A Method of Classification for Multisource Data in Remote Sensing Based on Interval-Valued Probabilities

Hakil Kim
Philip H. Swain

TR-EE 90-48
July 1990

Laboratory for Applications of Remote Sensing
and

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

A METHOD OF CLASSIFICATION FOR MULTISOURCE DATA IN REMOTE SENSING BASED ON INTERVAL-VALUED PROBABILITIES

Hakil Kim and Philip H. Swain

TR-EE 90-48

July 1990

Laboratory for Applications of Remote Sensing

and

School of Electrical Engineering

Purdue University

West Lafayette, Indiana 47907

This work was supported by NASA Grant No. NAGW-925 "Earth Observation Research - Using Multistage EOS-like Data" (Principal Investigators: David A. Landgrebe and Chris Johannsen).

ACKNOWLEDGMENTS

The financial support of the National Aeronautics and Space Administration (NASA) through Grant No. NAGW-925 is gratefully acknowledged.

The Anderson River SAR/MSS data set was acquired, preprocessed, and loaned to us by the Canada Centre for Remote Sensing, Department of Energy Mines, and Resources, of the Government of Canada.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

ABSTRACT

The importance of utilizing multisource data in ground-cover classification lies in the fact that improvements in classification accuracy can be achieved at the expense of additional independent features provided by separate sensors. However, it should be recognized that information and knowledge from most available data sources in the real world are neither certain nor complete. We refer to such a body of uncertain, incomplete, and sometimes inconsistent information as "evidential information."

The objective of this research is to develop a mathematical framework within which various applications can be made with multisource data in remote sensing and geographic information systems. The methodology described in this report has evolved from "evidential reasoning," where each data source is considered as providing a body of evidence with a certain degree of belief. The degrees of belief based on the body of evidence are represented by "interval-valued (IV) probabilities" rather than by conventional point-valued probabilities so that uncertainty can be embedded in the measures.

There are three fundamental problems in the multisource data analysis based on IV probabilities: (1) how to represent bodies of evidence by IV probabilities, (2) how to combine IV probabilities to give an overall assessment of the combined body of evidence, and (3) how to make a decision when the statistical evidence is given by IV probabilities.

This report first introduces an axiomatic approach to IV probabilities, where the IV probability is defined by a pair of set-theoretic functions which satisfy some pre-specified axioms. On the basis of this approach the report focuses on representation of statistical evidence by IV probabilities and combination of multiple bodies of evidence.

Although IV probabilities provide an innovative means for the representation and combination of evidential information, they make the decision process rather complicated. It entails more intelligent strategies for

making decisions. This report also focuses on the development of decision rules over IV probabilities from the viewpoint of statistical pattern recognition.

The proposed method, so called "evidential reasoning" method, is applied to the ground-cover classification of a multisource data set consisting of Multispectral Scanner (MSS) data, Synthetic Aperture Radar (SAR) data, and digital terrain data such as elevation, slope, and aspect. By treating the data sources separately, the method is able to capture both parametric and nonparametric information and to combine them.

Then the method is applied to two separate cases of classifying multiband data obtained by a single sensor. In each case, a set of multiple sources is obtained by dividing the dimensionally huge data into smaller and more manageable pieces based on the global statistical correlation information. By a Divide-and-Combine process, the method is able to utilize more features than the conventional Maximum Likelihood method.

CHAPTER 1

INTRODUCTION

1.1. Background

Since the developments of the digital computer and sensor systems made it possible to apply the quantitative approach to remote sensing in 1960s, information concerning the surface of the Earth and its environment has been largely extracted from the multispectral data obtained by a single sensor.

Within the last decade, as remote sensing and other data acquisition technologies have advanced, there has been a trend towards exploiting remotely sensed multispectral data in conjunction with related data from other sources for the purpose of extracting higher level information from multi-attribute data bases. For instance, the topographic information obtained from digital terrain data has been successfully used together with remotely sensed data in land cover analysis [Fleming et al. (1979), Franklin et al. (1986), Jones et al. (1988), Strahler et al. (1978)]. More recently, many researchers in the geographic information processing community have started reconsidering the possibility of utilizing remotely sensed data within geographic information systems (GIS) [Healey et al. (1988), Quarmby et al. (1988)]. Figure 1.1 depicts a typical multi-attribute database in remote sensing and GIS. In general, the information obtained from multiple sources is robust and more reliable than that from a single source. Furthermore, it may resolve ambiguities which might arise from single source analysis.

To a large extent, the methods which have been used for the analysis of multisource data have been *ad hoc* or often based on qualitative interpretation techniques, drawing heavily on the expertise and intuition of application scientists. Whereas techniques for collecting and storing data from multiple sources (e.g., multispectral scanner, side-looking radar, digital terrain model,

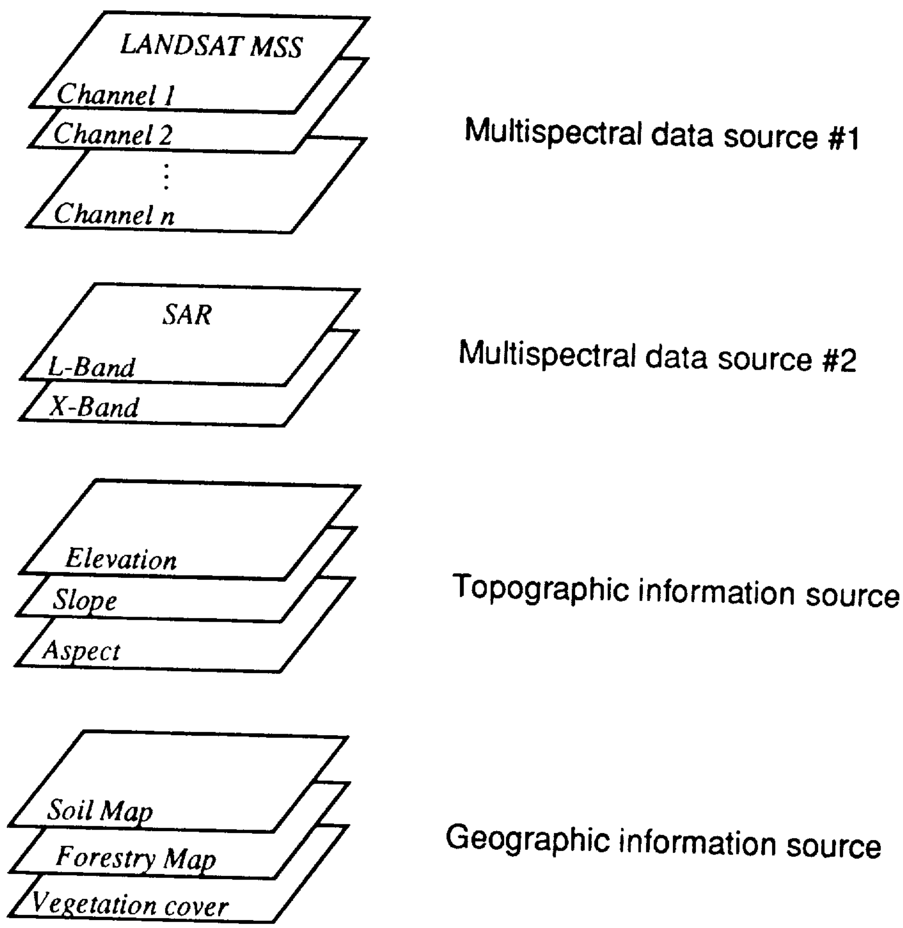


Figure 1.1 A Multi-Attribute Database in Remote Sensing and GIS.

etc.) have evolved rapidly, techniques for extracting and analyzing information from such complex data bases are still in the beginning stage. With the advancement in designing sensor systems and the increasing availability of ancillary data, interest in extracting the great wealth of higher level information contained in geographic and remote sensing contexts has led to extensive demand for computer-based, automated (or semi-automated) methods for the analysis of multisource data. Their development will be hastened more and more by proliferation of various and sophisticated remote sensing platforms and sensors in the next decades.

Unlike the situation in which we are dealing with purely spectral data from a single sensor, there are some conceivable problems in devising means for multisensor and multisource data analysis. Firstly, there is a difficulty in describing the disparate range of data types which have different units of measurement. The types of data to be combined cannot be assumed to be commensurable. For example, multispectral data represent the energy emanating from the scene of interest in different wavelengths while elevation data represent the altitude of the scene. Moreover, map-based ancillary data such as a soil map may even be nominal in nature. The situation becomes more complicated when the multi-attribute data bases include geometric characteristics such as lines, shapes, or sizes.

Secondly, since spatial variation of the attribute in a geographic context, such as vegetation cover, soil type, or slope aspect, has an effect on the spectral responses obtained from remote sensors, there are possibly significant but unknown interactions among multiple data sources. For example, in the visible/infrared spectral range the reflected energy measured by a sensor depends on properties such as the pigmentation, moisture content and cellular structure of vegetation, the mineral and moisture contents of soils, and the level of sedimentation of water. However, when there is insufficient knowledge concerning the interactions among data sources, the observations obtained from the data sources have been treated as independent variables. Such an independence assumption should be adopted with caution in the case of a statistical multisource data analysis because the data sources which seem to be apparently uninteracting are unlikely to be statistically independent.

Thirdly, while it is often reasonable to adopt the multivariate Gaussian distribution to model the probability function of multispectral data alone, this parametric model is not generally applicable to accommodate geographic or topographic data combined with multispectral data when the representation of their joint probability function is unknown.

Finally, there is an important factor which must be considered in combining multiple sources. Since various data sources are in general not equally reliable, the data sources usually provide a wide range of degrees of support for an observation, sometimes even in an inconsistent manner. Such information regarding the relative reliabilities of the sources should be included in the multisource data analysis.

These problems have been the motivation for the development of the techniques by which inferences can be drawn systematically from complex data bases composed of disparate, unequally reliable sources, regardless of their data types and interactions with the other sources.

1.2. Related Works

During the last decade, there have been a number of different approaches to the analysis of multisource data in remote sensing and geographic information systems. The approaches listed in this section are not exhaustive of the related works but are representative.

First of all, the "stacked vector" approach is the most straightforward method in which all data sources are considered simultaneously by organizing the respective measurements into a single vector. The resulting compound vectors are treated as data from a single source. Although this approach has been successfully applied to combined multispectral data and terrain data [Hoffer et al. (1975)], its use is limited to the situation where the sources are similar and their interactions are easily modeled.

The "layered" approach employed by Fleming et al. (1979) is more general in the sense that it can deal with multiple sources of diverse data types by treating them separately. This approach has been used for mapping forest cover types based on multispectral data and topographic data. Its idea is to classify major cover types based on the multispectral data, and then further subdivide the cover types to the species level based on the remaining data. Hutchinson (1982) has developed a similar approach, so called "ambiguity reduction" method, whose basic strategy is to stratify the data based on one (or more) of data sources, assess the results, and resort to the other sources to resolve the remaining ambiguities. A major disadvantage of these two approaches is that different groupings or orderings of the sources may produce different results. Furthermore, their mathematical schemes cannot incorporate the reliabilities and interactions of the sources into the classification process.

Swain et al. (1985) proposed an approach which can handle an arbitrary number of independent data sources. In their mathematical framework, the global membership function is derived from Bayes' formula by applying two different statistical independence assumptions. Due to the commutative property of the global membership function, different orderings of the sources in combination do not have an effect on final results. This method has been extended by Lee et al. (1987) and Benediktsson et al. (1989a) so that the relative quality of the sources can be accounted for in the global member-ship function.

Although their procedures in combining information from multiple data sources are different, the numerical representations of information in the above approaches are commonly based on the Bayesian inference, where posterior probabilities are defined by the multiplication of prior probabilities and observational probabilities. It is very important to recognize that in dealing with multispectral data combined with other forms of geographic data, the methods employed must be able to cope with uncertainties which arise both from intrinsic randomness of data and from ambiguities in modeling and combining disparate sources.

Recently, learning procedures based on neural networks have been applied to the classification of remotely sensed multisource data [Benediktsson et al. (1989b)]. Since it is nonparametric in nature, the neural network approach is most useful when the distribution functions of data are not known. However, this approach usually involves a large amount of computational complexity in training due to an iterative procedure.

Meanwhile, in the artificial intelligence and knowledge engineering community, there have been a number of attempts to build plausible models for automated reasoning with multiple information sources [Cohen (1985), McDermott and Doyle (1979), Shafer (1976a), Zadeh (1965)]. Such attempts have been embodied as "inference techniques under uncertainty" [Duda et al. (1976), Dubois and Prade (1980), Ginsberg (1984), Lowrance and Garvey (1982)] and used in various areas of science and engineering [Blonda et al. (1989), Duda et al. (1979), Garvey (1987), Garvey et al. (1981), Kim et al. (1986), Moon (1989), Shortliffe (1976)]. Applications to multisource geographic and remote sensing data have been rudimentary at best.

1.3. Statement of Problem

The importance of utilizing multisource data in ground-cover classification lies in the fact that it is generally correct to assume that improvements in terms of classification accuracy can be achieved at the expense of additional independent features provided by separate sensors or other forms of data sources. However, it should be recognized that information and knowledge from most available sources of data in the real world are neither certain nor complete. We refer to such a body of uncertain, incomplete, and sometimes inconsistent information as "evidential information."

In order for any methodology for multisource data classification to be implemented as a quantitative, computer-based technique, the methodology must be able to: (1) represent the partial information provided by the individual sensors as numerical measures, and (2) combine the measures by a combination rule to produce the overall assessment of the total evidence.

Consider the problem of classifying a pixel $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T$ to one of n classes denoted by ω_j for $j=1, \ldots, n$, where $\mathbf{x}_i$ ($i = 1, \ldots, m$) is the feature obtained from the i^{th} source denoted by S_i and the superscript T denotes the vector transposition. Suppose each data source S_i supports A denoting the event of $\mathbf{X}$ belonging to a certain class ω with a degree of belief $B(A|\mathbf{x}_i) = b_i$. Throughout the report, the term "degree of belief" or "belief measure" will be used for any kind of numerical measure representing one's belief states regarding the events. Then, the first problem above is equivalent to the construction of belief measures based on evidential information provided by each data source.

As we mentioned earlier, evidential information is characteristically uncertain and incomplete. Therefore, the classical Boolean logic is not adequate for representing evidence because it cannot have intermediate states between "True" and "False." In other words, the Boolean expressions never capture any notion of the relative strength of partial beliefs. Bayesian probabilities have been frequently used to represent partial beliefs. Yet this is possible only when there is a sufficient amount of data to estimate the statistical parameters of an assumed probability model. Further, there is no appropriate way for representing "total ignorance" in a Bayesian framework because the Bayesian probabilities should be "additive", that is,

$$P(A) + P(\bar{A}) = 1 \tag{1.3.1}$$

where $\bar{A}$ is the complementary event of A. To illustrate the consequence of this requirement, suppose there is no evidence available either for or against the occurrence of two exclusive and exhaustive events. In the Bayesian framework, both events are equally assigned a probability of $\frac{1}{2}$, which seems quite different from specifying that nothing is known regarding the occurrence of the events.

Once the belief measures based on individual sources are given, the next problem is: whether we can find a combined degree of belief $B(A \mid \mathbf{x}_1, \cdots, \mathbf{x}_m)$, or equivalently, whether we can build a numerical formula $\mathcal{F}$ such that

$$B(A \mid \mathbf{x}_1, \cdots, \mathbf{x}_m) = \mathcal{F}(b_1, \ldots, b_m) \tag{1.3.2}$$

If the data sources are not believed to be equally reliable, the relative reliabilities of the sources must be considered in computing the combined degree of belief, i.e.,

$$B(A \mid \mathbf{x}_1, \cdots, \mathbf{x}_m) = \mathcal{F}(b_1, \ldots, b_m ; a_1, \ldots, a_m) \tag{1.3.3}$$

where a_i's denote the relative reliabilities of the sources.

When the numerical representation of belief and the formulation of combining function depend on the expertise and intuition of human analysts, the solutions to the above problems are said to be *ad hoc*.

1.4. Objective of the Research

The objective of the research is to develop a mathematical framework for dealing effectively with multisource data in remote sensing and GIS and to provide a preliminary demonstration of its value. The methodology described in this report has evolved from "evidential reasoning," where each data source is considered as providing a body of evidence concerning propositions with certain degrees of belief. The degrees of belief based on the body of evidence are represented by "interval-valued (IV) probabilities" rather than by conventional additive probabilities so that uncertainty can be embedded in the measures.

There are three fundamental problems in the multisource data analysis based on IV probabilities: (1) how to represent bodies of evidence by IV probabilities, (2) how to combine IV probabilities to give an overall assessment of the combined body of evidence, and (3) how to make decisions based on IV probabilities.

There have been various approaches to IV probabilities in the areas of philosophy of science and statistics. The primary focus of this report is on the unification of various concepts of IV probabilities so that IV probabilities can be readily accessible to representation and combination of multiple bodies of evidence without any conceptual ambiguities. This report pursues an axiomatic approach to IV probabilities, where IV probabilities are defined axiomatically based on the least of the common properties which are consistently required in the various approaches. Secondarily, this report focuses on formal methods of representing statistical evidence by IV probabilities, first based on acceptable models in robust estimation of probabilities, and then using the likelihood function of observed data.

We do not propose any brand-new rule for combining multiple evidence. Instead, some existing rules are investigated in terms of their inferencing mechanisms when they are expressed as set-theoretic functions. Although IV probabilities provide an innovative means for the representation of evidential information, they make the decision process rather complicated. We need more intelligent strategies for making decisions. This report addresses the development of decision rules over IV probabilities as the counterparts of conventional decision rules in statistics.

