. Also, we have studied the change in the solution

when the property of the classifier changes: dichotomy vs. trichotomy, and single-variable vs. multivariate measurement. In Chap. V, we considered the case when there are multiple homogeneous classifiers and provided a supervisory strategy by using logical fusion rules. It was found that there is an optimal fusion rule that yields minimal probability of misclassification, and the optimal fusion rule changes as the available prior information changes. In Chap. VI, we considered multiple heterogeneous classifiers and proposed a novel classification architecture. It was shown that the nested architecture with a primary trichotomous classifier and a secondary dichotomous classifier outperforms a single dichotomous classifier.

7.2 Concluding remarks

7.2.1 Lessons learned

There are several messages that this dissertation delivers:

1. Gathering more information does not always decrease the probability of misclassification. The mechanism behind this phenomenon is the trade-off between the ability to sense truth out of truth (the rate of true positives, i.e., σ_T) and to sense falsehood out of falsehood (the rate of true negatives, i.e., σ_F).

2. For a dichotomous classifier with a single-variable measurement, if the distribution of a measurable property of each sub-population is Gaussian, then the optimal threshold is always at the intersection of the two distributions weighted by their prior information.

3. For a classifier with multivariate measurements, the solution can be found essentially in the same way as for a classifier with a single-variable measurement by adding an additional unknown, the sieving parameter.

4. For two homogeneous classifiers, there exists an optimal fusion rule that mini-

mizes the probability of misclassification and the optimal rule changes depending on the level of prior information at hand.

5. A nested architecture with two heterogeneous classifiers always outperforms a single dichotomous classifier.

7.2.2 Key contributions

- We show that increasing the amount of information, in the sense of Shannon's, *generally* implies improving classification performance, when classification decisions are made by the maximum likelihood rule and the classification performance is the probability of misclassification. We show the phenomenon for classifiers under two mechanisms: 1. workload-independent classifiers 2. workload-dependent classifiers. We demonstrate that, however, increasing the amount of information does *not always* imply improving classification performance, and that is indeed so for both classifiers with different mechanisms.

- We pose and solve the problem of trichotomous thresholding with a single variable measurement, where the classification decision is based on three options (true, false, or *unknown*) and determined by two thresholds. Then, we generalize the problem to a multivariate measurement and provide solutions.

- We propose a novel single and team classification model that depends on the individual classifier's confusion matrix and *a priori* information in a static environment. We show that the individual classifier's decision in the team can be fused by various logical operators and verify that the single classifier is a special case of the fused model. We show that there are fusion rules that improve the team performance compared to the individual performance.

- We propose a novel classifier architecture that uses a trichotomous classifier with workload-independent performance that turns over the data classified as un-

known to a binary classifier with workload-dependent performance. We demonstrate that the novel classifier architecture gives superior classification performance (the probability of misclassification) compared to a single dichotomous classifier. We relate the classifier's performance to the inherent difficulty of the classification task at hand (classifiability), and compare the performance of different classifiers.

7.3 Future directions

- **Performance measure**

  Our analysis throughout this work is based on the probability of misclassification as the single performance measure. As we have stated in Chap. I, there are other measures that may be considered in assessing the performance of classification, such as time-criticality and decision confidence.

  The time-criticality in classification becomes important in a mission where the situation changes rapidly. For instance, a threat may be able to move and hide away once it was spotted. One way to address the time-criticality is to consider each object of interest that is subject to classification as a task, and a classifier as a server that services tasks. A service time for a classifier is the time counted from when it receives to when it finishes the task. Queueing theory [132] can be a good candidate to formalize the problem.

  A simpler way of considering time-criticality is by using workload. Since workload can be an indirect indication of how much time the classifier needs to spend on an object of interest, a new cost function can be formalized as a convex combination of the probability of misclassification and the workload of the classifier.

An optimization problem with an objective function as

$$\min_{\tau} \lambda \cdot P_m + (1 - \lambda) \cdot W,$$

can be posed, where $\lambda \in [0, 1]$.

Another important performance measure is the decision confidence. Often in practice, it is very difficult to assess whether the classification is right or wrong because in order to do so, we need to know the ground truth. On the other hand, decision confidence is a measure that can be assessed based on the number of measurements, situation awareness, and etc. Thus, decision confidence can be more practical than the probability of misclassification as the performance measure.

- **Kinematic classification**

  The problem of kinematic classification arises when a classifier is able to move (thus, *kinematic classification*). Suppose that a mobile agent is located in the same area where the objects of interest are located. One of the subsystems of the agent provides information with respect to the objects where the information quality is dependent on the relative position between the agent and the objects, such as the range and/or the relative azimuth. Based on the collected information, the classification subsystem of the agent decides on the object property. Since the classification performance is determined by the information quality and the quality of information is dictated by the relative position between the agent and the object, the agent plans a path such that the probability of misclassification is minimized.

  To incorporate the fact that the information quality is dependent on the relative position between the agent and the object, let the sensor performance be

quantified by,

$$
\begin{aligned}
P(Y^i = Y_1 | X = T) &= 1 - \sigma_{T_i}, \\
P(Y^i = Y_2 | X = T) &= \sigma_{T_i}, \\
P(Y^i = Y_1 | X = F) &= \sigma_{F_i}, \\
P(Y^i = Y_2 | X = F) &= 1 - \sigma_{F_i}.
\end{aligned} \tag{7.1}
$$

The index $i \in \mathcal{N}$ represents the i-th sampling instance of the sensing device. Note that each $\sigma_{(\cdot)}$ is now a function of the relative position and azimuth between the mobile agent and the object.

Let ϕ denote a steering variable that determines the agent's motion. The goal of the problem is to solve an optimization problem with an objective function,

$$\min_{\phi} P_m.$$

As future work, we address two aspects of the problem of kinematic classification that are,

1. Kinematic classification (free measurements),
2. Costly kinematic classification (costly measurements).

- **Classification with learning**

  When situation changes rapidly, classifiers with learning can accommodate such changes. Many developers of classifiers put emphasis on the learning aspect as the classification task becomes more complex and time-varying. Our approach can be formalized with learning as follows:

  Let $w \in \mathcal{R}$ and $X \in \{T, F\}$. Calibration provides a pair of data $\{w_k, X\}$ where the subscript k denotes the sampling instants. Given the sequence of pairs of

data up to the instant k, the mean and the variance are learned as

$$\hat{m}_{T_k} = \frac{1}{k}\sum_{i=1}^{k} w_i, \tag{7.2}$$

$$\hat{s}^2_{T_k} = \frac{1}{k-1}\sum_{i=1}^{k}[w_i - \hat{m}_{T_k}]^2. \tag{7.3}$$

The mean and the variance for $X = F$ can be learned similarly. Further investigation of the learning aspect is left as future work.