In this report, the problem of multisource data analysis in remote sensing and GIS is viewed as an application area for the use of artificial intelligence and knowledge engineering techniques.

1.5. Thesis Organization

This report is made up of seven chapters. In this introductory chapter, the problems in the analysis of multisource data have been addressed, and the objective of the research has been stated. In the following chapter, after reviewing various approaches to IV probabilities, an axiomatic approach to IV probabilities is introduced. Chapter 3 describes how belief functions for statistical evidence can be constructed in the form of IV probabilities. Chapter 4 examines subjective Bayesian rules and Dempster's rule for combining evidence in the sense of satisfying some desirable properties which agree with human intuition. Particularly, attention is paid to the inference mechanisms of Dempster's rule. In Chapter 5, decision rules over IV probabilities are defined on the basis of well-known decision principles in statistics, such as the Likelihood Principle and the Minimax Principle. For the purpose of general assessments of its ability in capturing and utilizing information in multisource data, the approach is applied to the problems of ground-cover classification based on multispectral data in conjunction with other sources of data in remote sensing. The experimental results are presented in Chapter 6 and compared to the performance of a traditional maximum posterior probability classification method. Finally, Chapter 7 concludes the report by summarizing and suggesting directions for further research.

CHAPTER 2

APPROACHES TO INTERVAL-VALUED PROBABILITIES

2.1. Introduction

Interval-valued probabilities are, in general, a more adequate scheme than point-valued probabilities to express one's state of knowledge in the sense of handling uncertain, incomplete evidential information. IV probabilities can be thought of as a generalization of conventional additive probabilities, with the lower and upper extremes of the interval corresponding to an event being bounds for the unknown actual probability of the event. The endpoints of IV probabilities are called the "upper probability" and the "lower probability."

There have been various works introducing the concepts of IV probabilities in the areas of philosophy of science and statistics. For example, Koopman (1940) derives the upper and lower probabilities based on the intuitively evident laws of consistency governing all comparisons in partial ordering of non-numerical probabilities. Smith (1961) proposes a system of IV probabilities by considering the strength of one's belief in betting odds as an interval. Good (1962) considers the upper and lower probabilities of an event by analogy with the outer and inner measures of a non-measurable set. Dempster (1967) formulates a system of upper and lower probabilities induced by a set-theoretic multivalued mapping. Suppes and Zanotti (1977) show how a random relation generates upper and lower probabilities in the set-theoretic image space. And Walley and Fine (1982) present a frequentist account of IV probabilities based on a finite event algebra.

Among the above approaches, only Dempster's and Walley and Fine's models are useful for parametric statistical inference. Dempster's work and Shafer's mathematical theory of evidence [Shafer (1976a)], together called "Dempster-Shafer theory," have shown their usefulness in various evidential reasoning systems [Garvey (1987), Garvey et al. (1981), Zhang and Chen (1987)]. Walley and Fine's approach provides the fundamental concepts of a frequentist theory of statistics for IV probabilities. Their results indicate that an objectivist or frequency-oriented view of probability does not necessitate an additive probability concept, and that IV probability models can represent a type of indeterminacy not captured by additive probabilities. In the following two sections, both approaches will be briefly reviewed.

Although the mathematical rationales behind the approaches listed above are different, there are some common properties of IV probabilities which are consistently required. This chapter introduces an axiomatic approach to IV probabilities, where IV probabilities are defined by a pair of set-theoretic functions satisfying the common properties, so that conceptual ambiguities can be avoided.

2.2. Dempster-Shafer Theory

In his 1960's works, Dempster (1967, 1968) proposed a generalized scheme of statistical inference about a parameter space by introducing upper and lower probabilities induced by a multivalued mapping. His scheme has been further developed and recast as a "mathematical theory of evidence" by Shafer. In this section, after briefly recalling the concepts of Dempster's upper and lower probabilities, we discuss the formal framework of Shafer's theory in the aspect of evidential reasoning.

Suppose we have a pair of spaces $\mathcal{X}$ and Ω denoting respectively a sample space and a finite parameter space. Let Γ be a multivalued mapping which assigns a subset $\Gamma_{\mathbf{x}} \subset \Omega$ to every $\mathbf{x} \in \mathcal{X}$ and let μ be a probability measure assigning probabilities to the members of the class Ψ of subsets of $\mathcal{X}$. Then, $(\mathcal{X}, \Psi, \mu)$ is a probability space, and this model corresponds to a random experiment where the outcome cannot be precisely observed but can only be located in a subset of all possible outcomes.

For any $A \subset \Omega$, define

$$A^* = \{ \mathbf{x} \in \mathcal{X} \mid \Gamma_{\mathbf{x}} \cap A \neq \varnothing \} \qquad (2.2.1)$$

and

$$A_* = \{ \mathbf{x} \in \mathcal{X} \mid \Gamma_{\mathbf{x}} \subset A, \Gamma_{\mathbf{x}} \neq \varnothing \} \tag{2.2.2}$$

A^* consists of those $\mathbf{x} \in \mathcal{X}$ which can possibly correspond under Γ to an $\omega \in \Omega$, while A_* consists of those $\mathbf{x} \in \mathcal{X}$ which must lead to an $\omega \in \Omega$. Then, the upper probability and the lower probability of A are defined respectively as :

$$P^*(A) = \frac{\mu(A^*)}{\mu(\Omega^*)} \tag{2.2.3}$$

$$P_*(A) = \frac{\mu(A_*)}{\mu(\Omega_*)} \tag{2.2.4}$$

where $\Omega^* = \Omega_*$ is the domain of Γ. Note that $P^*(A)$ and $P_*(A)$ are defined only if $\mu(\Omega^*) \neq 0$. Since A^* consists of those $\mathbf{x} \in \mathcal{X}$ which can possibly correspond under Γ to an $\omega \in A$, $\mu(A^*)$ may be regarded as the largest possible amount of probability which can be transferred to the outcomes $\omega \in A$ from the measure μ. Similarly, A_* consists of those $\mathbf{x} \in \mathcal{X}$ which must lead to an $\omega \in A$. So, $\mu(A_*)$ represents the minimal amount of probability which can be transferred to the outcomes $\omega \in A$. The denominator $\mu(\Omega^*) = \mu(\Omega_*)$ in eq. (2.2.3) and eq. (2.2.4) is a normalizing factor. The normalization is necessary in the case where there is any $\mathbf{x} \in \mathcal{X}$ which does not map into any subset of Ω. In this case, the subset $\{ \mathbf{x} \in \mathcal{X} \mid \Gamma_{\mathbf{x}} = \varnothing \}$ must be removed from $\mathcal{X}$, and the measure of the remaining set Ω^* should be renormalized to unity.

Dempster has assumed that the actual probability measure of A, P(A), lies in the interval $[P_*(A), P^*(A)]$ such that

$$P_*(A) \leq P(A) \leq P^*(A) \tag{2.2.5}$$

The degree of uncertainty concerning the true value of P(A) is represented by the width, $P^*(A) - P_*(A)$, of the interval.

In Shafer's theory, Ω is called the "frame of discernment" containing a finite number of exhaustive and mutually exclusive propositions. 2^Ω denotes the set of all possible subsets of Ω. His theory of evidence may begin by defining "basic probability assignment":

$$m : 2^{\Omega} \rightarrow [0, 1] \tag{2.2.6}$$

where m satisfies the following conditions,

(1) $m(\varnothing) = 0,$ (2.2.7)

(2) $\sum_{A \subseteq \Omega} m(A) = 1$ (2.2.8)

Given a basic probability assignment m over 2^{Ω}, Shafer's "belief function" $\mathcal{Bel}: 2^{\Omega} \rightarrow [0 , 1]$ is obtained as:

$$\mathcal{Bel}(A) = \sum_{B \subseteq A} m(B) \tag{2.2.9}$$

It satisfies the following conditions:

(1) $\mathcal{Bel}(\varnothing) = 0$ (2.2.10)

(2) $\mathcal{Bel}(\Omega) = 1$ (2.2.11)

(3) For every integer n and every collection $A_1, \ldots, A_n$ of subsets of Ω,

$$\mathcal{Bel}(A_1 \cup \ldots \cup A_n) \geq \sum_{i} \mathcal{Bel}(A_i) - \sum_{i<j} \mathcal{Bel}(A_i \cap A_j) + \ldots + (-1)^{n+1} \mathcal{Bel}(A_1 \cap \ldots \cap A_n) \tag{2.2.12}$$

The basic probability assignment which produces a given belief function is uniquely recovered from the belief function by the inverse formula of eq. (2.2.9) [see Shafer (1976a)]:

$$m(A) = \sum_{B \subset A} (-1)^{|A-B|} \mathcal{Bel}(B) \quad \text{for all } A \subset \Omega \tag{2.2.13}$$

where $|C|$ denotes the cardinality of a set C.

The basic probability number of a set $A \subset \Omega$, $m(A)$, may be understood as the exact measure of belief that the knowledge source has committed to A. A is called a "focal element" of the belief function $\mathcal{Bel}$ over Ω if $m(A) > 0$. The measure ascribed to the frame of discernment, $m(\Omega)$, represents the degree of ignorance, i.e., the portion of belief that could not be assigned to any smaller subset of Ω based on the evidence at hand. It may be committed to some subsets with the help of additional information. $\mathcal{Bel}(A)$ represents the measure of the total belief committed to A. In fact, eq. (2.2.9) reflects the basic intuition that a portion of belief committed to a proposition is also committed to any other proposition it implies.

While $\mathcal{Bel}(A)$ describes one's belief about A, it does not reveal to what extent one doubts A, i.e., to what extent one believes the negation of A, $\bar{A}$. Once $\mathcal{Bel}(\bar{A})$ is known, the upper probability of A is defined as:

$$\mathcal{Pl}(A) = 1 - \mathcal{Bel}(\bar{A}) \tag{2.2.14}$$

In the evidential reasoning based on the Shafer's theory, $\mathcal{Bel}(A)$ is called "degree of support" representing the extent to which a given body of evidence supports A, while $\mathcal{Pl}(A)$ is called "degree of plausibility" representing the extent to which the body of evidence fails to refute A.

2.3. A Frequentist Theory of Upper and Lower Probabilities

Walley and Fine (1982) give a limiting frequentist interpretation of P_* and P^* as "lim inf" and "lim sup" of relative frequencies in hypothetical unlinked repetitions of an experiment, which is a generalization of the usual limiting frequentist interpretation of additive probabilities. Their results provide the statistical basis whereby IV probability models of random experiments can be inferred from observations made on unlinked repetition. In this section briefly described is the link between relative frequencies and IV probabilities.

Let $\mathcal{B}$ be a Boolean algebra of subsets of Ω. Suppose that propensities of events $A \in \mathcal{B}$ in independent, identically distributed (iid) repetitions $\varepsilon_1, \ldots, \varepsilon_n$ are represented through the lower probability P_*. To provide a connection between frequency and propensity, P_* is inferred or estimated from relative frequency data. Let r_i denote the relative frequencies of all events in $\varepsilon_1, \ldots, \varepsilon_n$. More reliable information regarding the underlying marginal probability P_* can be obtained on the basis of the outcomes of the repeated experiments than the relative frequencies observed at any particular single experiment ε_i. Walley and

Fine propose an estimator

$$\underline{r}_n = \min \{ r_j(A) : k(n) \leq j \leq n \} \quad \text{for all } A \in 2^{\Omega} \tag{2.3.1}$$

where k is some function such that $k(n) \to \infty$ and $\frac{k(n)}{n} \to 0$ as $n \to \infty$ (e.g., $k(n) = [\sqrt{n}]$).

Although it is not "optimal" in any sense, the above minimum estimator makes use of the additional information concerning the past evolution of the sequence of relative frequencies. The estimator has asymptotic properties in a sequence of infinite trials, and parallels the Bernoulli's law of large numbers. There is no explicit description of $\overline{r}_n$ in terms of relative frequencies. However, the upper probability is given in terms of upper and lower "envelopes" which will be described in the next section.

2.4. Axiomatic Approach

A system of IV probability derived from the definitions and specifications of a particular mathematical or statistical concept may cause complications resulting from the need to satisfy underlying assumptions of the system. In the axiomatic approach, IV probabilities are formulated by defining the upper and lower probabilities of the interval as set-theoretic functions which satisfy some pre-specified axioms.

Definition 2.1. [Suppes (1974)] Let $\mathcal{B}$ be a Boolean algebra of subsets of Ω. The interval-valued probability $[\mathcal{L}, \mathcal{U}]$ over $\mathcal{B}$ is defined by the set-theoretic functions

$$\text{lower probability function} \quad \mathcal{L} : \mathcal{B} \to [0,1] \tag{2.4.1}$$

$$\text{upper probability function} \quad \mathcal{U} : \mathcal{B} \to [0,1] \tag{2.4.2}$$

satisfying the following conditions:

I $\mathcal{U}(A) \geq \mathcal{L}(A) \geq 0$ for all $A \in \mathcal{B}$ (2.4.3)

II $\mathcal{U}(\Omega) = \mathcal{L}(\Omega) = 1$ (2.4.4)

III For any $A, B \in \mathcal{B}$ and $A \cap B = \varnothing$,

$\mathcal{L}(A \cup B) \geq \mathcal{L}(A) + \mathcal{L}(B)$ (Super-additivity of $\mathcal{L}$) (2.4.5)

$\mathcal{U}(A \cup B) \geq \mathcal{U}(A) + \mathcal{U}(B)$ (Sub-additivity of $\mathcal{U}$) (2.4.6)

$\mathcal{L}(A \cup B) \leq \mathcal{L}(A) + \mathcal{U}(B) \leq \mathcal{U}(A \cup B)$ (Mixed-additivity of $\mathcal{L}$ and $\mathcal{U}$) (2.4.7)

These conditions are the least requirements on $\mathcal{L}$ and $\mathcal{U}$ for further development of the theory of IV probability. The following lemma sets forth some significant properties of IV probabilities as simple consequences of the above definition.

Lemma 2.1. For any $A, B \in \mathcal{B}$, the interval-valued probability $[\mathcal{L}, \mathcal{U}]$ has the following properties:

(i) $\mathcal{L}(A) + \mathcal{U}(\bar{A}) = 1$ (2.4.8)

(ii) $\mathcal{L}(\varnothing) = \mathcal{U}(\varnothing) = 0$ (2.4.9)

(iii) If $A \subset B$ then $\mathcal{L}(A) \leq \mathcal{L}(B)$ and $\mathcal{U}(A) \leq \mathcal{U}(B)$ (2.4.10)

(iv) $\mathcal{L}(A) + \mathcal{L}(B) \leq 1 + \mathcal{L}(A \cap B)$ (2.4.11)

(v) $\mathcal{U}(A) + \mathcal{U}(B) \geq 1 + \mathcal{U}(A \cap B)$ (2.4.12)

Proof. (i) follows immediately from eq. (2.4.4) and eq. (2.4.7). (ii) is obtained by eq. (2.4.4) and eq. (2.4.8). For (iii), if $A \subset B$ then by eq. (2.4.7)

$$\mathcal{U}(B) = \mathcal{U}(A \cup (B-A)) \geq \mathcal{U}(A) + \mathcal{L}(B-A)$$

and by eq. (2.4.5)

$$\mathcal{L}(B) = \mathcal{L}(A \cup (B-A)) \geq \mathcal{L}(A) + \mathcal{L}(B-A)$$

Since $\mathcal{L}(B-A) \geq 0$ from eq. (2.4.3),

$$\mathcal{U}(A) \leq \mathcal{U}(B) \quad \text{and} \quad \mathcal{L}(A) \leq \mathcal{L}(B)$$

For (iv),

$$\begin{aligned} \mathcal{L}(A) + \mathcal{L}(B) &\le 1 - \mathcal{U}(\bar{A}) + \mathcal{L}(\bar{A} \cap B) && \text{(By eq. (2.4.8) \& eq. (2.4.10))} \\ &= 1 - \mathcal{U}(\bar{A}) + 1 - \mathcal{U}(A \cap \bar{B}) && \text{(By eq. (2.4.8))} \\ &\le 2 - \mathcal{U}(\bar{A} \cup \bar{B}) && \text{(By eq. (2.4.6))} \\ &= 1 + \mathcal{L}(A \cap B) && \text{(By eq. (2.4.8))} \end{aligned}$$

Likewise, (v) can be proved. ■

The following definition given by Huber (1973) connects the upper and lower probabilities to the supremum and infimum of a class of probability measures. This connection becomes essential later in Section 3.2 where IV probabilities are constructed by some models in robust estimation of probability measures.

Definition 2.2. Let $\mathcal{M}$ be the set of all probability measures on a Boolean algebra $\mathcal{B}$ of all subsets of Ω and $\mathcal{P}$ an arbitrary non-empty subset of $\mathcal{M}$. $[\mathcal{L}, \mathcal{U}]$ is said to be "representable" by $\mathcal{P}$ if $\mathcal{L}$ and $\mathcal{U}$ can be defined as:

$$\mathcal{L}(A) = \inf \{ \pi(A) : \pi \in \mathcal{P} \} \tag{2.4.13}$$

and

$$\mathcal{U}(A) = \sup \{ \pi(A) : \pi \in \mathcal{P} \} \tag{2.4.14}$$

for all $A \in \mathcal{B}$. In this particular case $\mathcal{L}$ and $\mathcal{U}$ are called a “lower envelope” and an “upper envelope” respectively.

It has been proven by Huber and Strassen (1973) that if $[\mathcal{L}, \mathcal{U}]$ is an envelope, then it is an IV probability. The converse is not always true. The following example from Huber (1981) illustrates such a case. In fact, $[\mathcal{L}, \mathcal{U}]$ being an IV probability does not imply even the existence of the class $\mathcal{P}$ of probability measures.

Example 2.1. Let Ω have cardinality $|\Omega| = 4$, and assume that $\mathcal{L}(A)$ and

$\mathcal{U}(A)$ depend only on the cardinality of $A \subset \Omega$, according to the following table:

$\|A\|$	0	1	2	3	4
$\mathcal{L}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	1
$\mathcal{U}$	0	$\frac{1}{2}$	$\frac{1}{2}$	1	1

Then $[\mathcal{L}, \mathcal{U}]$ satisfies the IV probability's conditions in Definition 2.1, but there is only a single additive set function between $\mathcal{L}$ and $\mathcal{U}$, namely $P(A) = \frac{|A|}{4}$; hence $[\mathcal{L}, \mathcal{U}]$ is not representable.

The following definition and lemma result in interesting subclasses of IV probabilities by requiring relatively stronger constraints on $\mathcal{L}$ and $\mathcal{U}$:

Definition 2.3. [Choquet (1953)] The lower probability function $\mathcal{L}$ in Definition 2.1 is said to be "monotone of order *n* " or briefly "*n*-monotone", where n (≥ 2) is a positive integer, if for every collection $A_1, A_2, \ldots, A_n$ of subsets of Ω

$$\mathcal{L}(A_1 \cup \ldots \cup A_n) \geq \sum_i \mathcal{L}(A_i) - \sum_{i<j} \mathcal{L}(A_i \cap A_j) \; + \ldots + \; (-1)^{n+1} \mathcal{L}(A_1 \cap \ldots \cap A_n) \qquad (2.4.15)$$

The conjugate upper probability function $\mathcal{U}$ is said to be "alternating of order *n*" or "*n*-alternating" and satisfies

$$\mathcal{U}(A_1 \cup \ldots \cup A_n) \leq \sum_i \mathcal{U}(A_i) - \sum_{i<j} \mathcal{U}(A_i \cap A_j) \; + \ldots + \; (-1)^{n+1} \mathcal{U}(A_1 \cap \ldots \cap A_n) \qquad (2.4.16)$$

It is known that if $\mathcal{L}$ ($\mathcal{U}$) is monotone (alternating) of order n, then it is also monotone (alternating) of order k for any integer $2 \leq k \leq n$. In particular, when $k=2$, $\mathcal{L}$ and $\mathcal{U}$ have the following properties:

$$\mathcal{L}(A_1 \cup A_2) \geq \mathcal{L}(A_1) + \mathcal{L}(A_2) - \mathcal{L}(A_1 \cap A_2) \qquad \text{(2-monotone)} \qquad (2.4.17)$$

$$\mathcal{U}(A_1 \cup A_2) \leq \mathcal{U}(A_1) + \mathcal{U}(A_2) - \mathcal{U}(A_1 \cap A_2) \qquad \text{(2-alternating)} \qquad (2.4.18)$$

The following lemma shows that $[\mathcal{L}, \mathcal{U}]$ satisfying the above equations is an IV probability.

Lemma 2.2. If $\mathcal{L}$ and $\mathcal{U}$ are respectively 2-monotone and 2-alternating and satisfy the following conditions for all $A \in \mathcal{B}$:

(i) $\mathcal{U}(A) \geq \mathcal{L}(A) \geq 0$ (2.4.19)

(ii) $\mathcal{U}(\Omega) = \mathcal{L}(\Omega) = 1$ (2.4.20)

(iii) $\mathcal{L}(A) + \mathcal{U}(\bar{A}) = 1$ (2.4.21)

then $[\mathcal{L}, \mathcal{U}]$ is an IV probability. The converse is not necessarily true.

Proof. To prove this lemma, we only need to show that $\mathcal{L}$ and $\mathcal{U}$ are super-additive, sub-additive, and mixed-additive as in Definition 2.1. For any $A, B \in \beta$, if $A \cap B = \emptyset$, from eq. (2.4.17) "2-monotone" implies "super-additive", and from eq. (2.4.18) "2-alternating" implies "sub-additive." When $A \cap B = \emptyset$, $\bar{B} = (\overline{A \cup B}) \cup B$. Using eq. (2.4.5) and eq. (2.4.20),

$$\mathcal{L}(\bar{A}) = \mathcal{L}((\overline{A \cup B}) \cup B) \geq \mathcal{L}(\overline{A \cup B}) + \mathcal{L}(B) = 1 - \mathcal{U}(A \cup B) + \mathcal{L}(B)$$

$$\therefore\ \mathcal{U}(A \cup B) \geq \mathcal{U}(A) + \mathcal{L}(B)$$

Likewise,

$$\mathcal{U}(\bar{B}) = \mathcal{U}(A \cup (\overline{A \cup B})) \leq \mathcal{U}(A) + \mathcal{U}(\overline{A \cup B}) = \mathcal{U}(A) + 1 - \mathcal{L}(A \cup B)$$

$$\therefore\ \mathcal{L}(A \cup B) \leq \mathcal{U}(A) + \mathcal{L}(B)$$

Hence, $\mathcal{L}$ and $\mathcal{U}$ have mixed-additivity, and the above lemma is proved. ■

By comparing eq. (2.4.15) with eq. (2.2.11), Shafer's belief function $\mathcal{B}el$ is n-monotone. Consequently, $\mathcal{P}l$ is n-alternating. According to the above lemma, $\mathcal{B}el$ along with $\mathcal{P}l$ formulates a subclass of IV probabilities. We can summarize the implicative relationship among IV probabilities and its subclasses as follows:

$\mathcal{L}$ is n-monotone and $\mathcal{U}$ is n-alternating for $n > 2 \Rightarrow \mathcal{L}$ is 2-monotone and $\mathcal{U}$ is 2-alternating $\Rightarrow [\mathcal{L}, \mathcal{U}]$ is an envelope $\Rightarrow [\mathcal{L}, \mathcal{U}]$ is an IV probabilities.

In practical applications, 2-monotone and 2-alternating IV probabilities seem to be sufficiently general and mathematically amenable to develop an alternative statistical inferencing scheme to Bayesian inferencing.

2.5. Summary

In this chapter, we have discussed the axiomatic approach to IV probabilities whose mathematical framework is the theoretical basis of the contents treated in the rest of this report. The axiomatic IV probability was represented first by the pair of set functions and then by the supremum and infimum of a class of probability measures. Subclasses of IV probabilities were introduced.

IV probabilities as a generalization of additive probabilities give rise to some advantages such as representing a certain type of indeterminacy or uncertainty not captured by additive probabilities. The choice between deterministic, additive probability and IV probability models will depend on our background knowledge concerning the context of particular applications, and especially the amount and reliability of the information available to help in specifying the model.

In this chapter, the contribution of this research is in a unification of various concepts of IV probabilities so that IV probabilities can be readily accessible to representation and combination of multiple bodies of evidence. Lemmas 2.1 and 2.2 are originally formulated and proved in this report.

CHAPTER 3

REPRESENTATION OF BELIEF FOR STATISTICAL EVIDENCE

3.1. Introduction

When a body of evidence is based on the outcomes of statistical experiments known to be governed by any (objective) probability models, it is called "statistical evidence." One of the fundamental problems in applying IV probabilities to real-world problems is how to represent a body of statistical evidence by IV belief functions. In fact, the utility of any existing system of IV probabilities is limited by the lack of effective approaches to quantitative representation of bodies of evidence. Throughout this chapter a lower probability and an upper probability are respectively called a "support function ($\mathcal{Sp}$)" and a "plausibility function ($\mathcal{Pl}$)" implying that they provide belief measures for the class of subsets of a finite space Ω based on a body of evidence.

The most extreme type of interval-valued belief function is the "vacuous belief function" defined as

$$\mathcal{Sp}(\mathrm{A}) = \begin{cases} 0 & \text{if } \mathrm{A} \neq \Omega \\ 1 & \text{if } \mathrm{A} = \Omega \end{cases} \tag{3.1.1}$$

and

$$\mathcal{Pl}(\mathrm{A}) = \begin{cases} 0 & \text{if } \mathrm{A} = \varnothing \\ 1 & \text{if } \mathrm{A} \neq \varnothing \end{cases} \tag{3.1.2}$$

The vacuous belief function assigns [0,1] to every non-empty subset A of Ω, and [1,1] to Ω itself. Its only focal element is Ω. It is a natural model for representing complete ignorance – no evidence about Ω at all.

The next simple type is a "simple support function", a belief function based on "homogeneous" evidence – a body of evidence which precisely and unambiguously supports a single non-empty subset of Ω. Suppose $\mathcal{S}p$ is a simple support function focused on a subset A, and let $\mathcal{S}p(\mathrm{A}) = s$ ($0 \le s \le 1$). Then the support function for any $\mathrm{B} \subseteq \Omega$ is given by

$$\mathcal{S}p(\mathrm{B}) = \begin{cases} 0 & \text{if } \mathrm{B} \not\supseteq \mathrm{A} \\ s & \text{if } \mathrm{B} \supseteq \mathrm{A} \text{ but } \mathrm{B} \neq \Omega \\ 1 & \text{if } \mathrm{B} = \Omega \end{cases} \tag{3.1.3}$$

It can be easily shown that a simple support function is 2-monotone. The conjugate plausibility function of the above support function is given by

$$\mathcal{P}\ell(\mathrm{B}) = \begin{cases} 1-s & \text{if } \mathrm{A} \cap \mathrm{B} = \emptyset \\ 1 & \text{if } \mathrm{A} \cap \mathrm{B} \neq \emptyset \end{cases} \tag{3.1.4}$$

The effect of the evidence represented by the simple support function in eq. (3.1.3) is limited to providing a degree of support s for A and any subset B of Ω implied by A.

The next section introduces a possible way of constructing interval-valued belief functions based on some models in robust statistics. Shafer (1976b) presents two different methods for constructing belief functions based on a body of statistical evidence: the "linear plausibility method" and the "simplicial plausibility method." Section 3.3 examines the characteristics of the belief function in the linear plausibility method and provides its generalized scheme by weakening an assumption underlying it. The result of the second method, which is the same as that of Dempster's structure of the second kind [Dempster (1968)], is outside the scope of this report because it applies to an infinite space Ω which parametrizes all multinomial distributions and consequently presents formidable computational difficulties. Section 3.4 discusses the quantitative representation of source reliability in the context of pixel classification of multiple data sources.

3.2. Belief Functions based on Robust Estimation of Probability Measures

In robust statistics, the true underlying probability distribution is assumed to lie in a certain neighborhood of an idealized model distribution. The neighborhood describes inaccuracies in the specification of the true distribution. This section illustrates how belief functions in the form of IV probabilities can be constructed by the supremum and infimum of a class of probability measures describing the neighborhood, as defined in eq. (2.4.13) and eq. (2.4.14).

Definition 3.1. [Huber (1973)] Consider any set functions λ and υ on $\mathcal{B}$. υ is said to "dominate" λ, denoted by $\upsilon \gg \lambda$, when $\upsilon(A) \geq \lambda(A)$ for all $A \in \mathcal{B}$.

Let $\mathcal{P}_\upsilon = \{ \pi \in \mathcal{M} \mid \upsilon \gg \pi \}$ be the set of all probability measures π dominated by υ. The following lemma from Huber and Strassen (1973) shows the existence of a 2-alternating upper probability in $\mathcal{P}_\upsilon$.

Lemma 3.1. Let υ be 2-alternating. Then for every $A \in \mathcal{B}$ there exists a $\pi \in \mathcal{P}_\upsilon$ such that $\pi(A) = \upsilon(A)$. This implies that υ coincides with the upper probability determined by $\mathcal{P}_\upsilon$.

Most of the proposals listed in Huber (1981), such as ε-contamination, total variation, Prohorov distance, Kolmogorov distance, and Lèvy distance, for formalizing the notion of an inexactly specified probability measure lead to a set $\mathcal{P}_\upsilon$ defined by a certain 2-alternating set function. The following models are the ones which make sense in arbitrary probability spaces.

Let ε and δ be fractions between 0 and 1, and P_0 denote an idealized model distribution as an estimation of the actual distribution:

A. ε-contamination or **gross error model :**

$$\mathcal{P}_\upsilon = \{ \pi \in \mathcal{M} \mid \pi = (1-\varepsilon)P_0 + \varepsilon\Lambda,\ \Lambda \in \mathcal{M} \} \tag{3.2.1}$$

For any non-empty set $A \in \mathcal{B}$,

$$\upsilon(A) = \sup_\pi \mathcal{P}_\upsilon = (1-\varepsilon)P_0(A) + \varepsilon \tag{3.2.2}$$

B. Total variation model :

$$\mathcal{P}_\upsilon = \{ \pi \in \mathcal{M} \mid |\pi(A) - P_0(A)| \leq \varepsilon \text{ for all } A \in \mathcal{B} \} \qquad (3.2.3)$$

For any non-empty set $A \in \mathcal{B}$,

$$\upsilon(A) = \sup_\pi \mathcal{P}_\upsilon = \min \{P_0(A) + \varepsilon, 1\} \qquad (3.2.4)$$

For both cases, υ is the 2-alternating upper probability function, and the conjugate lower probability function is obtained as $(1-\upsilon^c)$, where the superscript c denotes the complement.

The ε-contamination model assumes that the actual probability has a gross error with an arbitrary (unknown) distribution, instead of a strict parametric model. The total variation model formalizes the possibility of unknown small deviations from the idealized model P_0 by assigning a tolerance to it.

In being applied to real problems, both models demand additional labor to find an optimal value of ε. Different ε's will result in various IV probabilities. Most of the algorithms for robust parameter estimation based on the above models adopt iterative procedures [Eom (1986), Huber (1981)]. The iterative procedures not only cost tremendous computational complexity but also raise another problem of proving convergence of estimators.

In the following section, IV belief functions are derived from the likelihood functions of observed data. Compared to the ones described in this section, they require much less computation and have readily usable mathematical formulas.

3.3. Belief Functions based on Likelihood Principle

The belief functions described in this section depend on two underlying assumptions. Before the assumptions are listed, it is necessary to define the "consonance" of belief functions.

Definition 3.2. [Shafer (1976a)] A belief function is said to be "consonant" if its focal elements are nested, i.e., if for $A_i \subseteq \Omega$ (i=1, ...,r) such that $m(A_i) > 0$ for all i and $\sum_{i=1}^{r} m(A_i) = 1$, $A_i \subset A_j$ for any $i < j$.

A simple support function is consonant while the converse is not necessarily true. The following lemma describes the nature of consonant belief functions.

Lemma 3.2. [Shafer (1976a)] Suppose $\mathcal{Sp}$: $2^\Omega \to [0, 1]$ is a support function and $\mathcal{Pl}$: $2^\Omega \to [0, 1]$ is the conjugate plausibility function. Then the following assertions are all equivalent:

(1) $\mathcal{Sp}$ is consonant.

(2) $\mathcal{Sp}(A \cap B) = \min \{ \mathcal{Sp}(A), \mathcal{Sp}(B) \}$ for all $A, B \subset \Omega$.

(3) $\mathcal{Pl}(A \cup B) = \max \{ \mathcal{Pl}(A), \mathcal{Pl}(B) \}$ for all $A, B \subset \Omega$.

(4) $\mathcal{Pl}(A) = \max \{ \mathcal{Pl}(\{\omega\}) : \omega \in A \}$ for all non-empty $A \in \Omega$.

Example 3.1. Let $\Omega = \{ \omega_1, \omega_2, \omega_3 \}$. Suppose a body of evidence E provides basic probability numbers $m(\{\omega_1\}) = 0.5$, $m(\{\omega_1, \omega_2\}) = 0.2$, $m(\Omega) = 0.3$, and $m(A) = 0$ for all other subsets A of Ω. Then the support function $\mathcal{Sp}$ of E is consonant and given as:

$\mathcal{Sp}(\{\omega_1\}) = 0.5$ $\quad$ $\mathcal{Sp}(\{\omega_2\}) = 0$ $\quad$ $\mathcal{Sp}(\{\omega_3\}) = 0$

$\mathcal{Sp}(\{\omega_1, \omega_2\}) = 0.7$ $\quad$ $\mathcal{Sp}(\{\omega_1, \omega_3\}) = 0.5$ $\quad$ $\mathcal{Sp}(\{\omega_2, \omega_3\}) = 0$

$\mathcal{Sp}(\Omega) = 1$

The plausibility function $\mathcal{Pl}$ of E is given as:

$\mathcal{Pl}(\{\omega_1\}) = 1$ $\quad$ $\mathcal{Pl}(\{\omega_2\}) = 0.5$ $\quad$ $\mathcal{Pl}(\{\omega_3\}) = 0.3$

$\mathcal{Pl}(\{\omega_1, \omega_2\}) = 1$ $\quad$ $\mathcal{Pl}(\{\omega_1, \omega_3\}) = 1$ $\quad$ $\mathcal{Pl}(\{\omega_2, \omega_3\}) = 0.5$

$\mathcal{Pl}(\Omega) = 1$

Now, suppose that the observations of a statistical experiment are governed by one of a finite set of probability models $\{ p_\omega \mid \omega \in \Omega \}$, where p_ω is an ordinary probability density function on $\mathcal{X}$ given ω. The linear plausibility function based on this body of evidence is derived from the following assumptions:

(1) the degree of plausibility of a singleton $\{\omega \mid \omega \in \Omega\}$ is proportional to p_ω;

(2) the plausibility function is consonant.