- **Deceptive strategies**

  Deception is an effort to cause wrong-doings in the opponent. Deceptive strategies have been studied, in missile guidance for instance [129], and sometimes they are effective. Knowing the mechanism of a classifier and the role of information in classification, a deception problem can be posed. Assessing the impact of wrong calibration and wrong measurement in classification will help us understand how one should devise and counter a deceptive strategy.

7.4 A list of publications

The thesis is a comprehensive work based on the following papers that have been published in journals and presented at conferences:

Journal Articles

(J1) B. Hyun, P. Kabamba, A. Girard, "Optimally-informative path planning for dynamic Bayesian classification," *Optimization Letters,* Springers, 2011, Accepted for publication.

(J2) B. Hyun, C.J. Park, W. Wang, A.R. Girard, "Heterogeneous human operator team in classification tasks: modeling and supervisory control using Discrete Event Systems," *Special Issue for the Proceedings of IEEE,* 2011, Accepted for

publication.

(J3) B. Hyun, M. Faied, P. Kabamba, A. Girard, "On the Independence of information and classification performance," *IEEE Transactions on Systems, Man, and Cybernetics - Part B,* 2011, Submitted.

(J4) A. Klesh, B. Hyun, T. Huntsberger, G. Woodward, P. Kabamba, A. Girard, "Tactical area search with strapped-down anisotropic sensors," *Journal of Advanced Robotics,* 2011, Submitted.

(J5) B. Hyun, M. Faied, P. Kabamba, A. Girard, "Mixed-initiative nested classification - a study of classifiablity, thresholding, and workload," *IEEE Transactions on Systems, Man, and Cybernetics - Part B,* 2011, In preparation.

Conference Proceedings

(C1) B. Hyun, J. Jackson, A. Klesh, A. Girard, P. Kabamba, "Robotic exploration with non-isotropic sensors," *In Proceedings of AIAA Guidance, Navigation, and Control Conference and Exhibit,* Chicago, IL, Aug. 2009.

(C2) B. Hyun, C.J. Park, W. Wang, A. Girard, "Discrete event modeling of heterogeneous human operator team in classification task," *In Proceedings of American Control Conference,* Baltimore, MD, Jun. 2010.

(C3) B. Hyun, P. Kabamba, A. Girard, "Sequential Bayesian classification decisions for mobile sensors," *In Proceedings of IEEE Conference on Decision and Control,* Atlanta, GA, Dec. 2010.

(C4) B. Hyun, P. Kabamba, M. Faied, A. Girard, "Classification with synergistic teams," *In Proceedings of 18th World Congress of the International Federation of Automatic Control (IFAC),* Milan, Italy, Aug. 2011

(C5) B. Hyun, M. Faied, P. Kabamba, A. Girard, "On the independence of information and classification performance for workload-independent classifiers," *IEEE Conference on Decision and Control,* Orlando, FL, 2011, Submitted.

(C6) B. Hyun, M. Faied, P. Kabamba, A. Girard, "Mixed-initiative nested classification by optimal thresholding," *IEEE Conference on Decision and Control,* Orlando, FL, 2011, Submitted.

Abstracts

(A1) B. Hyun, P. Kabamba, A. Girard, "Optimally-informative path planning for dynamic Bayesian classification," *2nd International Conference on the Dynamics of Information Systems,* Feb 2010.

(A2) B. Hyun, L. Bertuccelli, N. Beckers, "Workload assessment in search scheduling using blink rate," *2010 HFES Student Research Conference, New England Chapter,* Cambridge, MA, Oct 2010.

Posters

(P1) B. Hyun, P. Kabamba, A. Girard, "Sequential Bayesian classification decisions for mobile sensors," *Michigan/AFRL Collaborative Center in Control Science (MACCCS) annual review meeting,* Ann Arbor, MI, Aug 2010.

(P2) B. Hyun, A. Girard, B. Kuipers, "Learning-to-Grasp: from an Infant to a Troublemaker," *IEEE the 9th International Conference on Development and Learning,* Aug 2010.

APPENDICES

APPENDIX A

YERKES-DODSON LAW

Unlike machine classifiers, human operator performance is subject to various human factors, such as workload, fatigue, boredom, stress, etc. Here, we model the human as a workload-dependent classifier. The workload-dependence is depicted by the Yerkes-Dodson law [110] that states that there is an optimal region of workload that allows humans to exhibit a maximum performance. Figure A.1 illustrates the concept.

Note that the Yerkes-Dodson law is not a definitive rule, meaning that depending on human subjects and situations, the performance-workload relationship may exhibit a different trend.

We model the Yerkes-Dodson law as a quadratic function of the workload throughout the thesis. The model is derived based on several assumptions. Let $f : [0, 1] \rightarrow [0.5, 1]$ be a quadratic function such that

$$\begin{aligned} \sigma &= f(W) \\ &= aW^2 + bW + c, \end{aligned} \tag{A.1}$$

where $\sigma \in [0.5, 1]$ is the performance variable, $W \in [0, 1]$ is the workload variable and $a, b, c \in \mathcal{R}$ are some coefficients. The three unknown coefficients are determined by the following assumptions:

- When the workload is either at its minimum or maximum, the performance is at minimum, i.e.,

$$f(W = 0) = 0.5, \tag{A.2a}$$

$$f(W = 1) = 0.5. \tag{A.2b}$$

- The optimal performance $\sigma^* \in [0.5, 1]$ is obtained at the median of the workload range, i.e.,

$$f(W = 0.5) = \sigma^*. \tag{A.3}$$

Given three unknowns and three equations, the coefficients a, b, c can be uniquely determined. The Yerkes-Dodson law is modeled as follows:

$$\sigma = -(4\sigma^* - 2)W^2 + (4\sigma^* - 2)W + 0.5. \tag{A.4}$$

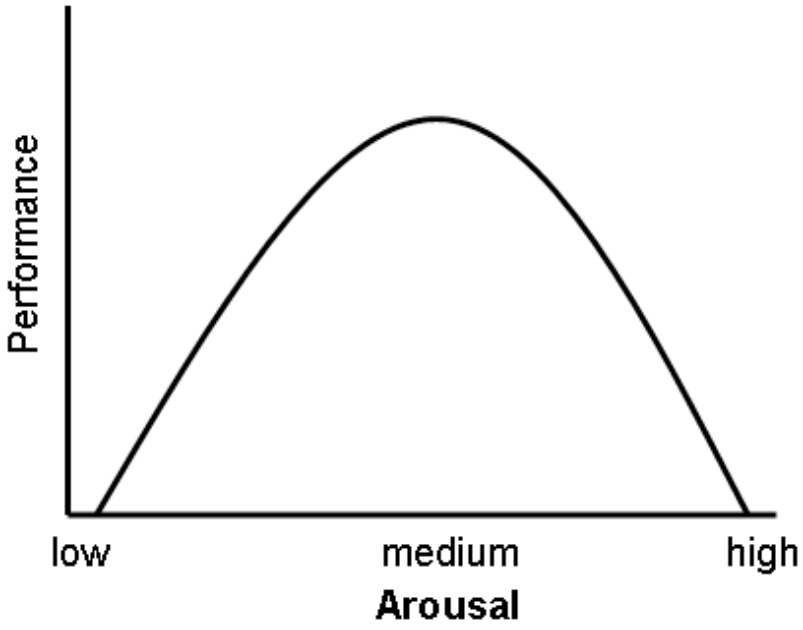


Figure A.1: Illustration of the Yerkes-Dodson law

APPENDIX B

PROOFS FOR CHAPTER III

First, we show that the following is true.