The first assumption corresponds to our intuition that an observation $\mathbf{x} \in \mathcal{X}$ favors those elements of Ω which assigns the greater chance to $\mathbf{x}$. Shafer claims that $\mathbf{x}$ should determine a plausibility function $\mathcal{Pl}_{\mathbf{x}}$ obeying

$$\mathcal{Pl}_{\mathbf{x}}(\{\omega\}) = C \cdot p_\omega(\mathbf{x}) \quad \text{for all } \omega \in \Omega \tag{3.3.1}$$

where the constant C does not depend on ω. He further shows that the first assumption, together with the second assumption of consonance, determines a unique consonant plausibility function as

$$\mathcal{Pl}_{\mathbf{x}}(A) = \frac{\max\{p_\omega(\mathbf{x}) : \omega \in A\}}{\max\{p_\omega(\mathbf{x}) : \omega \in \Omega\}} \quad \text{for all non-empty } A \subset \Omega \tag{3.3.2}$$

When A is a singleton, say $\{\omega'\}$, the consonant plausibility function gives the relative likelihood of ω' to the most likely element in Ω. The conjugate support function is obtained by

$$\mathcal{Sp}_{\mathbf{x}}(A) = 1 - \frac{\max\{p_\omega(\mathbf{x}) : \omega \in \overline{A}\}}{\max\{p_\omega(\mathbf{x}) : \omega \in \Omega\}} \quad \text{for all non-empty } A \subset \Omega \tag{3.3.3}$$

The next theorem derives the consonant basic probability assignment .

Theorem 3.1. Suppose that $\Omega^o = \{ \omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(n)} \}$ is an ordered set of Ω such that $p_{\omega^{(i)}} > p_{\omega^{(j)}}$ for any $1 \le i < j \le n$. If $\mathcal{Sp}_{\mathbf{x}}$ based on the statistical evidence is consonant, then it has the focal elements

$$A_k = \{ \omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(k)} \} \quad \text{for } k = 1, \ldots, n \tag{3.3.4}$$

Proof. Let $m_{\mathbf{x}}$ denote the basic probability function of $\mathcal{S}p_{\mathbf{x}}$. For a singleton subset A of Ω^o,

$$m_{\mathbf{x}}(A) = \mathcal{S}p_{\mathbf{x}}(A) = \begin{cases} \dfrac{p_{\omega^{(1)}}(\mathbf{x}) - p_{\omega^{(2)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} & \text{if } A = \{\omega^{(1)}\} \\ 0 & \text{otherwise} \end{cases} \tag{3.3.5}$$

Thus, $A_1 = \{\omega^{(1)}\}$ is the smallest focal element of $\mathcal{S}p_{\mathbf{x}}$. For any $A \subset \Omega^o$ ($|A| = 2$), eq. (2.2.13) gives

$$m_{\mathbf{x}}(A) = \begin{cases} \dfrac{p_{\omega^{(2)}}(\mathbf{x}) - p_{\omega^{(3)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} & \text{if } A = \{\omega^{(1)},\omega^{(2)}\} \\ 0 & \text{otherwise} \end{cases} \tag{3.3.6}$$

Let $A = \{ \omega^{(1)}, \ldots, \omega^{(i-1)},\omega^{(i+1)}, \ldots, \omega^{(k)} \}$ for $3 \le k \le n$.

$$\begin{aligned} m_{\mathbf{x}}(A) &= \sum_{B \subset A} (-1)^{|A-B|} \mathcal{S}p_{\mathbf{x}}(B) \\ &= \sum_{B \subset (A-\{\omega^{(k)}\})} [(-1)^{|A-B-1|} \mathcal{S}p_{\mathbf{x}}(B) + (-1)^{|A-B-2|} \mathcal{S}p_{\mathbf{x}}(B \cup \{\omega^{(k)}\})] \\ &= -\, m_{\mathbf{x}}(A-\{\omega^{(k)}\}) + \sum_{B \subset (A-\{\omega^{(k)}\})} [(-1)^{|A-B|} \mathcal{S}p_{\mathbf{x}}(B \cup \{\omega^{(k)}\})] \\ &= -\, m_{\mathbf{x}}(A-\{\omega^{(k)}\}) + m_{\mathbf{x}}(A-\{\omega^{(k)}\}) \\ &= 0 \end{aligned}$$

For $A_k = \{ \omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(k)} \}$ ($3 \le k \le n$-1), eq. (2.2.13) gives non-zero basic probability numbers

$$m_{\mathbf{x}}(A_k) = \frac{p_{\omega^{(k)}}(\mathbf{x}) - p_{\omega^{(k+1)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} \tag{3.3.7}$$

And,

$$m_{\mathbf{x}}(\Omega^o) = 1 - \sum_{k=1}^{n-1} m_{\mathbf{x}}(A_k) = \frac{p_{\omega^{(n)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} \tag{3.3.8}$$

Hence, the basic probability function $m_{\mathbf{x}}$ of $\mathcal{S}p_{\mathbf{x}}$ is given as

$$m_{\mathbf{x}}(A) = \begin{cases} \dfrac{p_{\omega^{(k)}}(\mathbf{x}) - p_{\omega^{(k+1)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} & \text{for } A = \{\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(k)}\} \ (1 \le k \le n-1) \\ \dfrac{p_{\omega^{(n)}}(\mathbf{x})}{p_{\omega^{(1)}}(\mathbf{x})} & A = \Omega^{\circ} \ (=\Omega) \\ 0 & \text{otherwise} \end{cases} \tag{3.3.9}$$

and the focal elements of $\mathcal{S}p$ are $A_k = \{ \omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(k)} \}$ for $k = 1, \ldots, n$. ■

Although the consonant belief function described above is simple to implement, its application is limited to the particular cases where the consonance assumption is satisfied. Indeed, Shafer made a remark regarding his method; "..... these assumptions must be regarded as conventions for establishing degrees of support, conventions that can be justified only by their general intuitive appeal and by their success in dealing with particular examples." [Shafer (1976a)]

A generalized scheme of the consonant support and plausibility functions can be formulated by weakening the consonance assumption.

Definition 3.3. A support function $\mathcal{S}p$: $2^{\Omega} \rightarrow [0, 1]$ is said to be "partially consonant" if there exists a partition $\{ \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_r \}$ of Ω and $\mathcal{S}p$ is consonant in every $\mathcal{W}_k$ for $k=1, \ldots, r$.

In the problem of classifying remotely sensed data, Ω represents a set of information classes. The information classes in remote sensing can be partitioned into major ground-cover types, e.g., soil, vegetation, and water [Swain et al. (1978)]. This hierarchical structure of the information classes motivates the partitioning of Ω for partial consonance.

The following theorem and lemma derive the partially consonant basic probability assignment and the corresponding interval-valued probabilities.

Theorem 3.2. Suppose that $\mathcal{Sp}$ is partially consonant on a partition $\{ \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_r \}$ of Ω. Let $\mathcal{W}_k^{\rho} = \{\omega_k^{(1)}, \omega_k^{(2)}, \ldots, \omega_k^{(n_k)}\}$ denote an ordered set of $\mathcal{W}_k$ such that $p_{\omega_k^{(i)}} > p_{\omega_k^{(j)}}$ for any $1 \le i < j \le n_k$, where $\sum_{k=1}^{r} n_k = n$. Then the basic probability function m of $\mathcal{Sp}$ is given as

$$m(A)=\begin{cases} C_p \cdot \{p_{\omega_k^{(l)}} - p_{\omega_k^{(l+1)}}\} & \text{for } A=\{\omega_k^{(1)}, \ldots, \omega_k^{(l)}\}\ (1 \le l \le n_k - 1) \\ C_p \cdot p_{\omega_k^{(n_k)}} & \text{for } A = \mathcal{W}_k^{\rho} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } 1 \le k \le r \tag{3.3.10}$$

where

$$C_p = [\sum_{k=1}^{r} \max \{ p_\omega : \omega \in \mathcal{W}_k \}]^{-1} \tag{3.3.11}$$

Proof. Since $\mathcal{Sp}$ is partially consonant on $\{ \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_r \}$, it is consonant in every $\mathcal{W}_k$ for k=1, ..., r. Using eq. (3.3.9), we can derive eq. (3.3.10). To prove this theorem, it is sufficient to derive eq. (3.3.11).

$$1 = \sum_{A \subseteq \Omega} m(A) = \sum_{k=1}^{r} \{ \sum_{A \subseteq \mathcal{W}_k} m(A) \} = C_p \cdot \sum_{k=1}^{r} \max\{ p_\omega : \omega \in \mathcal{W}_k \}$$

$$\therefore\ C_p = [\sum_{k=1}^{r} \max \{ p_\omega : \omega \in \mathcal{W}_k \}]^{-1}$$

Thus the theorem is proved. ■

Lemma 3.3. The partially consonant plausibility function and support function corresponding to eq. (3.3.10) are

$$\mathcal{Pl}(A) = C_p \cdot \sum_{k=1}^{r} \max\{p_\omega : \omega \in A \cap \mathcal{W}_k\} \tag{3.3.12}$$

$$= \sum_{k=1}^{r} \max\{\mathcal{Pl}(\{\omega\}) : \omega \in A \cap \mathcal{W}_k\} \tag{3.3.13}$$

$$\mathcal{Sp}(A) = \sum_{k=1}^{r} [\max\{p_\omega : \omega \in \mathcal{W}_k\} - \max\{p_\omega : \omega \in \bar{A} \cap \mathcal{W}_k\}] \tag{3.3.14}$$

Proof. Use eq. (2.2.13) and eq. (2.2.14). ■

Partial consonance is weaker than consonance in the sense that it includes consonance when r = 1, i.e., the partition of Ω is Ω itself. In the other extreme case where $r=n$, i.e., the partition consists of n singleton subsets of Ω, the partially consonant support function becomes the Bayesian probability function ($\mathcal{Sp}(\{\omega_i\}) = \mathcal{Pl}(\{\omega_i\}) = m(\{\omega_i\})$ for i=1, ...,n). While partial consonance gives a flexibility to Shafer's linear plausibility method, it raises the problem of how to determine the optimal partition of Ω; i.e., the partition which will give the best classification accuracy. In practice, the partition must be chosen based on relationship among the classes in the application at hand.

Example 3.2. Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Suppose that a single observation $\mathbf{x}$ provides $p_{\omega_1}(\mathbf{x}) = 0.5$, $p_{\omega_2}(\mathbf{x}) = 0.3$, $p_{\omega_3}(\mathbf{x}) = 0.15$, and $p_{\omega_4}(\mathbf{x}) = 0.05$. Table 3.1 shows the values of $m_\mathbf{x}$, $\mathcal{Sp}_\mathbf{x}$, and $\mathcal{Pl}_\mathbf{x}$ for all subsets of Ω in both cases of consonance and partial consonance on the partition $\{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$.

It is very interesting that both intervals given by the belief functions contain the additive probability ($p_A(\mathbf{x})$) for every A except $\{\omega_1, \omega_2\}$ and $\{\omega_3, \omega_4\}$ in partial consonance. Compared with the consonance case, the partially consonant belief function always provides intervals of less width, correspondingly less degrees of uncertainty. It means that the assumption of partial consonance requires more knowledge about a given body of evidence.

Note that low $\mathcal{Sp}$ do not necessarily imply low $\mathcal{Pl}$ whereas high $\mathcal{Sp}$ always imply high $\mathcal{Pl}$. We can also observe two relations: (1) $\mathcal{Sp}(A) + \mathcal{Sp}(\bar{A}) \leq 1$, and (2) $\mathcal{Pl}(A) + \mathcal{Pl}(\bar{A}) \geq 1$ for every A. The first relation indicates that it is hardly possible for both A and $\bar{A}$ to be well supported, and the second one is interpreted as either one of A and $\bar{A}$ or possibly both must be highly plausible.

The belief functions described in this section are considered to be based on the Likelihood Principle because they are expressed in terms of likelihood functions, eq. (3.3.2), (3.3.3), (3.3.12), and (3.3.14). They are obtained by transforming the assessment of statistical evidence already in the form of point-valued likelihood functions into interval-valued probability models.

Table 3.1. Consonant and Partially Consonant Belief Functions based on a Single Observation.

		Consonance			Partial Consonance		
A	$p_A(\mathbf{x})$	$m_\mathbf{x}$	$\mathcal{S}p_\mathbf{x}$	$\mathcal{P}\ell_\mathbf{x}$	$m_\mathbf{x}$	$\mathcal{S}p_\mathbf{x}$	$\mathcal{P}\ell_\mathbf{x}$
$\{\omega_1\}$	0.50	0.4	0.4	1.0	0.31	0.31	0.77
$\{\omega_2\}$	0.30	0.0	0.0	0.6	0.00	0.00	0.46
$\{\omega_3\}$	0.15	0.0	0.0	0.3	0.15	0.15	0.23
$\{\omega_4\}$	0.05	0.0	0.0	0.1	0.00	0.00	0.08
$\{\omega_1, \omega_2\}$	0.80	0.3	0.7	1.0	0.46	0.77	0.77
$\{\omega_1, \omega_3\}$	0.65	0.0	0.4	1.0	0.00	0.46	1.00
$\{\omega_1, \omega_4\}$	0.55	0.0	0.4	1.0	0.00	0.31	0.85
$\{\omega_2, \omega_3\}$	0.45	0.0	0.0	0.6	0.00	0.15	0.69
$\{\omega_2, \omega_4\}$	0.35	0.0	0.0	0.6	0.00	0.00	0.54
$\{\omega_3, \omega_4\}$	0.20	0.0	0.0	0.3	0.08	0.23	0.23
$\{\omega_1, \omega_2, \omega_3\}$	0.95	0.2	0.9	1.0	0.00	0.92	1.00
$\{\omega_1, \omega_2, \omega_4\}$	0.85	0.0	0.7	1.0	0.00	0.77	0.85
$\{\omega_1, \omega_3, \omega_4\}$	0.70	0.0	0.4	1.0	0.00	0.54	1.00
$\{\omega_2, \omega_3, \omega_4\}$	0.50	0.0	0.0	0.6	0.00	0.23	0.69
Ω	1.00	0.1	1.0	1.0	0.00	1.00	1.00

3.4. Representation of Source Reliability

Since information sources in remote sensing and GIS are in general not equally reliable, they usually provide various degrees of support for an event. In order to incorporate a relative quality factor, so-called "degree of reliability," of individual data sources into the combination of multiple evidence, reliability should be represented quantitatively. Although the belief functions in the form of IV probabilities are useful to represent the uncertainty in describing the degrees of support for individual events, they do not take into account the relative source reliability representing a body of evidence as a whole.

As a simple example, consider a problem of classifying a pixel using two data sources as depicted in Figure 3.1. Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the vectors of the pixel obtained from Source 1 and Source 2 respectively. Based on Source 1 alone, the pixel seems to belong to ω_1 while according to the other source it is more likely to come from ω_2. If there is a priori information concerning how reliable each data source is, it would be reasonable to make a decision on the classification of the pixel using the source reliabilities as well as the probabilistic information from both sources.

Benediktsson and Swain (1989) have used three statistical measures, overall classification accuracy, weighted average separability, and equivocation, to quantify reliability of sources in the classification of multisource data. Which measure should be applied to a particular problem depends on the meaning of the reliability of a source in the context of the problem, that is, the sense in which the source is called reliable. For the problem of multisource data classification, it is quite natural that a source is called reliable when it gives higher classification accuracy. Measuring reliability of a source based on classification accuracy is straightforward. It is usually computed from the overall classification accuracy over a representative set of training samples.

A statistical separability measure such as Jeffries-Matusita (J-M) distance, Bhattacharyya distance, or (Transformed) Divergence is an alternative to the numerical representation of source reliability assuming that a data source provides higher classification accuracy when information classes are more separable in the source. For example, the J-M distance defined as follows is a measure of statistical separability of pairs of classes:

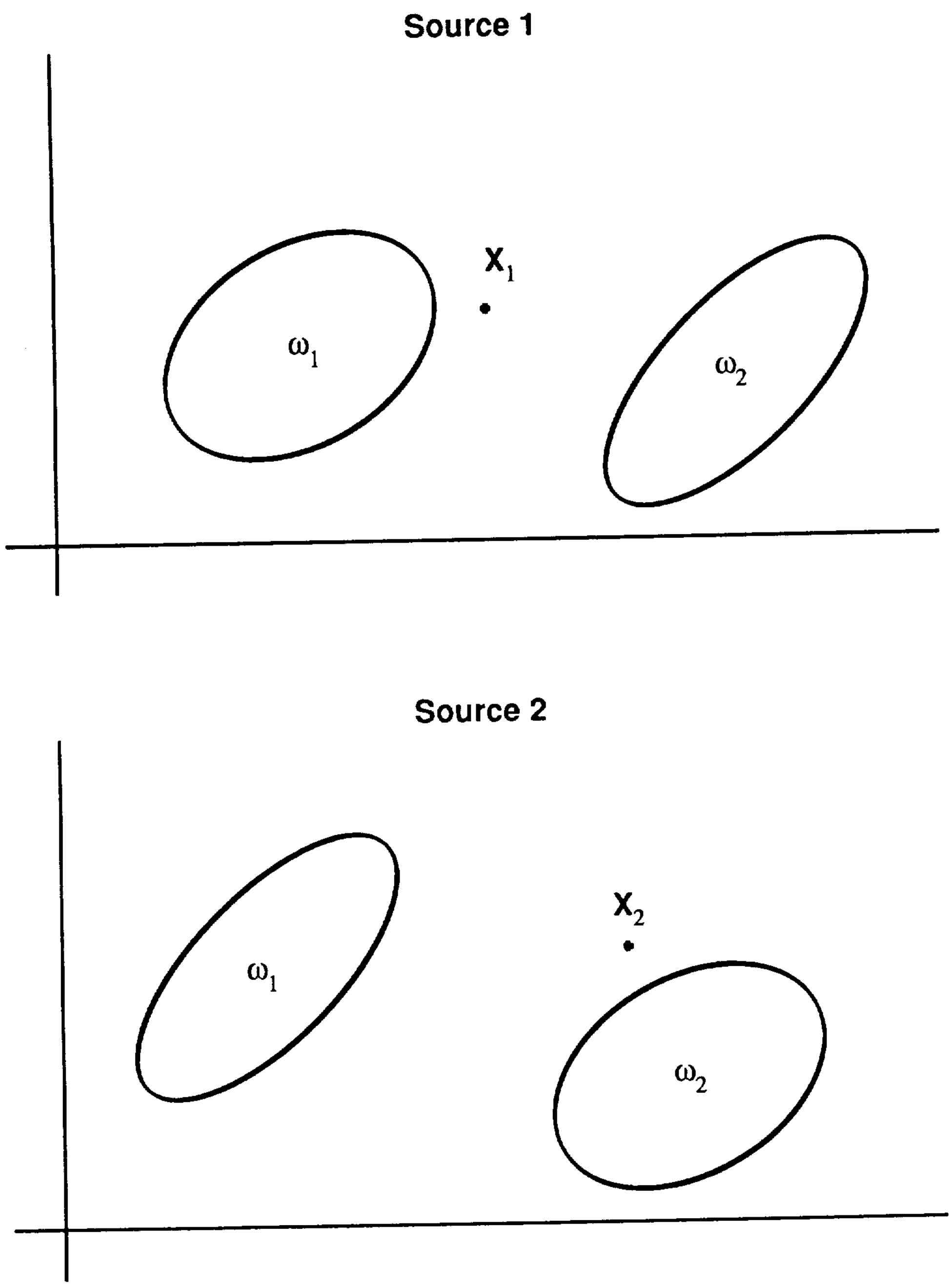


Figure 3.1 An Example of Conflicting Evidence in Multisource Data Classification.

$$\mathcal{J}_{ij} = \left\{ \int_{X} [\sqrt{p(\mathbf{X}|\omega_i)} - \sqrt{p(\mathbf{X}|\omega_j)}\,]^2 d\mathbf{X} \right\}^{\frac{1}{2}} \tag{3.4.1}$$

where $p(\mathbf{X}|\omega_i)$ is the probability density function of class ω_i. When each class is assumed to have a normal density function $\mathcal{N}(M_i,\Sigma_i)$ ($i = 1, \ldots, n$), the above equation is reduced to

$$\mathcal{J}_{ij} = \sqrt{2(1 - \exp(-\beta_{ij}))} \tag{3.4.2}$$

where β_{ij} is the Bhattacharyya distance between ω_i and ω_j defined as:

$$\beta_{ij} = \frac{1}{8}(M_i - M_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}(M_i - M_j) + \frac{1}{2}\log_e\left[\frac{\left|\frac{(\Sigma_i + \Sigma_j)}{2}\right|}{\sqrt{|\Sigma_i|\cdot|\Sigma_j|}}\right] \tag{3.4.3}$$

The average J-M distance over all class pairs is given as:

$$\mathcal{J}_{av} = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} P(\omega_i)\cdot P(\omega_j)\cdot \mathcal{J}_{ij} \tag{3.4.4}$$

where $P(\omega_i)$ is the prior probability of ω_i.

For the normal distribution case, Transformed Divergence between ω_i and ω_j is defined as:

$$\mathcal{D}^t_{ij} = 2\left[1 - \exp\left(\frac{-\mathcal{D}_{ij}}{8}\right)\right] \tag{3.4.5}$$

where

$$\mathcal{D}_{ij} = \frac{1}{2}\mathrm{tr}\left[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})\right] + \frac{1}{2}\mathrm{tr}\left[(\Sigma_i^{-1} + \Sigma_j^{-1})(M_i - M_j)(M_i - M_j)^T\right] \tag{3.4.6}$$

Then the average Transformed Divergence over all class pairs is given as:

$$\mathcal{D}^t_{av} = \sum_{i=1}^{i=n}\sum_{j=1}^{j=n} P(\omega_i)\cdot P(\omega_j)\cdot \mathcal{D}^t_{ij} \tag{3.4.7}$$

Equivocation is the class separability measure corresponding to Shannon's entropy measure [Devijver and Kittler (1982)]. Benediktsson et al.

(1989) use equivocation to measure the reliability with which classes identifiable by means of each data source can be used to identify the information classes of interest in a given application.

The three measures briefly reviewed above are related indirectly to the classification accuracy of the source. The source reliability can have a little different meaning in the mathematical framework of the theory of evidence. In the previous example of Figure 3.1, assume that Source 1 is a main data source and Source 2 an ancillary data source, and that the main source gives higher classification accuracy over training samples. Then Source 2 can be considered as reliable as Source 1 if there is little overall conflict between them in providing evidence for classifying observations. And its reliability will decrease according to the extent of conflict with Source 1. The following definition gives a notion of quantifying source reliability based on a measure of the extent of the conflict between the belief functions provided by two entirely distinct bodies of evidence.

Definition 3.4. [Shafer (1976a)] Assume that $\mathcal{B}el_1$ and $\mathcal{B}el_2$ are belief functions provided by two bodies of evidence. Let m_1 and m_2 denote the basic probability assignments of $\mathcal{B}el_1$ and $\mathcal{B}el_2$, respectively. The measure of conflict between $\mathcal{B}el_1$ and $\mathcal{B}el_2$ is defined as:

$$k = \sum_{A_i \cap B_j = \emptyset} m_1(A_i) \cdot m_2(B_j) \tag{3.4.8}$$

k is a fraction between 0 and 1. When $\mathcal{B}el_1$ and $\mathcal{B}el_2$ have no conflict, $k=0$. If they are completely contradictory, $k=1$. After k is computed for every pixel, the average measure of conflict between the sources is obtained as:

$$\mathcal{K} = E[k] = \int_0^1 k \cdot p(k)\, dk \tag{3.4.9}$$

where $p(k)$ is the probability density function of k.

In order to illustrate their uses and compare the performances, the average J-M distance ($\mathcal{J}_{av}$), the average Transformed Divergence ($\mathcal{D}^t_{av}$), and the average measures of conflict between pairs of sources in the Anderson River data set were computed. The data set has 6 sources as shown in Table 3.2. For more detail about this data set, see Section 6.2. For this experiment, six information classes are defined. Each class has 100 training samples uniformly scattered over the test fields. The first row in Table 3.2 shows the overall classification accuracy (OCA) over the training samples using the Maximum Likelihood classification. Although most of the classes are not normally distributed in the topographic data sources (see Figures 6.9 through 6.12), they were assumed to be so in the calculations. The maximum values of $\mathcal{J}_{av}$ and $\mathcal{D}^t_{av}$ are $\sqrt{2}$ and 2, respectively. When they are directly used as measures of source reliability, they should be divided by the corresponding maximum value so that their maximum is 1. Table 3.2 shows that the separability measures agree with the overall classification accuracy in ranking the sources for their relative reliabilities. Based on the measures in Table 3.2, the sources can be ranked from best to worst as A/B MSS, Elevation, SAR-Shallow, SAR-Steep, Aspect, and Slope.

Table 3.2 Overall Classification Accuracy (OCA), Average J-M Distance ($\mathcal{J}_{av}$), and Average Transformed Divergence ($\mathcal{D}^t_{av}$) of Sources in Anderson River Data Set (Training Samples).

	A/B MSS	SAR Shallow	SAR Steep	Aspect	Elevation	Slope
OCA (%)	83.5	34.7	33.5	30.3	45.8	29.2
$\mathcal{J}_{av}$	1.09	.57	.49	.35	.66	.21
$\mathcal{D}^t_{av}$	1.58	.52	.40	.32	.82	.08

The average measures of conflict between pairs of sources in the same data set were computed for the training samples and the combined training and test samples, and the results are listed in Table 3.3 and 3.4, respectively. The type of the belief function used was the consonant belief function. Since the probability density function of k in eq. (3.4.8) was not known, the histogram approach was used to estimate $p(k)$. The results show that Elevation and SAR-

Shallow sources have less conflict with A/B MSS in providing bodies of evidence, compared to the remaining sources. Knowing that A/B MSS source gives the highest overall classification accuracy, relative degrees of reliability of the other sources can be assigned according to their measures of conflict with A/B MSS such that the less conflicting, the more reliable. Thus the sources can be ranked from best to worst as A/B MSS, Elevation, SAR-Shallow, Aspect,

Table 3.3 Average Measures of Conflict between Pairs of Sources in Anderson River Data Set. (Using Consonant Belief Function with Training Samples)

	SAR Shallow	SAR Steep	Aspect	Elevation	Slope
A/B MSS	.388	.586	.543	.327	.565
SAR Shallow		.269	.387	.429	.404
SAR Steep			.436	.437	.341
Aspect				.588	.463
Elevation					.543

Table 3.4 Average Measures of Conflict between Pairs of Sources in Anderson River Data Set. (Using Consonant Belief Function with All Samples)

	SAR Shallow	SAR Steep	Aspect	Elevation	Slope
A/B MSS	.407	.585	.538	.351	.550
SAR Shallow		.284	.385	.453	.385
SAR Steep			.437	.462	.344
Aspect				.572	.428
Elevation					.513

Slope, and SAR-Steep. The average measure of conflict agrees with the separability measures and OCA only in ranking the first three sources (A/B MSS, Elevation, and SAR-Shallow). In the multisource data classification with this data set, the remaining sources (SAR-Steep, Aspect, and Slope) will be considered as equally reliable as the 4th.

There are two problems in quantifying source reliability based on the average measure of conflict. First, the values of the average measures of conflict will vary depending on what kind of belief function is used in eq. (3.4.8). However, as long as the belief function represents the body of evidence properly, the ranking of the sources in terms of their relative reliabilities will remain the same. Second, even the ranking of the sources depends on the prior information regarding which is the most reliable source. For example, in Table 3.4, if SAR-Shallow were assumed to be the most reliable, then the second most reliable source would be SAR-Steep instead of A/B MSS.

One of the advantages of the measure of conflict is that it provides the relative reliabilities between all pairs of sources. When the "most reliable source" changes from one to another due to the meaning of the reliability in the context of a problem, the measure of conflict gives the ranking of the sources according to the new most reliable source.

Furthermore, the measure of conflict can be computed for test samples as well as training samples. In the above case, there is not much difference between the measures of conflict for the training samples and the entire sample because the training samples are uniformly distributed over the entire sample. On the other hand, when training samples are limited and poor representatives of test samples, there may be difference between the measures of conflict obtained from the training samples and from the entire sample.

Both the separability measures and the measure of conflict give information for ranking multiple sources in the sense of their relative reliabilities, but a quantitative method of computing the absolute reliabilities of the sources is still unknown.

Once the relative reliabilities of the data sources are given, they are included in the multisource data analysis by "discounting" belief functions [Shafer (1976a)]. Suppose α denotes the relative reliability assigned to a given source, where $0 \le \alpha \le 1$. By discounting, the basic probability number of every subset A of Ω is reduced from $m(A)$ to $\alpha \cdot m(A)$ and the basic probability number of Ω increases from $m(\Omega)$ to $m(\Omega)+\alpha$.

3.5. Summary

This chapter has focused on the construction of interval-valued belief functions for statistical evidence and the quantitative representation of source reliability. Belief functions can be obtained in the form of IV probabilities from the supremum and infimum of a class of probability measures. Two models for robust estimation of probability measure, the ε-contamination model and the total variation model, were introduced to formalize the class of probability measures. Then the IV belief functions based on the Likelihood Principle were constructed. Although they require some underlying assumptions (consonance or partial consonance), they have mathematically simple and readily usable formulas. The required assumptions are not difficult to satisfy in practical applications of this approach.

In order to include the relative reliabilities of sources in a multisource data analysis, the attempts to quantitatively represent the degree of reliability by the average Jeffries-Matusita distance, the average Transformed Divergence, and the average measure of conflict between pairs of sources were made. Their performances were compared by applying them to an actual multisource data set.

In the experiments described in Chapter 6, the belief functions based on the Likelihood Principle will be implemented, and the multiple sources will be ranked based on the average J-M distance and the average measure of conflict.

In this chapter, the contribution of this research is in the representation of statistical evidence by IV probabilities such as consonant and partially consonant IV probabilities. Theorems 3.1 and 3.2, Definition 3.3, and Lemma 3.3 are originally formulated and proved in this report.

CHAPTER 4

COMBINATION OF BELIEF FOR STATISTICAL EVIDENCE

4.1. Introduction

To base inferences and decisions on all available information, it is necessary to combine the information from various sources. The role of rules for combining evidence is to integrate the conditional knowledge about states of nature based on each body of evidence into combined knowledge based on the total evidence. Combination rules may be formulated in various ways; they may depend on the characteristics of the problem, the experience of the knowledge engineer, and the mathematical theories on which the rules are founded.

Various procedures for the formation of a consensus of opinions have been suggested in the group decision problems [French (1981), Genest (1986), and Winkler (1968)], some on pragmatic grounds, others justified axiomatically. The following formulas are most typical ones among them.

Consider the situation where there are m sources of information, each providing its subjective probability π_i (i=1, ...,m) over $\mathcal{B}$. Here π_i can be any kind of additive probability measure according to the context of problems.

Linear Opinion Pool defines the overall probability measure π as a weighted mean of π_i's:

$$\pi(A) = \sum_{i=1}^{m} \gamma_i \cdot \pi_i(A) \quad \text{for all } A \in \mathcal{B} \tag{4.1.1}$$

where γ_i (i = 1, ..., m) are positive weights assigned to each source and satisfying $\sum_{i=1}^{m} \gamma_i = 1$.

Independent Opinion Pool assumes that the information sources are "independent" and defines the overall probability measure simply as a product of the individual measures:

$$\pi(A) = \kappa \cdot \left[\prod_{i=1}^{m} \pi_i(A) \right] \quad \text{for all } A \in \mathcal{B} \tag{4.1.2}$$

where κ is an appropriate normalizing constant so that $\pi(\cdot)$ become additive.

Logarithmic Opinion Pool is a generalization of the independent opinion pool. The overall probability measure is given as:

$$\pi(A) = \kappa \cdot \left[\prod_{i=1}^{m} \{\pi_i(A)\}^{\alpha_i} \right] \quad \text{for all } A \in \mathcal{B} \tag{4.1.3}$$

where α_i is any positive real number representing the relative reliability of the i^{th} source.

A deficiency of the linear opinion pool is that the individual probabilities do not reinforce the others. The combined measure given in (4.1.1) is always between the maximum and the minimum values of π_i,

$$\min_{i=1,\ldots,m} \pi_i(A) \leq \pi(A) \leq \max_{i=1,\ldots,m} \pi_i(A) \quad \text{for all } A \in \mathcal{B} \tag{4.1.4}$$

The other two schemes have the "zero probability property", viz.,

$$\text{If } \pi_i(A) = 0 \text{ for any } i, \quad \text{then } \pi(A) = 0 \tag{4.1.5}$$

which makes the combined measure too sensitive to a small probability measure. More in-depth discussions are found in French (1985) and Berger (1985).

In rule-based inferencing systems, several subjective Bayesian updating rules have been proposed to modify the probabilities of hypotheses as each piece of evidence is provided. These rules are derived by applying one or two statistical independence assumptions to Bayes' rule and successfully used in rule-based expert systems such as PROSPECTOR [Duda et al. (1979)] and MYCIN [Shortliffe (1976)]. However, there have been some controversies over the inconsistency between the independence assumptions and their updating rules.

During the last decade Dempster's rule has been receiving more attention from many researchers in various areas of science and engineering. It is a generalization of Bayesian inference, including the subjective Bayesian updating rules as the special cases for which the domain-specific knowledge is precise.

The objective of this chapter is to investigate the inferencing mechanisms of the subjective Bayesian updating rules and Dempster's rule in combining multiple evidence when they are formulated as set-theoretic functional equations. They are given a behavioral interpretation in terms of the desirable properties which agree with human intuition. The independence assumptions underlying them and the robustness to small variations in probability measures are studied.

4.2. Properties of Combination Rules

For computer-based, quantitative techniques of multisource data analysis the rules for combining evidence must be formulated as functional equations computing the degree of belief based on the total evidence from degrees of belief based on each single piece of evidence.

As given earlier, Ω consists of a finite number of exhaustive and mutually exclusive events and $\mathcal{B}$ is a Boolean algebra of all subsets of Ω. Let $\mathcal{E}$ be a set of multiple bodies of evidence $\{ E_1, E_2, ..., E_m \}$ and $B(A_j|E_i) = b_i$ (i=1, ..., m) denote the degree of conditional belief for $A_j \in \mathcal{B}$ given a body of evidence E_i. Then a rule for combining evidence expresses the degree of belief based on the total evidence, $B(A_j|E_1 \& E_2 \& \ldots \& E_m)$, as a function on the set of evidence given the knowledge of $B(A_j|E_i)$ for i=1, ..., m. Several properties of combining rules are proposed by Cheng and Kashyap (1986) to provide guidelines for constructing the rules as numerical formulas. In this section those properties are formally stated.