Lemma B.1. *If* $0.5 \leq \sigma_i \leq 1$ *for* $i \in \{T, F\}$ *and* $0 \leq u \leq 1$*, then* $f_1 \leq f_2$ *for all* σ_i *and* u*.*

Proof. The ratio of f_1 to f_2 is given as

$$\frac{f_1}{f_2} = \left(\frac{1-\sigma_T}{\sigma_F}\right)\left(\frac{1-\sigma_F}{\sigma_T}\right). \tag{B.1}$$

Since $0.5 \leq \sigma_i \leq 1$ for $i \in \{T, F\}$, it is true that $\left(\frac{1-\sigma_i}{\sigma_i}\right) \leq 1$ holds. Therefore,

$$\frac{f_1}{f_2} \leq 1 \Rightarrow f_1 \leq f_2. \tag{B.2}$$

□

Proof for Theorem III.6

Assume u is fixed. By the multivariable chain rule,

$$\Delta I(X;Y) = \frac{\partial I(X;Y)}{\partial \sigma_T}\Delta\sigma_T + \frac{\partial I(X;Y)}{\partial \sigma_F}\Delta\sigma_F. \tag{B.3}$$

The partial derivative of $I(X;Y)$ with respect to σ_T is given as

$$\frac{\partial I(X;Y)}{\partial \sigma_T} = u \log\left(\frac{1+1/f_1}{1+1/f_2}\right), \tag{B.4}$$

while the partial derivative of $I(X;Y)$ with respect to σ_F is given as

$$\frac{\partial I(X;Y)}{\partial \sigma_F} = (1-u) \log\left(\frac{1+f_2}{1+f_1}\right). \tag{B.5}$$

Due to Lemma B.1, it is true that

$$\frac{\partial I(X;Y)}{\partial \sigma_i} \geq 0, \tag{B.6}$$

holds for all σ_i and u for $i \in \{T, F\}$. Therefore, $I(X;Y)$ is a monotonically increasing function with respect to σ_T and σ_F. □

The following is a direct consequence of Lemma B.1.

Corollary B.2. *$f_1 > 1 \wedge f_2 \leq 1$ is false for all σ_i and u for $i \in \{T, F\}$.*

Proof for Theorem III.7

Depending on the range of f_1 and f_2, P_m can be expressed as

$$P_m = \begin{cases} 1-u & \text{if } f_1 > 1 \wedge f_2 > 1, \\ (1-\sigma_F)(1-u) + (1-\sigma_T)u & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ \sigma_F(1-u) + \sigma_T u & \text{if } f_1 > 1 \wedge f_2 \leq 1, \\ u & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{B.7}$$

Due to Corollary B.2, the results for the condition $f_1 > 1 \wedge f_2 \leq 1$ do not hold. With careful examinations, it can be shown that P_m is continuous at the boundary conditions, i.e., $f_1 = 1$ and $f_2 = 1$.

The partial derivatives of P_m with respect to σ_T and σ_F are given as

$$\frac{\partial P_m}{\partial \sigma_T} = \begin{cases} 0 & \text{if } f_1 > 1 \wedge f_2 > 1, \\ -u & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ 0 & \text{if } f_1 \leq 1 \wedge f_2 \leq 1, \end{cases} \tag{B.8}$$

and

$$\frac{\partial P_m}{\partial \sigma_F} = \begin{cases} 0 & \text{if } f_1 > 1 \wedge f_2 > 1, \\ -1 + u & \text{if } f_1 \leq 1 \wedge f_2 > 1, \\ 0 & \text{if } f_1 \leq 1 \wedge f_2 \leq 1. \end{cases} \tag{B.9}$$

Thus,

$$\frac{\partial P_m}{\partial \sigma_i} \leq 0, \tag{B.10}$$

for all σ_i and u for $i \in \{T, F\}$. Therefore, P_m is a monotonically decreasing function with respect to σ_T and σ_F. □

APPENDIX C

ANALYTICAL SOLUTIONS OF GAUSSIAN CUMULATIVE PROBABILITY DISTRIBUTION

Dichotomous thresholding

Let $\tau \in \mathcal{R}$ be the threshold variable. Let the rates of true positives and negatives be evaluated as:

$$\sigma_T = \int_{-\infty}^{\tau} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{C.1a}$$

$$\sigma_F = \int_{\tau}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{C.1b}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, and $c_i = \sqrt{2s_i^2}$, $i \in \{T, F\}$. The closed-form solutions to Eq. (C.1) are

$$\sigma_T = \begin{cases} \lim_{w\to-\infty} \left(-\frac{1}{2} a_T c_T \sqrt{\pi} \left(\operatorname{erf}\left(\frac{w+b_T}{c_T}\right) - \operatorname{erf}\left(\frac{|\tau|+b_T}{c_T}\right)\right)\right) & \text{if } \tau > 0, \\ \lim_{w\to-\infty} \left(-\frac{1}{2} a_T c_T \sqrt{\pi} \left(\operatorname{erf}\left(\frac{w+b_T}{c_T}\right) + \operatorname{erf}\left(\frac{|\tau|-b_T}{c_T}\right)\right)\right) & \text{if } \tau < 0, \\ \lim_{w\to-\infty} \left(\frac{1}{2} a_T c_T \sqrt{\pi} \left(-\operatorname{erf}\left(\frac{w+b_T}{c_T}\right) + \operatorname{erf}\left(\frac{b_T}{c_T}\right)\right)\right) & \text{if } \tau = 0, \end{cases} \tag{C.2}$$

and

$$\sigma_F = \begin{cases} \lim_{w\to\infty}\left(\frac{1}{2}a_F c_F\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_F}{c_F}\right) - \text{erf}\left(\frac{|\tau|+b_F}{c_F}\right)\right)\right) & \text{if } \tau > 0, \\ \lim_{w\to\infty}\left(\frac{1}{2}a_F c_F\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_F}{c_F}\right) + \text{erf}\left(\frac{|\tau|-b_F}{c_F}\right)\right)\right) & \text{if } \tau < 0, \\ \lim_{w\to\infty}\left(-\frac{1}{2}a_F c_F\sqrt{\pi}\left(-\text{erf}\left(\frac{w+b_F}{c_F}\right) + \text{erf}\left(\frac{b_F}{c_F}\right)\right)\right) & \text{if } \tau = 0. \end{cases} \tag{C.3}$$

Trichotomous thresholding

Let $\tau_1 \in \mathcal{R}$ and $\tau_2 \in \mathcal{R}$ be the threshold variables such that the cumulative probability distributions are,

$$\sigma_T = \int_{-\infty}^{\tau_1} a_T e^{-(w+b_T)^2/c_T^2} dw, \tag{C.4a}$$

$$\sigma_F = \int_{\tau_2}^{\infty} a_F e^{-(w+b_F)^2/c_F^2} dw, \tag{C.4b}$$

where $a_i = 1/\sqrt{2\pi s_i^2}$, $b_i = -m_i$, and $c_i = \sqrt{2s_i^2}$ with $i \in \{T, F\}$. The closed-form solutions to Eq. (C.4) are

$$\sigma_T = \begin{cases} \lim_{w\to-\infty}\left(-\frac{1}{2}a_T c_T\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_T}{c_T}\right) - \text{erf}\left(\frac{|\tau_1|+b_T}{c_T}\right)\right)\right) & \text{if } \tau_1 > 0, \\ \lim_{w\to-\infty}\left(-\frac{1}{2}a_T c_T\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_T}{c_T}\right) + \text{erf}\left(\frac{|\tau_1|-b_T}{c_T}\right)\right)\right) & \text{if } \tau_1 < 0, \\ \lim_{w\to-\infty}\left(\frac{1}{2}a_T c_T\sqrt{\pi}\left(-\text{erf}\left(\frac{w+b_T}{c_T}\right) + \text{erf}\left(\frac{b_T}{c_T}\right)\right)\right) & \text{if } \tau_1 = 0, \end{cases} \tag{C.5}$$

$$\sigma_F = \begin{cases} \lim_{w\to\infty}\left(\frac{1}{2}a_F c_F\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_F}{c_F}\right) - \text{erf}\left(\frac{|\tau_2|+b_F}{c_F}\right)\right)\right) & \text{if } \tau_2 > 0, \\ \lim_{w\to\infty}\left(\frac{1}{2}a_F c_F\sqrt{\pi}\left(\text{erf}\left(\frac{w+b_F}{c_F}\right) + \text{erf}\left(\frac{|\tau_2|-b_F}{c_F}\right)\right)\right) & \text{if } \tau_2 < 0, \\ \lim_{w\to\infty}\left(-\frac{1}{2}a_F c_F\sqrt{\pi}\left(-\text{erf}\left(\frac{w+b_F}{c_F}\right) + \text{erf}\left(\frac{b_F}{c_F}\right)\right)\right) & \text{if } \tau_2 = 0. \end{cases} \tag{C.6}$$

APPENDIX D

DERIVATION OF THE PROBABILITY OF MISCLASSIFICATION FOR TWO CLASSIFIERS

Here we derive the probability of misclassification for two classifiers. From Chap. III, we have defined that the probability of misclassification is the sum of probabilities of two faulty outcomes: false positive and false negative:

$$P_m^2 = P(O_s = T \wedge X = F) + P(O_s = F \wedge X = T). \tag{D.1}$$

Note that the superscript 2 in P_m^2 is to denote "two" classifiers. Let $Y^i \in \{Y_1, Y_2\}$ denote the object property where $i \in 1, 2$ indicates the classifier number. For instance, $Y^1 = Y_2$ denotes an event where the classifier 1 have observed an object property Y_2.

Then, by the theorem of total probability, each term in Eq. (D.1) expands as follows:

$$\begin{aligned} P(O_s = T \wedge X = F) = {} & P(O_s = T \wedge X = F | Y^1 = Y_1 \wedge Y^2 = Y_1) P(Y^1 = Y_1 \wedge Y^2 = Y_1) \\ & + P(O_s = T \wedge X = F | Y^1 = Y_1 \wedge Y^2 = Y_2) P(Y^1 = Y_1 \wedge Y^2 = Y_2) \\ & + P(O_s = T \wedge X = F | Y^1 = Y_2 \wedge Y^2 = Y_1) P(Y^1 = Y_2 \wedge Y^2 = Y_1) \\ & + P(O_s = T \wedge X = F | Y^2 = Y_1 \wedge Y^2 = Y_2) P(Y^1 = Y_2 \wedge Y^2 = Y_2), \end{aligned} \tag{D.2}$$

and

$$\begin{aligned} P(O_s = F \wedge X = T) = {} & P(O_s = F \wedge X = T | Y^1 = Y_1 \wedge Y^2 = Y_1) P(Y^1 = Y_1 \wedge Y^2 = Y_1) \\ & + P(O_s = F \wedge X = F | Y^1 = Y_1 \wedge Y^2 = Y_2) P(Y^1 = Y_1 \wedge Y^2 = Y_2) \\ & + P(O_s = F \wedge X = T | Y^1 = Y_2 \wedge Y^2 = Y_1) P(Y^1 = Y_2 \wedge Y^2 = Y_1) \\ & + P(O_s = F \wedge X = T | Y^2 = Y_1 \wedge Y^2 = Y_2) P(Y^1 = Y_2 \wedge Y^2 = Y_2). \end{aligned} \tag{D.3}$$

Assuming that the classifier decision O_s and the object status X are conditionally independent given two classifiers Y_1 and Y_2, i.e.,

$$\begin{aligned} & P(O_s = F \wedge X = T | Y^1 = Y_1 \wedge Y^2 = Y_1) = \\ & P(O_s = F | Y^1 = Y_1 \wedge Y^2 = Y_1) \cdot P(X = T | Y^1 = Y_1 \wedge Y^2 = Y_1), \end{aligned} \tag{D.4}$$

Eq. (D.1) can be expressed as

$$
\begin{aligned}
P_m^2 = {} & P(O_s = T | Y^1 = Y_1 \wedge Y^2 = Y_1) P(X = F \wedge Y^1 = Y_1 \wedge Y^2 = Y_1) \\
& + P(O_s = T | Y^1 = Y_1 \wedge Y^2 = Y_2) P(X = F \wedge Y^1 = Y_1 \wedge Y^2 = Y_2) \\
& + P(O_s = T | Y^1 = Y_2 \wedge Y^2 = Y_1) P(X = F \wedge Y^1 = Y_2 \wedge Y^2 = Y_1) \\
& + P(O_s = T | Y^1 = Y_2 \wedge Y^2 = Y_2) P(X = F \wedge Y^1 = Y_2 \wedge Y^2 = Y_2) \\
& + P(O_s = F | Y^1 = Y_1 \wedge Y^2 = Y_1) P(X = T \wedge Y^1 = Y_1 \wedge Y^2 = Y_1) \\
& + P(O_s = F | Y^1 = Y_1 \wedge Y^2 = Y_2) P(X = T \wedge Y^1 = Y_1 \wedge Y^2 = Y_2) \\
& + P(O_s = F | Y^1 = Y_2 \wedge Y^2 = Y_1) P(X = T \wedge Y^1 = Y_2 \wedge Y^2 = Y_1) \\
& + P(O_s = F | Y^1 = Y_2 \wedge Y^2 = Y_2) P(X = T \wedge Y^1 = Y_2 \wedge Y^2 = Y_2). \qquad \text{(D.5)}
\end{aligned}
$$

Let the conditional probabilities for each classifier defined as

$$
\begin{aligned}
& P(Y^i = Y_1 | X = T) = 1 - \sigma_{T_i}, \\
& P(Y^i = Y_2 | X = T) = \sigma_{T_i}, \\
& P(Y^i = Y_1 | X = F) = \sigma_{F_i}, \\
& P(Y^i = Y_2 | X = F) = 1 - \sigma_{F_i}, \\
& i \in \{1, 2\}. \qquad \text{(D.6)}
\end{aligned}
$$

The posterior probability $P(X = X_0 | Y^1 = Y_0^1 \wedge Y^2 = Y_0^2)$ is summarized in Table D.1.