Definition 4.1. Let $\mathcal{F}$ denote a function representing a rule for combining evidence. $\mathcal{F}$ is said to be "decomposable" if there exists a function f such that

$$\mathcal{F}(b_1, \ldots, b_m) = f(f(\ldots f(f(b_1, b_2), b_3), \ldots), b_m) \qquad (4.2.1)$$

where f is called a "binary operator" of $\mathcal{F}$.

In general, $\mathcal{F}$ and f (if it exists) are assumed to be continuous except at the endpoints. This corresponds to the idea that the human reasoning process is not abrupt.

If we assume that the final degree of belief depends only on the set of evidence and not on the order in which the pieces of evidence are combined, different orderings of evidence in combination should produce the same result. The properties in the following definitions are essential to any combination rule for exchangeability of the order of evidence and for decomposability of its numerical function into a binary operator.

Definition 4.2. $\mathcal{F}$ is "commutative" if it has a binary operator f such that

$$f(b_i, b_j) = f(b_j, b_i) \qquad (4.2.2)$$

for any pair of i, j ($1 \leq$ i, j $\leq m$).

Definition 4.3. $\mathcal{F}$ is "associative" if it has a binary operator f such that

$$f(f(b_i, b_j), b_k) = f(b_i, f(b_j, b_k)) \qquad (4.2.3)$$

for all i, j, and k ($1 \leq$ i, j, k $\leq m$).

In every numerical representation of belief, a stronger belief is represented by a larger number. Imagine that two degrees of belief provided by different pieces of evidence, say b_i and b_j, are to be combined respectively with another degree of belief b_k. Suppose $b_i > b_j$, i.e., b_i represents a relatively stronger belief than b_j, then it is natural that the combination of b_i with b_k produces a larger number than the combination of b_j with b_k. The next definition gives the mathematical expression of this property.

Definition 4.4. $\mathcal{F}$ is said to be "monotonous" if its binary operator f satisfies the condition

$$\text{if } b_i \geq b_j, \quad \text{then } f(b_i, b_k) \geq f(b_j, b_k) \tag{4.2.4}$$

for any b_k.

Monotonicity is a rather general property compared to commutativity and associativity because it should hold even for combining functions which do not have binary operators. It is true that when one piece of evidence is replaced by one providing stronger belief, $\mathcal{F}$ should produce a larger value.

Definition 4.5. $\mathcal{F}$ is "positively reinforcing" if

$$\mathcal{F}(b_1, \ldots, b_m) \geq \max_{i=1,\ldots,m} \{b_i\} \tag{4.2.5}$$

or its binary operator f satisfies

$$f(b_i, b_j) \geq \max\{ b_i, b_j \} \tag{4.2.6}$$

Definition 4.6. $\mathcal{F}$ is "negatively reinforcing" if

$$\mathcal{F}(b_1, \ldots, b_m) \leq \min_{i=1,\ldots,m} \{b_i\} \tag{4.2.7}$$

or its binary operator f satisfies

$$f(b_i, b_j) \leq \min\{ b_i, b_j \} \tag{4.2.8}$$

Positive (Negative) reinforcement means that the belief based on the total evidence is stronger (weaker) than the belief based on any single piece of evidence.

In the following two sections, the definitions of desirable properties of a combination rule play a role of interpreting inferencing mechanisms of the subjective Bayesian updating rules and Dempster's rule of combination.

4.3. Subjective Bayesian Updating Rules

The three different subjective Bayesian updating rules have been obtained by applying one or two statistical independence assumptions to Bayes' rule.

Global independence over $\mathcal{E} = \{ E_1, E_2, ..., E_m \}$ is defined as:

$$P(E_1 \& E_2 \& \ldots \& E_m) = \prod_{i=1}^{m} P(E_i) \tag{4.3.1}$$

Conditional independence over $\mathcal{E}$ given a proposition is defined as:

$$P(E_1 \& E_2 \& \ldots \& E_m \mid A_j) = \prod_{i=1}^{m} P(E_i|A_j) \quad \text{for all } j=1,\ldots,n \tag{4.3.2}$$

Conditional independence over $\mathcal{E}$ given the negation of a proposition is defined as:

$$P(E_1 \& E_2 \& \ldots \& E_m \mid \bar{A}_j) = \prod_{i=1}^{m} P(E_i|\bar{A}_j) \quad \text{for all } j=1,\ldots,n \tag{4.3.3}$$

Using Bayes' rule, the posterior probability of A_j given the combined body of evidence can be written as

$$P(A_j \mid E_1 \& E_2 \& \ldots \& E_m) = \frac{P(E_1 \& E_2 \& \ldots \& E_m \mid A_j) \cdot P(A_j)}{P(E_1 \& E_2 \& \ldots \& E_m)} \tag{4.3.4}$$

Under the assumption of conditional independence in eq. (4.3.2), the Bayes' formula in eq. (4.3.4) can be written as:

$$P(A_j \mid E_1 \& E_2 \& \ldots \& E_m) = \frac{P(A_j) \prod_{i=1}^{m} \frac{P(A_j \mid E_i)}{P(A_j)}}{\sum_{k=1}^{n} \left\{ P(A_k) \prod_{i=1}^{m} \frac{P(A_k \mid E_i)}{P(A_k)} \right\}} \tag{4.3.5}$$

This rule has been used by Cheng and Fu (1985) in a rule-based reasoning system for diagnosing diseases.

The global independence assumption in equation (4.3.1) together with the conditional independence in equation (4.3.2) rewrites Bayes' rule as

$$P(A_j \mid E_1 \& E_2 \& \ldots \& E_m) = \prod_{i=1}^{m} \frac{P(E_i \mid A_j)}{P(E_i)} = \prod_{i=1}^{m} \frac{P(A_j \mid E_i)}{P(A_j)} \tag{4.3.6}$$

Swain et al. (1985) have used this formula to construct a global membership function. Also, the rule for combining measures of belief and disbelief in MYCIN has been obtained from the binary form (m=2) of eq. (4.3.6) after translating probabilities to its own measures of belief and disbelief.

Also, applying both conditional independence assumptions to Bayes' rule, we can derive the following combining function

$$P(A_j \mid E_1 \& E_2 \& \ldots \& E_m) = \frac{\prod_{i=1}^{m} P(A_j \mid E_i)}{\prod_{i=1}^{m} P(A_j \mid E_i) + \prod_{i=1}^{m} P(\bar{A}_j \mid E_i)} \tag{4.3.7}$$

which is the updating rule used in PROSPECTOR, a rule-based computer consultant system intended to aid geologists in evaluating the favorability of an exploration site for occurrences of ore deposits of particular types. Interestingly, this rule is a special case of the rule in eq. (4.3.5) when $P(A_j) = \frac{1}{n}$ for all j. Nevertheless, it is more appealing because this rule expresses the combined measure in terms of only the conditional probabilities of individual bodies of evidence. Note that the rules expressed in eq. (4.3.5) and eq. (4.3.6) include the effect of prior probabilities in combining bodies of evidence.

All of the subjective Bayesian updating rules described in this section are decomposable. The binary operator of each rule can be easily obtained by setting $m = 2$. In the following, we will take a closer look at the characteristics of the rule expressed in eq. (4.3.7).

For a subset A of Ω, set $P(A|E_1) = p_1$ and $P(A|E_2) = p_2$. Since $P(\cdot)$ is additive, $P(\bar{A}|E_i) = 1-p_i$ for i =1, 2. The binary operator of the rule in equation (4.3.7) is given as:

$$f_A(p_1, p_2)\ (= P(A \mid E_1\&E_2)) = \frac{p_1 \cdot p_2}{p_1 \cdot p_2 + (1-p_1)\cdot(1-p_2)} \tag{4.3.8}$$

The above binary operator has the following properties:

(1) Positively reinforcing when $p_1, p_2 \geq \frac{1}{2}$, and negatively reinforcing when $p_1, p_2 \leq \frac{1}{2}$. Not defined in terms of reinforcement when $p_1 \leq \frac{1}{2}$ and $p_2 \geq \frac{1}{2}$, or $p_1 \geq \frac{1}{2}$ and $p_2 \leq \frac{1}{2}$.

(2) When $p_1 = \frac{1}{2}$, $f_A(p_1, p_2) = p_2$; $\frac{1}{2}$ is the identity of the binary operator. Since the rule deals with additive probabilities, $\frac{1}{2}$ represents the total ignorance of evidence for the rule.

(3) When $p_1 = 0$ (or 1), $f_A(p_1, p_2) = 0$ (or 1) except $p_2 = 1$ (or 0); 0 and 1 are the annihilators of the binary operator, that is, when E_1 provides complete certainty either for A ($p_1 = 1$) or for $\bar{A}$ ($p_1 = 0$), the other body of evidence cannot affect the combined belief measure.

(4) $f_A(0, 1)$ and $f_A(1, 0)$ are not defined; this rule cannot combine two bodies of evidence which are completely contradictory.

Figure 4.1 is a graphical interpretation of the binary operator based on set-theoretic operations. In the figure, the upper-left rectangle represents the degree of belief for A based on the combined evidence while the lower-right rectangle represents the degree of belief against A based on the combined evidence. The upper-right and lower-left rectangles represent the measure which fails to be committed to either A or $\bar{A}$.

The question now is which independence assumption is empirically

	$P(A\|E_2) = p_2$	$P(\overline{A}\|E_2) = 1 - p_2$
$P(A\|E_1) = p_1$	$A \cap A = A$ $p_1 \cdot p_2$	$A \cap \overline{A} = \varnothing$ $p_1 \cdot (1-p_2)$
$P(\overline{A}\|E_1) = 1 - p_1$	$\overline{A} \cap A = \varnothing$ $(1-p_1) \cdot p_2$	$\overline{A} \cap \overline{A} = \overline{A}$ $(1-p_1) \cdot (1-p_2)$

Figure 4.1 Graphical Interpretation of Binary Operator of Subjective Bayesian Updating Rule in Equation (4.3.7)

more reasonable and yields a better updating scheme. Controversially, it has been shown that there is inconsistency between some independence assumptions and their updating rules. We will begin the discussion with the following lemmas which were stated and proven by Pednault et al. (1981), and Johnson (1986), respectively.

Lemma 4.1. If Ω consists of n ($n > 2$) mutually exclusive and exhaustive propositions, i.e., if $\sum_{j=1}^{n} P(A_j) = 1$ and $P(A_i \,\&\, A_j) = 0$ for $i \neq j$, then equations (4.3.2) and (4.3.3) together imply equation (4.3.1).

When $n = 2$ ($\Omega = \{ A, \bar{A} \}$), the above lemma does not hold.

Lemma 4.2. If Ω consists of n mutually exclusive and exhaustive propositions, where $n > 2$, and if equations (4.3.2) and (4.3.3) are assumed, then there is at most one piece of evidence that produces updating for the proposition.

Lemma 4.2 says that under the above conditions regarding Ω, at most one piece of evidence can alter the probability of any given proposition; thus, although updating is possible, multiple updating for any of the propositions is impossible. The following lemma is from Cheng et al. (1986).

Lemma 4.3. Suppose that $\Omega = \{ A, \bar{A} \}$. If equations (4.3.1), (4.3.2), and (4.3.3) are assumed, then there is at most one piece of evidence that produces updating for each proposition.

As a consequence of the above lemmas, in order for probabilities of two or more mutually exclusive and exhaustive propositions to be updated and allow multiple pieces of evidence to influence a decision, one of the conditional independence assumptions should be eliminated. In fact, Charniak (1983) and Johnson recommend the updating scheme in eq. (4.3.5) for inference about any number of mutually exclusive and exhaustive propositions.

4.4. Dempster's Rule of Combination

Dempster's rule is a generalized scheme of Bayesian inference to aggregate bodies of evidence provided by multiple information sources. Let m_1 and m_2 be the basic probability assignments associated respectively with the belief functions $\mathcal{Bel}_1$ and $\mathcal{Bel}_2$ which are inferred from two entirely distinct bodies of evidence E_1 and E_2. For all A_i, B_j, and $X_k \subset \Omega$, Dempster's rule (or Dempster's orthogonal sum) gives a new belief function denoted by

$$\mathcal{Bel} = \mathcal{Bel}_1 \oplus \mathcal{Bel}_2 \tag{4.4.1}$$

The basic probability assignment associated with the new belief function is defined as:

$$m(X_k) = (1 - k)^{-1} \sum_{A_i \cap B_j = X_k} m_1(A_i) \cdot m_2(B_j) \quad (X_k \neq \varnothing) \tag{4.4.2}$$

where k is the measure of conflict between $\mathcal{Bel}_1$ and $\mathcal{Bel}_2$, as defined in Definition 3.4.

Dempster's rule computes the basic probability of X_k, $m(X_k)$, from the product of $m_1(A_i)$ and $m_2(B_j)$ by considering all A_i and B_j whose intersection is X_k. Once m is computed for every $X_k \subset \Omega$, the belief function is obtained by the sum of m's committed to X_k and its subsets. The denominator $(1-k)$ normalizes the result to compensate for the measure committed to the empty set so that the total probability mass has measure one. Consequently, Dempster's rule discards the conflict between E_1 and E_2 and carries their consensus to the new belief function.

There are several points of interest with regard to this rule. First, it requires that the basic probability assignments to be combined be based on entirely distinct bodies of evidence and refer to the same frame of discernment Ω. Secondly, it is both commutative and associative. Therefore, the order or grouping of evidence in combination does not affect the result, and a sequence of information sources can be combined either sequentially or pairwise. Finally, k in the above equation is the measure of conflict between E_1 and E_2, which represents the amount of the total probability that is committed to disjoint (or contradictory) subsets of Ω. If $\mathcal{k}$ is equal to one, this means that E_1 and E_2 are completely contradictory and the orthogonal sum of their basic probability assignments does not exist.

To exhibit the properties of Dempster's rule, suppose that there are only two focal elements A and $\bar{A}$ in Ω and the basic probability assignment m_i based on E_i is given as:

$$m_i(A) = p_i, \;\; m_i(\bar{A}) = q_i, \;\; m_i(\Omega) = 1-p_i-q_i \quad \text{for } i = 1, 2 \tag{4.4.3}$$

where $p_i + q_i \leq 1$, i.e., they are non-additive.

Then, the respective interval-valued belief function given $E_i(i=1,2)$ supports A with $[p_i, 1-q_i]$, and $\bar{A}$ with $[q_i, 1-p_i]$. Dempster's rule produces the new basic probability assignment m, and by equation (2.2.9) the support function for A and $\bar{A}$ based on the total evidence is given as:

$$\begin{aligned} \mathcal{S}p(A|E_1\&E_2) &= \frac{p_1 \cdot p_2 + p_1 \cdot (1-p_2-q_2) + (1-p_1-q_1) \cdot p_2}{1-p_1 \cdot q_2 - q_1 \cdot p_2} \\ &= 1 - \frac{(1-p_1) \cdot (1-p_2)}{1-p_1 \cdot q_2 - q_1 \cdot p_2} \end{aligned} \tag{4.4.4}$$

$$\begin{aligned} \mathcal{S}p(\bar{A}|E_1\&E_2) &= \frac{q_1 \cdot q_2 + q_1 \cdot (1-p_2-q_2) + (1-p_1-q_1) \cdot q_2}{1-p_1 \cdot q_2 - q_1 \cdot p_2} \\ &= 1 - \frac{(1-q_1) \cdot (1-q_2)}{1-p_1 \cdot q_2 - q_1 \cdot p_2} \end{aligned} \tag{4.4.5}$$

Figure 4.2 shows the graphical interpretation of Dempster's rule for the above case. The probability mass committed to Ω represents the uncertainty concerning the support for A and $\bar{A}$. The conjugate plausibility function $\mathcal{P}\ell$ is obtained by equation (2.2.14). In general, Dempster's rule has the following properties:

(1) Commutativity and associativity.

(2) $[\mathcal{S}p, \mathcal{P}\ell] \oplus [0, 1] = [\mathcal{S}p, \mathcal{P}\ell]$; $[0, 1]$ plays the role of identity for the rule.

	$m_2(A) = p_2$	$m_2(\bar{A}) = q_2$	$m_2(\Omega) = 1-p_2-q_2$
$m_1(A) = p_1$	$A \cap A = A$ $p_1 \cdot p_2$	$A \cap \bar{A} = \varnothing$ $p_1 \cdot q_2$	$A \cap \Omega = A$ $p_1 \cdot (1-p_2-q_2)$
$m_1(\bar{A}) = q_1$	$\bar{A} \cap A = \varnothing$ $q_1 \cdot p_2$	$\bar{A} \cap \bar{A} = \bar{A}$ $q_1 \cdot q_2$	$\bar{A} \cap \Omega = \bar{A}$ $q_1 \cdot (1-p_2-q_2)$
$m_1(\Omega) = 1-p_1-q_1$	$\Omega \cap A = A$ $(1-p_1-q_1) \cdot p_2$	$\Omega \cap \bar{A} = \bar{A}$ $(1-p_1-q_1) \cdot q_2$	$\Omega \cap \Omega = \Omega$ $(1-p_1-q_1) \cdot (1-p_2-q_2)$

Figure 4.2 Graphical Interpretation of Dempster's Rule when $\Omega = \{ A, \bar{A} \}$

(3) When $p_i+q_i = 1$, i.e., they are additive, equation (4.4.4) is equal to equation (4.3.8), and the resulting belief function becomes additive.

(4) For any interval $[\mathcal{S}p, \mathcal{P}l] \neq [0, 0]$, $[\mathcal{S}p, \mathcal{P}l] \oplus [1, 1] = [1, 1]$, and for any interval $[\mathcal{S}p, \mathcal{P}l] \neq [1, 1]$, $[\mathcal{S}p, \mathcal{P}l] \oplus [0, 0] = [0, 0]$; [0, 0] and [1, 1] are annihilators for the rule.

(5) $[0, 0] \oplus [1, 1]$ is undefined; Dempster's rule cannot combine completely conflicting bodies of evidence.

(6) The combined interval is no wider than any interval to be combined, i.e.,

$$\frac{(1-p_1-q_1)\cdot(1-p_2-q_2)}{1-p_1\cdot q_2-q_1\cdot p_2} \leq 1-p_i-q_i \qquad \text{for } i = 1, 2 \qquad (4.4.6)$$

Since the width of an interval-valued belief measure corresponds to the measure of uncertainty, it seems intuitively reasonable that the value of the measure of uncertainty decreases as the amount of evidential information increases.

The only condition that Dempster's rule requires is that the bodies of evidence to be combined must be entirely distinct. In the context of the problem of multisource data classification, combining entirely distinct bodies of evidence is considered as a fusion of the individual observations provided by independent sensors. The meaning of independence here is that an observation from one sensor does not have any effect on an observation from any other sensor.

4.5. Robustness of Combination Rules

The previous two sections described the functional characteristics of the subjective Bayesian updating rules and Dempster's rule in terms of the desirable properties of combination rules. In this section, the binary operators of Dempster's rule (eq. (4.4.4)) and a subjective Bayesian updating rule (eq. (4.3.8)) are compared with respect to their sensitivity to small changes of the initial belief measures to be combined.

Suppose we are classifying a pixel denoted by a vector **X** into one of a set of mutually exclusive and exhaustive classes, ω_1, ω_2, and ω_3, based on two independent data sources. Let E_1 and E_2 denote the bodies of evidence provided by the two data sources, and $\Omega = \{A_1, A_2, A_3\}$ denote the frame of discernment, where A_i represents the event of **X** being classified to ω_i. Suppose that the basic probability assignment numbers based on each data source are given as:

$$m_1(A_1) = \delta,\ \ m_1(A_2) = 1-\delta-\rho,\ \ m_1(A_3) = \rho \tag{4.5.1}$$

and

$$m_2(A_1) = 1-\delta-\rho,\ \ m_2(A_2) = \delta,\ \ m_2(A_3) = \rho \tag{4.5.2}$$

Note that the above measures are additive, i.e., there is no measure of uncertainty. Hence, both data sources are believed to be completely reliable, and the information provided by the data sources is assumed to be exact and precise for representing the belief measures.

When $\delta = 0$ and $0 < \rho << 1$, there is strong conflict between the bodies of evidence provided by the data sources. The only agreement between them is that A_3 is highly improbable. In other words, **X** is hardly believed to belong to ω_3. On the contrary, the equation (4.3.8) – recall that it is a special case of Dempster's rule when the belief measures are additive – yields the combined measures as:

$$m(A_1|\, E_1\&E_2) = m(A_2|\, E_1\&E_2) = 0,\ \ m(A_3|\, E_1\&E_2) = 1$$

without regard to the value of ρ. The result expresses that ω_3 is the only possible class for **X**, which is completely against our intuition.

Now, in order to examine how sensitive the combination rule is to slight changes of initial measures, let δ be a non-zero small number. Then, we find

$$m(A_1|\, E_1\&E_2) = m(A_2|\, E_1\&E_2) = \frac{\delta(1-\delta-\rho)}{2\delta(1-\delta-\rho) + \rho^2} \tag{4.5.3}$$

$$m(A_3|\, E_1\&E_2) = \frac{\rho}{2\delta(1-\delta-\rho) + \rho^2} \tag{4.5.4}$$

Table 4.1 shows the results of the equation (4.3.8) for various small values of δ when $\rho = 0.1$.

Table 4.1 Result of Combination by Dempster's Rule for Additive Belief Measures.

	$\rho = 0.1$		
	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.05$
$m(A_1 \mid E_1\&E_2)$	0.076	0.320	0.447
$m(A_2 \mid E_1\&E_2)$	0.076	0.320	0.447
$m(A_3 \mid E_1\&E_2)$	0.848	0.360	0.106

By comparing the combined measures for $\delta = 0.001$ and 0.05, we can draw a conclusion that the extreme sensitivity may lead to totally different decisions when the numerical representation of belief is coarse. Recall that the measures of belief in the above example are additive. Will Dempster's rule show such sensitivity when the measures of belief are subadditive?

When the data sources are not completely reliable, which is true in most cases of real world data sources, the belief measures based on the partially reliable sources include the measure of uncertainty. Suppose both data sources are assigned the same amount of measure of uncertainty α, that is,

$$m_1(\Omega) = m_2(\Omega) = \alpha$$

where $0 < \alpha < 1$. α is assigned to the frame of discernment Ω to represent the partial ignorance of belief based on the incomplete data sources. Then, the initial measures in (4.5.1) and (4.5.2) which were additive are reduced as:

$$m_1(A_1) = (1-\alpha)\delta,\;\; m_1(A_2) = (1-\alpha)(1-\delta-\rho),\;\; m_1(A_3) = (1-\alpha)\rho \qquad (4.5.5)$$

and

$$m_2(A_1) = (1-\alpha)(1-\delta-\rho),\;\; m_2(A_2) = (1-\alpha)\delta,\;\; m_2(A_3) = (1-\alpha)\rho \qquad (4.5.6)$$

Now, the belief measures become non-additive, and they are represented in terms of interval-valued probabilities in Table 4.2. In this particular case, since all the focal elements are singleton, the width of their IV belief measures is the same.

Table 4.2 Interval-valued Belief Measures after Combination by Dempster's Rule for Non-additive Belief Measures.

	E_1		E_2	
	$\mathcal{Sp}$	$\mathcal{Pl}$	$\mathcal{Sp}$	$\mathcal{Pl}$
A_1	$(1-\alpha)\delta$	$(1-\alpha)\delta+\alpha$	$(1-\alpha)(1-\delta-\rho)$	$(1-\alpha)(1-\delta-\rho)+\alpha$
A_2	$(1-\alpha)(1-\delta-\rho)$	$(1-\alpha)(1-\delta-\rho)+\alpha$	$(1-\alpha)\delta$	$(1-\alpha)\delta+\alpha$
A_3	$(1-\alpha)\rho$	$(1-\alpha)\rho+\alpha$	$(1-\alpha)\rho$	$(1-\alpha)\rho+\alpha$

Dempster's rule yields the new basic probability assignment as:

$$m(A_1|\, E_1\&E_2) = m(A_2|\, E_1\&E_2) = \frac{(1-\alpha)\{\delta(1-\alpha)(1-\delta-\rho)+\alpha(1-\rho)\}}{(1-k)} \tag{4.5.7}$$

$$m(A_3|\, E_1\&E_2) = \frac{r(1-\alpha)\{(1-\alpha)\rho+2\alpha\}}{(1-k)} \tag{4.5.8}$$

and

$$m(\Omega|\, E_1\&E_2) = \frac{\alpha^2}{(1-k)} \tag{4.5.9}$$

where $k = (1-\alpha)^2\{1-\rho+2\delta(\rho+\delta-1)\}$.

Let $\alpha = 0.1$, which means that the data sources are highly reliable but still incomplete. For $\delta = 0$ and $\rho = 0.1$, the combined measures are:

$$m(A_1|\, E_1\&E_2) = m(A_2|\, E_1\&E_2) = 0.409\,, \quad m(A_3|\, E_1\&E_2) = 0.132$$

Compared to those which are additive, the non-additive measures, after being combined by Dempster's rule, are more in accordance with human intuition. Table 4.3 shows the results of Dempster's rule combining non-additive measures for various small values of δ.

Table 4.3 Result of Combination by Dempster's Rule for Non-additive Belief Measures.

	$\alpha = 0.1 \quad \rho = 0.1$		
	$\delta = 0.001$	$\delta = 0.01$	$\delta = 0.05$
$m(A_1 \| E_1 \& E_2)$	0.409	0.414	0.432
$m(A_2 \| E_1 \& E_2)$	0.409	0.414	0.432
$m(A_3 \| E_1 \& E_2)$	0.131	0.123	0.098
$m(\Omega \| E_1 \& E_2)$	0.051	0.049	0.038

By assigning a small amount of uncertainty to the data sources, we can avoid the extreme sensitivity of Dempster's rule to slight changes of measures provided by conflicting bodies of evidence.

Since the problem of extreme sensitivity of Dempster's rule was exposed by Zadeh (1979), Dubois and Prade (1985) proposed as an alternative a possibilistic rule of combination based on the theory of possibility which is related to the fuzzy set theory. Zadeh and Dubois et al. insist that the extreme sensitivity of Dempster's rule in combining additive probabilities is the effect of the normalization in its denominator. They think that the normalization suppresses an important aspect of information obtained from the conflicting bodies of evidence, so that Dempster's rule may yield highly counterintuitive results. According to the above example, however, the cause of the extreme sensitivity lies in incorrect representation of belief, not in Dempster's rule itself. Recall that the frame of discernment consists of mutually exclusive and exhaustive hypotheses. If two sources were completely reliable, there might be little conflict between the bodies of evidence provided by them. Conversely, if there were strong conflict between bodies of evidence, the sources providing the evidence could not be completely reliable, either or both of them should have non-zero measure of uncertainty. In conclusion, interval-valued probabilities are more adequate than conventional additive probabilities to represent belief.

4.6. Summary

In this chapter, after defining desirable properties for combination rules to be formulated as functional equations, the inferencing mechanisms of subjective Bayesian updating rules and Dempster's rule were examined in terms of their properties. The comparison revealed that Dempster's rule is a more general scheme to combine bodies of evidence providing the belief functions represented by interval-valued probabilities. It has been observed that in combining conflicting bodies of evidence, Dempster's rule produces more robust and consistent combined belief measures when the belief measures are interval-valued.

In this chapter, the contributions of this research are the formal definitions of the desirable properties of combination rules, interpretations of the inferencing mechanisms of the existing combination rules, and the analysis of the robustness of Dempster's rule in the aspect of its differential behavior according to slight changes of initial belief measures.

CHAPTER 5

DECISION MAKING BASED ON INTERVAL-VALUED PROBABILITIES

5.1. Introduction

Making a decision is the last step before evaluating the performance of a classifier in any pattern recognition problem. Over the past three decades, statistical decision theory has played an important role in the decision process of statistical pattern recognition techniques.

In conventional statistical methods for pattern recognition where statistical information is represented by point-valued probabilities, there is only one decision rule to use in deciding whether or not a given pattern belongs to some prespecified class of patterns. The decision rule gives an estimate of the unknown, true class of the pattern, and the estimate varies depending on the criterion underlying the decision rule. For example, the "Bayes decision rule" is devised in such a way that the "average risk" is minimized. The Maximum Posterior classification, which is the most common classification method in remote sensing, uses a "Bayes decision rule with 0–1 loss function."

In the previous chapters, representation and combination of statistical evidence in the form of interval-valued probabilities were studied. Although interval-valued probabilities provide an innovative means for the representation of evidential information, they make the decision process rather complicated and entail more intelligent strategies in making decisions. Based on the evidential interval bounded by degrees of support and plausibility, one has more than one choice for a decision rule. One can make a decision either based on any one of support or plausibility, or based on their average.

This chapter presents an account of basic elements in the decision theory for pattern recognition based on interval-valued probabilities. It will be noticed that under a certain condition those basic elements are a generalization of the elements of Bayesian decision theory. This chapter also formalizes the decision-making process and develops decision rules for the evidential intervals.

5.2. Interval-Valued Expectations

Let $[\mathcal{L}, \mathcal{U}]$ be an interval-valued probability defined in the Boolean algebra $\mathcal{B}$ of subsets of Ω, and V denote a real-valued function defined over $\Omega = \{\omega\}$. Dempster(1968) defines an "upper distribution function" and a "lower distribution function" respectively as:

$$F^*(v) = \mathcal{U}(\{\omega \mid V(\omega) \le v\})$$

$$\text{for } -\infty < v < \infty \qquad (5.2.1)$$

$$F_*(v) = \mathcal{L}(\{\omega \mid V(\omega) \le v\})$$

The pair $[F^*, F_*]$ defined above has the following properties:

(i) Both are nondecreasing, i.e.,

$$\text{if} \quad v_1 < v_2 \quad \text{then} \quad F^*(v_1) \le F^*(v_2) \quad \text{and} \quad F_*(v_1) \le F_*(v_2) \qquad (5.2.2)$$

(ii) Both are continuous from the right, i.e.,

$$\text{For } \varepsilon > 0, \quad \lim_{\varepsilon \to 0} F^*(v+\varepsilon) = F^*(v) \quad \text{and} \quad \lim_{\varepsilon \to 0} F_*(v+\varepsilon) = F_*(v) \qquad (5.2.3)$$

(iii) $F^*(+\infty) = F_*(+\infty) = 1, \quad F^*(-\infty) = F_*(-\infty) = 0$ (5.2.4)

(iv) If $F^*(v_0) = 0$ $(F_*(v_0) = 0)$
then $F^*(v) = 0$ $(F_*(v) = 0)$ for every $v \le v_0$ (5.2.5)

(v) $F^*(v) \ge F_*(v)$ for $-\infty < v < \infty$ (5.2.6)

The proof of the above properties is trivial. (i) – (iv) are the same as the properties of the ordinary distribution function. Refer to Papoulis(1984) for their proofs. And (v) is a direct consequence of eq. (2.4.3).

Further, Dempster defines his "upper expectation" and "lower expectation" as:

$$E^*(V) = \int_{-\infty}^{\infty} v \, dF_*(v)$$

$$E_*(V) = \int_{-\infty}^{\infty} v \, dF^*(v) \tag{5.2.7}$$

Note that the upper and lower stars are interchanged. It is necessary in order to keep the relation $E^*(V) \geq E_*(V)$. For any real-valued functions V and W defined over Ω, E^* and E_* have the following properties:

(i) $E^*(V) \geq E_*(V)$ (5.2.8)

(ii) If $V(\omega) \geq W(\omega)$ for all $\omega \in \Omega$

$E^*(V) \geq E^*(W)$ and $E_*(V) \geq E_*(W)$ (5.2.9)

Dempster's upper and lower expectations generalize the concepts of upper and lower probabilities. Speaking in detail, let Z_A be the indicator function of $A \subset \Omega$, i.e.,

$$Z_A(\omega) = \begin{cases} 1 & \text{for } \omega \in A \\ 0 & \text{otherwise} \end{cases} \tag{5.2.10}$$

Then, by the above definitions and the conjugate relationship of $\mathcal{U}$ and $\mathcal{L}$

$$E^*(Z_A) = \int_{-\infty}^{+\infty} z \, dF_*(z) = \int_0^1 z \cdot \mathcal{L}(\{\omega | Z_A(\omega) \leq z\}) \, dz$$

$$= \mathcal{L}(\Omega) - \mathcal{L}(\bar{A}) = 1 - \mathcal{L}(\bar{A}) = \mathcal{U}(A)$$

$$E_*(Z_A) = \int_{-\infty}^{+\infty} z \, dF^*(z) = \int_0^1 z \cdot \mathcal{U}(\{\omega | Z_A(\omega) \leq z\}) \, dz \tag{5.2.11}$$

$$= \mathcal{U}(\Omega) - \mathcal{U}(\bar{A}) = 1 - \mathcal{U}(\bar{A}) = \mathcal{L}(A)$$

For pattern recognition problems, it seems natural to define upper and lower probabilities respectively by upper and lower envelopes, i.e., the supremum and the infimum of a certain class of probability measures as expressed in Definition 2.2. As mentioned earlier in section 2.4, the envelopes are a subclass of the axiomatically defined interval-valued probabilities. Also, if $\mathcal{L}$ is 2-monotone and $\mathcal{U}$ is 2-alternating, then they are envelopes.

Suppose that $\mathcal{L}$ and $\mathcal{U}$ are given as

$$\begin{aligned} \mathcal{L}(A) &= \inf\{\pi(A) : \pi \in \mathcal{P}\} \\ & \qquad\qquad\qquad \text{for } A \in \mathcal{B} \qquad (5.2.12) \\ \mathcal{U}(A) &= \sup\{\pi(A) : \pi \in \mathcal{P}\} \end{aligned}$$

where $\mathcal{P}$ is the class of the probability measures dominated by $\mathcal{U}$. Then, the following lemma is proved by Wolfenson and Fine (1982).

Lemma 5.1. For an interval-valued probability $[\mathcal{L}, \mathcal{U}]$, the upper and lower expectations can be given as:

$$\begin{aligned} E^*(V) &= \sup_{\pi \in \mathcal{P}} E_\pi(V) \\ & \qquad\qquad\qquad\qquad (5.2.13) \\ E_*(V) &= \inf_{\pi \in \mathcal{P}} E_\pi(V) \end{aligned}$$

iff $\mathcal{L}$ is 2-monotone and $\mathcal{U}$ is 2-alternating, where V is a real-valued function over Ω and $E_\pi(V)$ is the expected value of V with respect to the probability measure π.