Table D.1: Summary of the posterior probabilities for two subsequent measurements

X_0	Y_0^1	Y_0^2	$P(X = X_0 \vert Y^1 = Y_0^1 \wedge Y^2 = Y_0^2)$
T	Y_1	Y_1	$\frac{(1-\sigma_{T_1})(1-\sigma_{T_2})u}{(1-\sigma_{T_1})(1-\sigma_{T_2})u+\sigma_{F_1}\sigma_{F_2}(1-u)}$
F	Y_1	Y_1	$\frac{\sigma_{F_1}\sigma_{F_2}(1-u)}{(1-\sigma_{T_1})(1-\sigma_{T_2})u+\sigma_{F_1}\sigma_{F_2}(1-u)}$
T	Y_1	Y_2	$\frac{(1-\sigma_{T_1})\sigma_{T_2}u}{(1-\sigma_{T_1})\sigma_{T_2}u+\sigma_{F_1}(1-\sigma_{F_2})(1-u)}$
F	Y_1	Y_2	$\frac{\sigma_{F_1}(1-\sigma_{F_2})(1-u)}{(1-\sigma_{T_1})\sigma_{T_2}u+\sigma_{F_1}(1-\sigma_{F_2})(1-u)}$
T	Y_2	Y_1	$\frac{(1-\sigma_{T_2})\sigma_{T_1}u}{(1-\sigma_{T_2})\sigma_{T_1}u+\sigma_{F_2}(1-\sigma_{F_1})(1-u)}$
F	Y_2	Y_1	$\frac{\sigma_{F_2}(1-\sigma_{F_1})(1-u)}{(1-\sigma_{T_2})\sigma_{T_1}u+\sigma_{F_2}(1-\sigma_{F_1})(1-u)}$
T	Y_2	Y_2	$\frac{\sigma_{T_1}\sigma_{T_2}u}{\sigma_{T_1}\sigma_{T_2}u+(1-\sigma_{F_1})(1-\sigma_{F_2})(1-u)}$
F	Y_2	Y_2	$\frac{(1-\sigma_{F_1})(1-\sigma_{F_2})(1-u)}{\sigma_{T_1}\sigma_{T_2}u+(1-\sigma_{F_1})(1-\sigma_{F_2})(1-u)}$

Let $f_{Y_0^1,Y_0^2} \in [0,\, \infty]$ denote the ratio of the posterior probabilities such that,

$$
\begin{aligned}
f_{1,1} &= \left(\frac{1-\sigma_{T_1}}{\sigma_{F_1}}\right)\left(\frac{1-\sigma_{T_2}}{\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right),\\
f_{1,2} &= \left(\frac{1-\sigma_{T_1}}{\sigma_{F_1}}\right)\left(\frac{\sigma_{T_2}}{1-\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right),\\
f_{2,1} &= \left(\frac{\sigma_{T_1}}{1-\sigma_{F_1}}\right)\left(\frac{1-\sigma_{T_2}}{\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right),\\
f_{2,2} &= \left(\frac{\sigma_{T_1}}{1-\sigma_{F_1}}\right)\left(\frac{\sigma_{T_2}}{1-\sigma_{F_2}}\right)\left(\frac{u}{1-u}\right).
\end{aligned}
\tag{D.7}
$$

Let $\delta_{O_{s0}} : \mathcal{R} \to \{0, 1\}$ such that

$$
\delta_T(f) = \delta_{O_s=T}(f) = \begin{cases} 1 & \text{if } f > 1 \\ 0 & \text{if } f \leq 1, \end{cases} \tag{D.8a}
$$

$$
\delta_F(f) = \delta_{O_s=F}(f) = \begin{cases} 1 & \text{if } f \leq 1 \\ 0 & \text{if } f > 1. \end{cases} \tag{D.8b}
$$

The conditional probabilities $P(O_s = O_{s0} \vert Y^1 = Y_0^1 \wedge Y^2 = Y_0^2)$ are summarized in Table D.2.

Table D.2: Summary of the conditional probabilities for two classifiers

O_s	Y_0^1	Y_0^2	$P(O_s = O_{s0} \mid Y^1 = Y_0^1 \wedge Y^2 = Y_0^2)$
T	Y_1	Y_1	$\delta_T(f_{1,1})$
F	Y_1	Y_1	$\delta_F(f_{1,1})$
T	Y_1	Y_2	$\delta_T(f_{1,2})$
F	Y_1	Y_2	$\delta_F(f_{1,2})$
T	Y_2	Y_1	$\delta_T(f_{2,1})$
F	Y_2	Y_1	$\delta_F(f_{2,1})$
T	Y_2	Y_2	$\delta_T(f_{2,2})$
F	Y_2	Y_2	$\delta_F(f_{2,2})$

Substituting the results in Table D.1 and D.2 into Eq. (D.5) yields

$$\begin{aligned} P_m^2 = {} & \delta_T(f_{1,1})\sigma_{F_1}\sigma_{F_2}(1-u) + \delta_T(f_{1,2})\sigma_{F_1}(1-\sigma_{F_2})(1-u) \\ & + \delta_T(f_{2,1})\sigma_{F_2}(1-\sigma_{F_1})(1-u) + \delta_T(f_{2,2})(1-\sigma_{F_1})(1-\sigma_{F_2})(1-u) \\ & + \delta_F(f_{1,1})(1-\sigma_{T_1})(1-\sigma_{T_2})u + \delta_F(f_{1,2})(1-\sigma_{T_1})\sigma_{T_2}u \\ & + \delta_F(f_{2,1})(1-\sigma_{T_2})\sigma_{T_1}u + \delta_F(f_{2,2})\sigma_{T_1}\sigma_{T_2}u. \end{aligned} \tag{D.9}$$

The expression can be reformulated in matrix form as

$$P_m^2 = \bar{\sigma}_1^T R_2 \bar{\sigma}_2, \tag{D.10}$$

where

$$\bar{\sigma}_i = \begin{bmatrix} \sigma_{F_i} & 1-\sigma_{F_i} & 1-\sigma_{T_i} & \sigma_{T_i} \end{bmatrix}^T, \; i = 1, 2$$

$$R_2 = \begin{bmatrix} \delta_T(f_{1,1})(1-u) & \delta_T(f_{1,2})(1-u) & 0 & 0 \\ \delta_T(f_{2,1})(1-u) & \delta_T(f_{2,2})(1-u) & 0 & 0 \\ 0 & 0 & \delta_F(f_{1,1})u & \delta_F(f_{1,2})u \\ 0 & 0 & \delta_F(f_{2,1})u & \delta_F(f_{2,2})u \end{bmatrix}.$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] United States Air Force Chief Scientist (AF/ST). Report on technology horizons - a vision for air force science & technology during 2010-2030, 2010. AF/ST-TR-10-01-PR.

[2] Department of Defense. Fy2009-2034 unmanned systems integrated roadmap, 2009.

[3] U.S. Army UAS Center of Excellence. Eyes of the army, u.s. army roadmap for unmanned aircraft systems 2010-2035, 2010.

[4] S.J. Zaloga, D. Rockwell, and P. Finnegan. World unmanned aerial vehicle systems, market profile and forecast, 2008.

[5] C.E. Nehme. *Modeling human supervisory control in heterogeneous unmanned vehicle systems*. PhD thesis, Department of Aeronautics and Astronautics, MIT, 2009.

[6] J.S. McCarley and C.D. Wickens. Human factors implications of uavs in the national airspace. Technical report, University of Illinois at Urbana-Champaign, 2005.

[7] A. Dimoka, G. Adomavicius, A. Gupta, and P.A. Pavlou. Reducing the cognitive overload in continuous combinatorial auctions: evidences from an fmri study. Temple University, Working paper, 2011.

[8] P.T. Kabamba, S.M. Meerkov, and F.H. Zeitz. Optimal path planning for unmanned combat aerial vehicles to defeat radar tracking. *Journal of Guidance, Control, and Dynamics*, 29(2), Mar.-Apr. 2006.

[9] Nazih N. Youssef. Radar cross section of complex targets. *Proceedings of IEEE*, 77(5), May 1989.

[10] W. Ross Stone. *Radar cross sections of complex objects*. IEEE Press, 1989.

[11] James Gleick. *The information*. Pantheon Books, 2011.

[12] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, (27:379–423, 623–656), 1948.

[13] Harry Nyquist. Certain factors affecting telegraph speed. *Bell System Technical Journal*, 3:324–346, 1924.

[14] Katsuhiko Ogata. *Discrete-time control systems*. Prentice Hall, 2nd ed., 1995.

[15] Robert V. L. Hartley. Transmission of information. *Bell System Technical Journal*, pages 535–563, 1928.

[16] R. A. Fisher. *Contribution to mathematical statistics (collection of papers published 1920-1943)*. Wiley, New York, NY, 1950.

[17] B. Grocholsky. *Information-theoretic control of multiple sensor platforms*. PhD thesis, University of Sydney, Australia, 2006.

[18] B. R. Frieden. *Physics from Fisher information*. Cambridge University Press, Cambridge, UK, 1998.

[19] C. Cai. *Information-driven sensor path planning and the treasure hunt problem*. PhD thesis, Department of mechanical engineering and materials science, Duke university, 2008.

[20] S. Kullback. *Information theory and statistics*. Wiley, New York, 1959.

[21] A. Rényi. On measures of entropy and information. In *the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

[22] American Society of Photogrammetry. *Manual of photogrammetry*. Asprs Pubns, 4th ed., 1980.

[23] J.L. Crassidis and J.L. Junkins. *Optimal Estimation of Dynamic Systems*. Chapman & Hall/CRC, 2004.

[24] Merrill Skolnik. *Introduction to radar systems*. McGraw-Hill Science/Engineering/Math, 3rd ed., 2002.

[25] W.J. Larson and J.R. Wertz. *Space mission analysis and design*. Microcosm Press, 3rd ed., 1999.

[26] C. Cai, S. Ferrari, and M. Qian. Bayesian network modeling of acoustic sensor measurements. *IEEE Sensors Conference*, pages 345–348, 2007.

[27] G. Zhang, S. Ferrari, and M. Qian. Information roadmap method for robotic sensor path planning. *Journal of Intelligent & Robotic Systems*, 56(1-2):69–98, 2009.

[28] R. Platt, R. Tedrake, L.P. Kaelbling, and T. Lozano-Perez. Belief space planning assuming maximum likelihood observations. In *Proceedings of Robotics: Science and Systems*, 2010.

[29] Z. Kim and R. Sengupta. Target detection and position likelihood using an aerial image sensor, May 2008.

[30] K. Laventall and J. Corteés. Coverage control by multi-robot networks with limited-range anisotropic sensors. *International Journal of Control*, 9(1):1–9, 2009.

[31] A. Klesh, A. Girard, and P. Kabamba. Path planning for cooperative time-optimal information collection. In *Proceedings of American Control Conference*, 2008.

[32] A. Klesh, P. Kabamba, and A. Girard. Optimal path planning for uncertain exploration. In *Proceedings of American Control Conference*, 2009.

[33] A.T. Klesh, B. Hyun, T.L. Huntsberger, G.M. Woodward, P.T. Kabamba, and A.R. Girard. Tactical area search with strapped-down anisotropic sensors. Control Group Report CGR 10-05, University of Michigan, Ann Arbor, MI, 2010.

[34] I. I. Hussein. Kalman filtering with optimal sensor motion planning. In *Proceedings of the American Control Conference*, 2008.

[35] V. A. Sujan and M. A. Meggiolaro. On the visual exploration of unknown environments using information theory based metrics to determine the next best view. *Mobile Robots: New Research*, 2006.

[36] R. Sim. To boldly go: Bayesian exploration for mobile robots, 2000.

[37] T. H. Chung, V. Gupta, J. W. Burdick, and R. M. Murray. On a decentralized active sensing strategy using mobile sensor platforms in a network. In *IEEE Conf. on Decision and Control*, pages 1914–1919, 2004.

[38] S. Martínez and F. Bullo. Optimal sensor placement and motion coordination for target tracking. *Automatica*, 42(4):661–668, 2006.

[39] C. Kreucher, K. Kastella, and A. O. Hero. Sensor management using an active sensing approach. *Signal Processing*, 85:608–624, 2005.

[40] R. Fierro, S. Ferrari, and C. Cai. An information-driven framework for motion planning in robotic sensor networks: complexity and experiments. In *Proceedings of the 47th IEEE Conference on Decision and Control, Cancun, Mexico*, 2008.

[41] P. Salaris, D. Fontanelli, L. Pallottino, and A. Bicchi. Shortest paths for a robot with nonholonomic and field-of-view constraints. *Transactions on Robotics*, 26(2):269–281, 2010.

[42] H. Chitsaz, S. Lavalle, D.J. Balkcom, and M.T. Mason. Minimum wheel-rotation paths for differential-drive mobile robots. *International Journal of Robotics*, 28(1):66–80, 2009.

[43] B. Grocholsky, H.F. Durrant-Whyte, and P. Gibbens. An information theoretic approach to decentralized control of multiple autonomous flight vehicles. In *Sensor Fusion and Decentralized Control in Robotic Systems III*, pages 348–359, 2000.

[44] A. Singh, A. Krause, C. Guestrin, W. Kaiser, and M. Batalin. Efficient planning of informative paths for multiple robots. *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2204–2211, 2007.

[45] A.J. Sinclair, R.J. Prazenica, and D.E. Jeffcoat. Optimal and feedback path planning for cooperative attack. *Journal of Guidance, Control, and Dynamics*, 31(6):1708–1715, 2008.

[46] Jinwhan Kim. *Dual control approach for automatic docking using monocular vision.* PhD thesis, Stanford University, Stanford, CA, USA, 2007.

[47] J. Ousingsawat and M.R. Campbell. Optimal cooperative reconnaissance using multiple vehicles. *Journal of Guidance, Control, and Dynamics*, 30(1), 2007.

[48] D. Georgiev, P.T. Kabamba, and D.M. Tilbury. Distributed risk minimization in teams with incomplete information. In *Risk Symposium 2007: Risk Analysis for Homeland Security and Defense Theory and Application, Santa Fe, NM*, 2007.

[49] A. Menezes and P.T. Kabamba. Information requirement for self-reproducing systems in lunar robotic colonies. In *Proceedings of 57th International Astronautical Congress, Valencia, Spain*, 2006. Paper IAC-06-A5-P.04.

[50] Robin Smith. *Logic, In J. Barnes (ed) The Cambridge companion to Aristotle.* Cambridge: Cambridge University Press, 1995.

[51] Pierre T. Kabamba. A primer on logic - analysis, refutation and proof of formal arguments. 2010.

[52] Alan M. Turing. On computable numbers, with an application to the enstcheidungsproblem. In *Proceedings of the London Mathematical Society*, volume 42, pages 230–65, 1936.

[53] Alan M. Turing. On computable numbers, with an application to the enstcheidungsproblem: A correction. In *Proceedings of the London Mathematical Society*, volume 43, pages 544–6, 1937.

[54] D.J. Hand. Intelligent data analysis: issues and opportunities. *Intelligent Data Analysis*, 2:67–79, 1998.

[55] S.K. Halgamuge and L. Wang. *Classification and clustering for knowledge discovery (Studies in computational intelligence).* Springer, 2010.

[56] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[57] James C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer, 1edn, 1981.

[58] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 1943.

[59] Frank Rosenblatt. Perceptron–a perceiving and recognizing automaton. Technical report, Cornell Aeronautical Laboratory, 1957. Report 85-460-1.

[60] B. Widrow and M.A. Lehr. Perceptrons, adalines, and backpropagation. In *in Handbook of Brain Theory and Neural Networks, M.A. Arbib, ed.*, pages 719–724. MIT Press, 1995.

[61] Bernard Widrow. Adaline: Smarter than sweet. *Stanford Today*, 1963.

[62] B. Widrow and M.A. Lehr. 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, Sep 1990.

[63] P.J. Werbos. *Beyond regression: new tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University, 1974.

[64] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20(3-4):121–136, 1975.

[65] Bradley R. Smith. *Neural network enhancement of closed-loop controllers for ill-modeled systems with unknown nonlinearities*. PhD thesis, Virginia Polytechnic Institute and State University, 1997.

[66] M.H. Kutner, C.J. Nachtsheim, J. Neter, and W. Li. *Applied linear statistical models*. McGraw-Hill Irwin, 5th edition, 2005.

[67] S.S. Gupta and L.-Y. Leu. On a classification problem: ranking and selection approach. Technical Report #89-27C, Department of Statistics, Purdue University, 1989.

[68] R.G. Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.*, pages 35–46, 1991.

[69] C.-I Chang, Y. Du, J. Wang, S.-M. Guo, and P.D. Thouin. Survey and comparative analysis of entropy and relative entropy thresholding techniques. *IEEE Proc. Visual Image Signal Process*, 153(6), Dec 2006.

[70] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[71] Y.-J. Wu, M.D. Alston, and P.M. Chau. Dynamic adaptation of quantization thresholds for soft-decision viterbi decoding with reinforcement learning neural network. *Journal of VLSI Signal Processing*, 6:77–84, 1993.

[72] P. Reynaud-Bouret and V. Rivoirard. Near optimal thresholding estimation of a poisson intensity on the real line. *Electronic Journal of Statistics*, 4:172–238, 2010.

[73] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), Jan 2000.

[74] C.F. Hester and D. Casasent. Multivariant technique for multiclass pattern recognition. *Applied Optics*, 19(11), Jun. 1980.

[75] M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Elsevier Signal Processing*, 2003.

[76] N. Cebron and M.R. Berthold. Adaptive prototype-based fuzzy classification. *Fuzzy Sets and Systems*, 159:2806–2818, 2008.

[77] N. Cebron and M.R. Berthold. Adaptive active classification of cell assay images. *PKDD*, pages 79–90, 2006.

[78] N. Cebron and M.R. Berthold. Active learning for object classification: from exploration to exploitation. *Data Min. Knowl. Discov.*, 18(2):283–299, 2009.

[79] C. Yu, T.G. Smith, S. Hidaka, M. Scheutz, and L.B. Smith. A data-driven paradigm to understand multimodal communication in human-human and human-robot interaction. In *Advances in Intelligent Data Analysis IX, 9th International Symposium, IDA 2010, Tucson, AZ, USA*, 2010.

[80] Gerd Gigerenzer. *Gut Feelings - The intelligence of the unconscious*. Penguin Books, 2007.

[81] R.W. Holsapple, P.R. Chandler, J.J. Baker, A.R. Girard, and M. Pachter. Autonomous decision making with uncertainty for an urban intelligence, surveillance and reconnaissance (isr) scenario. In *Proceedings of AIAA Guidance, Navigation and Control Conference and Exhibit, Honolulu, HI, USA*, 2008.

[82] A.R. Girard, S. Dharba, M. Pachter, and P.R. Chandler. Stochastic dynamic programming for uncertainty handling in uav operations. In *Proceedings of the American Control Conference, New York City, NY, USA*, 2007.

[83] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3), May/June 1992.

[84] Y.S. Huang and C.Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:90–94, 1995.

[85] C.Y. Suen, C. Nadal, T.A. Mai, R. Legault, and L. Lam. Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts. In *Proc. Int. Workshop on Frontiers in Handwriting Recognition, Montreal, Canada*, 1990.

[86] C. Nadal, R. Legault, and C.Y. Suen. Complementary algorithms for the recognition of totally unconstrained handwritten numerals. In *Proc. 10th Int. Conf. Pattern Recog.*, volume A, pages 434–449, 1990.

[87] T.K. Ho, J.J. Hull, and S.N. Srihari. Combination of structural classifiers. In *Proc. IAPR Workshop Syntactic and Structural Pattern Recog.*, pages 123–137, 1990.

[88] T.K. Ho, J.J. Hull, and S.N. Srihari. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:66–75, 1994.

[89] Ludmila I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[90] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

[91] Robert A. Jacobs. Methods for combining experts' probability assessments. *Neural Computation*, 7(5):867–888, 2008.

[92] R. Radner. Team decision problems. *The Annals of Mathematical Statistics*, 33(3):857–881, 1962.

[93] J. Marschak. Elements for a theory of teams. *Management Science*, 1(2):127–137, 1955.

[94] Y.-C. Ho and K.-C. Chu. Team decision theory and information structures in optimal control problem - part 1. *IEEE Transactions on Automatic Control*, 17(1), Feb 1972.

[95] Y.-C. Ho. Team decision theory and information structures. *Proceedings of the IEEE*, 68(6), Jun 1980.

[96] D. Georgiev, P.T. Kabamba, and D.M. Tilbury. A new model for team optimization: the effects of uncertainty on interaction. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: System and Humans*, 38(6), Nov 2008.

[97] R. Murphey. An introduction to collective and cooperative systems. In *Cooperative Control and Optimization*, pages 171–197. Kluwer Academic Publishers, 2002.

[98] S.H. Dandach, R. Carli, and F. Bullo. Accuracy and decision time for decentralized implementations of the sequential probability ratio test. In *Proceedings of the American Control Conference, Baltimore, MD, USA*, 2010.

[99] A. Pete, K.R. Pattipati, and D.L. Kleinman. Methods for fusion of individual decisions. In *Proceedings of the American Control Conference*, 1991.

[100] N.H. Machworth. The breakdown of vigilance during prolonged visual search. *Quarterly Jounal of Experimental Psychology*, 1:6–21, 1948.

[101] W. Edwards. Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, 1965.

[102] M. Stone. Models for choice-reaction time. *Psychometrika*, 25(3), Sep 1960.

[103] P.M. Fitts. Cognitive aspects of information processing: 3. set for speed versus accuracy. 71(6), 1966.

[104] R.W. Pew. The speed-accuracy operating characteristics. *Acta Psychologica*, 30, 1969.

[105] J.P. Veverka and M.E. Campbell. Operator decision modeling for intelligence, surveillance and reconnaissance type missions. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 754–759, 2005.

[106] Y. Boussemart, J. Las Fargeas, M.L. Cummings, and N. Roy. Comparing learning technique for hidden markov models of human supervisory control behavior. In *Proceedings of AIAA Infotech@Aerospace'09 Conference, Seattle, WA*, 2009.

[107] C.E. Nehme and M.L. Cummings. An analysis of heterogeneity in futuristic unmanned vehicle systems, report (hal2007-07). Technical report, MIT Humans and Automation Laboratory, Cambridge, MA., 2007.

[108] M. Seck, C. Frydman, N. Giambiasi, T. Ören, and L. Yilmaz. Use of a dynamic personality filter in discrete event simulation of human behavior under stress and fatigue. In *Proc. 1st International Conference on Augmented Cognition*, 2005.

[109] M. Seck, N. Giambiasi, C. Frydman, and L. Baâit. Devs for human behavior modelling in cgfs. *Journal of Defense Modeling and Simulation*, 4(3), Jul 2007.

[110] R.M. Yerkes and J.D. Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18:459–482, 1908.

[111] L. Kuchinke. *Implicit and explicit recognition of emotionally valenced words*. PhD thesis, Freie Universitat, Berlin, 2007.

[112] T.B. Sheridan. Dynamic decisions and work load in multitask supervisory control. *IEEE Transaction on Systems, Man, and Cybernetics*, 10(5), May 1980.

[113] C.E. Nehme, B. Mekdeci, J.W. Crandall, and M.L. Cummings. The impact of heterogeneity on operator performance in future unmanned vehicle systems. *The International Command and Control Journal*, 2(2), 2008.

[114] M.L. Cummings and P.J. Mitchell. Predicting controller capacity in remote supervision of multiple unmanned vehicles. *IEEE Systems, Man, and Cybernetics, Part A Systems and Humans*, 38(2):451–460, 2008.

[115] M.L. Cummings, S. Bruni, S. Mercier, and P.J. Mitchell. Automation architecture for single operator, multiple uav command and control. *The International Command and Control Journal*, 1(2), 2007.

[116] S.D. Scott, F. Sasangohar, and M.L. Cummings. Investigating supervisory-level activity awareness displays for command and control operations. In *Proc. of HSIS 2009: Human Systems Integration Symposium, Annapolis, MD*, Mar 2009.

[117] A.E. Sklar and N.B. Sarter. Good vibrations: tactile feedback in support of attention allocation and human-automation coordination in event-driven domain. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41(4):543–552, 1999.

[118] B. Donmez, M.L. Cummings, and H.D. Graham. Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics*, 51(5):718–729, 2009.

[119] B. Hyun, C. Park, W. Wang, and A. Girard. Discrete event modeling of heterogeneous human operator team in classification task. In *Proceedings of the American Control Conference*, 2010.

[120] K. Savla, T. Temple, and E. Frazzoli. Human-in-the-loop vehicle routing policies for dynamic environments. In *Proceedings of IEEE Conference on Decision and Control, Cancun, Mexico*, pages 1145–1150, Dec 2008.

[121] K. Savla and E. Frazzoli. A dynamical queue approach to intelligent task management for human operators. *Proceedings of the IEEE*, 2010. Submitted.

[122] L.F. Bertuccelli, N.W.M. Beckers, and M.L. Cummings. Developing operator models for uav search scheduling. In *Proceedings of AIAA Guidance, Navigation, and Control Exhibit, Toronto, Canada*, 2010.

[123] L.L. Scharf. *Statistical signal processing (detection, estimation, and time series analysis)*. Addison-Wesley, 1991.

[124] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, 2005.

[125] D.S. Bernstein and P. Tsiotras. A course in classical optimal control, 2009.

[126] Maurice R. Fréchet. Sur la loi de répartion de certaines grandeurs géographiques. *Journal de la societé de Statistique du Paris*, 82:114–122, 1941.

[127] René Gâteuax. Sur les fonctionnelles continues et les fonctionnelles analytiques. *Comptes rendus de l'academie des sciences (Paris)*, 157:325–327, 1913.

[128] René Gâteuax. Fonctions d'une infinité de variables indépendantes. *Bulletin de la Société Mathmatique de France*, 47:70–96, 1919.

[129] Pierre T. Kabamba. Guidance, navigation and avionics. Lecture notes, 2009.

[130] D. Bertsimas and J. Tsitsiklis. Simulated annealing. *Statistical Science*, 8(1):10–15, 1993.

[131] John H. Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to Biology, Control, and Artificial Intelligence*. A Bradford Book, 1992.

[132] Donald Gross. *Fundamentals of Queueing theory*. Wiley, 2009.