The upper and lower expectations in (5.2.13) have the following properties as well as the properties in (5.2.8) and (5.2.9):

(iii) $E_*(V) \le E_\pi(V) \le E^*(V)$ for any $\pi \in \mathcal{P}$, (5.2.14)

(iv) For any nonnegative function W over Ω,

$$E^*(a+bW) = \begin{cases} a + bE^*(W) & \text{if } b \ge 0 \\ a + bE_*(W) & \text{if } b < 0 \end{cases} \qquad (5.2.15)$$

$$E_*(a+bW) = \begin{cases} a + bE_*(W) & \text{if } b \geq 0 \\ a + bE^*(W) & \text{if } b < 0 \end{cases} \tag{5.2.16}$$

where a and b are constants.

This section introduces two different definitions of the interval-valued expectations; one which applies to any system of interval-valued probabilities, and the other which applies only to a system of 2-monotone and 2-alternating interval-valued probabilities. In general, the two definitions do not coincide in a class of all sets of probability measures over $\mathcal{B}$. Dempster (1968) already argued that for a general convex set $\mathcal{P}$, it can happen that

$$\int_{-\infty}^{\infty} v \, dF_*(v) < \inf_{\pi \in \mathcal{P}} E_\pi(V) \tag{5.2.17}$$

The second definition is not only unapt to a general system of interval-valued probabilities but also computationally intractable. For the expectations in eq. (5.2.13) to be useful, an explicit expression of π in $\mathcal{P}$ must be available.

5.3. Decision Rules based on Interval-Valued Probability

Consider a basic classification problem where an arbitrary pattern $\mathbf{x} \in \mathcal{X}$ from an unknown class is assigned to one of n classes in Ω. Let $\lambda(\omega_i|\omega_j)$ be a measure of the "loss" incurred when the decision ω_i is made and the true pattern class is in fact ω_j, where i, j = 1, ..., n. Also, let $\hat{\omega}(\mathbf{x})$ denote a decision rule that tells which class to choose for every pattern $\mathbf{x}$. Using the upper and lower expectations in eq. (5.2.7), the "upper expected loss" and the "lower expected loss" of making a decision $\hat{\omega}(\mathbf{x})=\omega_i$ are obtained as:

$$\begin{aligned} \mathcal{L}^*_i(\mathbf{x}) &= \sum_{j=1}^{n} \lambda(\omega_i \mid \omega_j)\, \mathcal{U}_{\mathbf{x}}(\omega_j) \\ \mathcal{L}_{*i}(\mathbf{x}) &= \sum_{j=1}^{n} \lambda(\omega_i \mid \omega_j)\, \mathcal{L}_{\mathbf{x}}(\omega_j) \end{aligned} \tag{5.3.1}$$

where $\mathcal{U}_{\mathbf{x}}$ and $\mathcal{L}_{\mathbf{x}}$ are respectively the upper and the lower probabilities for $\mathbf{x}$ being actually from ω_j.

Based on the interval-valued expected losses, the most desirable decision rule is the one which has the upper expected loss less than the lower expected losses of the others, i.e.,

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \ell^*_i(\mathbf{x}) \leq \ell_{*j}(\mathbf{x}) \quad \text{for } j=1,\ldots, n \tag{5.3.2}$$

This rule is called an "absolute rule."

The "Bayes-like rule" is the one which minimizes both the upper and the lower expected losses, i.e.,

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \ell^*_i(\mathbf{x}) \leq \ell^*_j(\mathbf{x}) \quad \text{and} \quad \ell_{*i}(\mathbf{x}) \leq \ell_{*j}(\mathbf{x}) \quad \text{for } j=1,\ldots, n \tag{5.3.3}$$

In particular, when λ is the "0–1 loss function", i.e.,

$$\lambda(\hat{\omega}(\mathbf{x})|\omega_j) = \begin{cases} 0 & \text{if } \hat{\omega}(\mathbf{x}) = \omega_j \\ 1 & \text{if } \hat{\omega}(\mathbf{x}) \neq \omega_j \end{cases} \tag{5.3.4}$$

the interval-valued expected loss in eq. (5.3.1) is simplified as:

$$\begin{aligned} \ell^*_i(\mathbf{x}) &= \sum_{j=1}^{n} \mathcal{U}_{\mathbf{x}}(\omega_j) - \mathcal{U}_{\mathbf{x}}(\omega_i) \\ \ell_{*i}(\mathbf{x}) &= \sum_{j=1}^{n} \mathcal{L}_{\mathbf{x}}(\omega_j) - \mathcal{L}_{\mathbf{x}}(\omega_i) \end{aligned} \tag{5.3.5}$$

Since the first terms in the right-hand sides are constant for i=1, ..., n, minimizing both $\ell^*_i(\mathbf{x})$ and $\ell_{*i}(\mathbf{x})$ corresponds to maximizing $\mathcal{U}_{\mathbf{x}}(\omega_i)$ and $\mathcal{L}_{\mathbf{x}}(\omega_i)$. Hence, the decision rule in eq. (5.3.2) becomes

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \mathcal{U}_{\mathbf{x}}(\omega_i) \geq \mathcal{U}_{\mathbf{x}}(\omega_j) \quad \text{and} \quad \mathcal{L}_{\mathbf{x}}(\omega_i) \geq \mathcal{L}_{\mathbf{x}}(\omega_j) \quad \text{for } j=1,\ldots, n \tag{5.3.6}$$

A problem with the above decision rules is that there does not always exist ω which satisfies the condition in eq. (5.3.2) or (5.3.3), which can lead to ambiguity. In comparing a pair of the interval-valued expected losses, there are three different kinds of relationships distinguished by their relative locations:

(1) **disjoint intervals**;

$$\ell^*_i(\mathbf{x}) \geq \ell_{*i}(\mathbf{x}) > \ell^*_j(\mathbf{x}) \geq \ell_{*j}(\mathbf{x}) \qquad (5.3.7)$$

(2) **overlapped intervals**;

$$\ell^*_i(\mathbf{x}) \geq \ell^*_j(\mathbf{x}) \geq \ell_{*i}(\mathbf{x}) \geq \ell_{*j}(\mathbf{x}) \qquad (5.3.8)$$

(3) **nested intervals**;

$$\ell^*_i(\mathbf{x}) \geq \ell^*_j(\mathbf{x}) > \ell_{*j}(\mathbf{x}) \geq \ell_{*i}(\mathbf{x}) \qquad (5.3.9)$$

The following example illustrates these intervals.

Example 5.1. Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. $[\mathcal{L}_\mathbf{x}, \mathcal{U}_\mathbf{x}]$ denotes the interval-valued probability function of subsets of Ω given a pattern $\mathbf{x}$. Suppose that the basic probability assignment $m_\mathbf{x}$ of $[\mathcal{L}_\mathbf{x}, \mathcal{U}_\mathbf{x}]$ is given as

$$m_\mathbf{x}(\{\omega_1\}) = 0.2 \quad m_\mathbf{x}(\{\omega_2\}) = 0.3 \quad m_\mathbf{x}(\{\omega_1, \omega_3\}) = 0.34 \quad m_\mathbf{x}(\{\omega_2, \omega_4\}) = 0.16$$

and $m_\mathbf{x}(\mathrm{A}) = 0$ for any other subsets A of Ω. Then, the interval-valued probabilities of the singletons are obtained as

	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_3\}$	$\{\omega_4\}$
$\mathcal{L}_\mathbf{x}$	0.2	0.3	0	0
$\mathcal{U}_\mathbf{x}$	0.54	0.46	0.34	0.16

For the 0–1 loss function, the expected loss interval of ω_2 is nested in ω_1's, ω_1 is overlapped with ω_3, and ω_4 is disjoint with respect to ω_1 and ω_2. The Bayes-like rule does not produce a decision.

The above example shows a simple case where the Bayes-like decision rule leads to ambiguity. In such an ambiguous situation, one may withhold the decision and wait for a new piece of information. Otherwise, the ambiguity may be resolved by resorting to the following rule, so-called "minimum average expected loss rule":

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \frac{\mathcal{L}^*_i(\mathbf{x}) + \mathcal{L}_{*i}(\mathbf{x})}{2} \leq \frac{\mathcal{L}^*_j(\mathbf{x}) + \mathcal{L}_{*j}(\mathbf{x})}{2} \quad \text{for } j=1,\ldots,n \tag{5.3.10}$$

For the 0–1 loss function, this rule is called "maximum average probability rule", and the decision is made according to

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \frac{\mathcal{U}_{\mathbf{x}}(\omega_i) + \mathcal{L}_{\mathbf{x}}(\omega_i)}{2} \geq \frac{\mathcal{U}_{\mathbf{x}}(\omega_j) + \mathcal{L}_{\mathbf{x}}(\omega_j)}{2} \quad \text{for } j=1,\ldots,n \tag{5.3.11}$$

As an alternative to the absolute rule and the Bayes-like rule, there are two other rules by which a decision is made according to individual measures of the interval, for instance, either the upper expected loss or the lower expected loss:

(1) **minimum upper expected loss rule**:

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \mathcal{L}^*_i(\mathbf{x}) \leq \mathcal{L}^*_j(\mathbf{x}) \quad \text{for } j=1,\ldots,n \tag{5.3.12}$$

For the 0–1 loss function, this rule may be renamed "maximum upper probability rule" or "maximum plausibility rule", and the decision is made according to

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \mathcal{U}_{\mathbf{x}}(\omega_i) \geq \mathcal{U}_{\mathbf{x}}(\omega_j) \quad \text{for } j=1,\ldots,n \tag{5.3.13}$$

(2) **minimum lower expected loss rule**:

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \mathcal{L}_{*i}(\mathbf{x}) \leq \mathcal{L}_{*j}(\mathbf{x}) \quad \text{for } j=1,\ldots,n \tag{5.3.14}$$

For the 0–1 loss function, this rule is called "maximum lower probability rule" or "maximum support rule", and the decision is made according to

$$\hat{\omega}(\mathbf{x}) = \omega_i \quad \text{if} \quad \mathcal{L}_{\mathbf{x}}(\omega_i) \geq \mathcal{L}_{\mathbf{x}}(\omega_j) \quad \text{for } j=1,\ldots,n \tag{5.3.15}$$

Although the above two rules always produce decisions and there is no ambiguous situation in making a decision according to the rules, they do not utilize all of the information represented by the IV probabilities. The performance of these rules will be compared with the minimum average expected loss rule in the next chapter by applying them to problems of ground-cover classification based on remotely sensed and geographic data.

5.4. Summary

The purpose of this chapter was to formalize the decision-making process for any system of interval-valued probabilities. In particular, the process was considered from the viewpoint of statistical decision theory.

First, two different definitions of interval-valued expectations were studied, and their statistical properties were compared with those of the ordinary expected value. Then the absolute rule and the Bayes-like rule for evidential intervals were developed based on the general interval-valued expectation. Since these rules are not always satisfied, they may require an extra step to resolve ambiguous situations. In order to resolve the ambiguous situations, this chapter proposed the minimum average expected loss rule. As alternatives to the absolute rule and the Bayes-like rule, the minimum upper expected loss rule and the minimum lower expected loss rule were proposed.

While the absolute rule and the Bayes-like rule make decisions based on both the upper and the lower expected losses, the minimum upper expected loss rule and the minimum lower expected loss rule make decisions based on either the upper or the lower expected loss. In the evidential reasoning, the lower probability and the upper probability represent respectively the minimal and the maximal degree of belief. Hence, the minimum lower expected loss rule may be chosen when the decision process needs to be conservative; and the minimum upper expected loss rule may be chosen when the decision maker is confident about the information represented by IV probabilities.

In this chapter, the contribution of the research is in the formal development of the decision-making process and the decision rules for interval-valued probabilities.

CHAPTER 6
EXPERIMENTAL RESULTS

6.1. Introduction

In this chapter, the methods presented in this report are applied to problems of ground-cover classification for multispectral data combined with other geographic data. The multisource data (MSD) classification based on the evidential reasoning (ER) method is implemented as the following procedure:

In the training stage,

1. Compute the global correlation coefficient matrix of multisource data and reform the data set if necessary. Throughout the experiments, the global correlation information will be used to confirm the "distinctness" of bodies of evidence as required by Dempster's rule.
2. For each class, select training pixels and compute statistics for each source.
3. Compute the separability measures of each source and the average measures of conflict between pairs of the sources as defined in Section 3.4. Rank the data sources and assign a degree of reliability to each source.

The steps in the test stage classifying "unknown" pixels will be described by considering an actual problem of classifying a test pixel to one of the classes in $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ based on two data sources denoted by S_1 and S_2.

$\mathbf{x}_i$: Test vector representing the test pixel obtained from S_i (i=1, 2).

α_i : Source reliability of S_i, $0 \leq \alpha_i \leq 1$.

$p_{\omega_j}(\mathbf{x}_i)$: Conditional probability density of $\mathbf{x}_i$ given ω_j.

m_i : Basic probability assignment based on S_i.

m : Basic probability assignment based on S_1 and S_2.

$\mathcal{Sp}$: Support function based on S_1 and S_2.

$\mathcal{Pl}$: Plausibility function based on S_i and S_2.

Suppose that $p_{\omega_j}(\mathbf{x}_i)$ for i=1, 2 and j=1, ..., 4 are obtained such that

$$p_{\omega_1}(\mathbf{x}_1) \geq p_{\omega_2}(\mathbf{x}_1) \geq p_{\omega_3}(\mathbf{x}_1) \geq p_{\omega_4}(\mathbf{x}_1)$$

$$p_{\omega_2}(\mathbf{x}_2) \geq p_{\omega_3}(\mathbf{x}_2) \geq p_{\omega_1}(\mathbf{x}_2) \geq p_{\omega_4}(\mathbf{x}_2)$$

(A) Using the consonant belief functions:

The focal elements based on S_1 are $\{\omega_1\}$, $\{\omega_1, \omega_2\}$, $\{\omega_1, \omega_2, \omega_3\}$, and Ω.

The focal elements based on S_2 are $\{\omega_2\}$, $\{\omega_2, \omega_3\}$, $\{\omega_2, \omega_3, \omega_1\}$, and Ω.

1. Compute $m_1(A)$ and $m_2(B)$ by using eq. (3.3.9), where A and B denote the focal elements of S_1 and S_2, respectively.
2. Multiply m_i by α_i for the subsets of Ω, and add α_i to $m_i(\Omega)$.
3. Compute $m = m_1 \oplus m_2$ by using eq. (4.4.2).
4 For each singleton ω_i, compute

$$\mathcal{Sp}(\{\omega_i\}) = m(\{\omega_i\}) \quad \text{and} \quad \mathcal{Pl}(\{\omega_i\}) = \sum_{A \cap \{\omega_i\} \neq \emptyset} m(A)$$

5 Classify the test pixel to a class according to one of the decision rules for IV probabilities in Chapter 5.

(B) Using the partially consonant belief functions:

Based on the relation in the hierarchical structure of the classes, suppose that Ω has a partition $\{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$.

The focal elements based on S_1 are $\{\omega_1\}$, $\{\omega_1, \omega_2\}$, $\{\omega_3\}$, and $\{\omega_3, \omega_4\}$.

The focal elements based on S_2 are $\{\omega_2\}$, $\{\omega_2, \omega_1\}$, $\{\omega_3\}$, and $\{\omega_3, \omega_4\}$.

1. Compute $m_1(A)$ and $m_2(B)$ by using eq. (3.3.10) and (3.3.11), where A and B denote the focal elements of S_1 and S_2, respectively.
2. Multiply m_i by α_i for the subsets of Ω, and add α_i to $m_i(\Omega)$.
3. Compute $m = m_1 \oplus m_2$ by using eq. (4.4.2).

4 For each singleton ω_i, compute

$$\mathcal{S}p(\{\omega_i\}) = m(\{\omega_i\}) \quad \text{and} \quad \mathcal{P}l(\{\omega_i\}) = \sum_{A \cap \{\omega_i\} \neq \emptyset} m(A)$$

5 Classify the test pixel to a class according to one of the decision rules for IV probabilities in Chapter 5.

Figure 6.1 is the block diagram of for classifying a pixel in the MSD classification based on the ER method.

The experiments have been performed over three different image data sets. Table 6.1 shows the names and types of data sources of the multisource data sets. More detailed descriptions will be given in the following sections. Each data set also has a geometrically registered, digitized ground truth map as a reference based on which the accuracies of all subsequent classifications will be evaluated.

The next section presents the experimental results of the proposed method applied to the Anderson River data set. The intention of the experiment is to assess the ability of the method in capturing and utilizing the information obtained from topographic data sources as well as multispectral data sources. In Section 6.3, the method is applied to the Indiana agricultural area data set which contains only a single multispectral data source. The purpose is to show the possibility that the MSD classification based on the evidential reasoning method can overcome the effects of the Hughes phenomenon [Hughes (1968)] which results in lowered classification accuracy for high-dimensional data with limited number of training samples. The goal is to show that improved classification can be obtained by decomposing a high-dimensional data source into smaller and more manageable pieces and treating them as multiple data sources. The possibility becomes more concrete in Section 6.4 where the method is applied to a simulated High Resolution Imaging Spectrometer (HIRIS) data set which is composed of 201 bands.

In every application, the classification accuracies of the MSD classification are compared with those of Maximum Likelihood (ML) classifications based on the stacked vector approach. Since the stacked vector approach treats

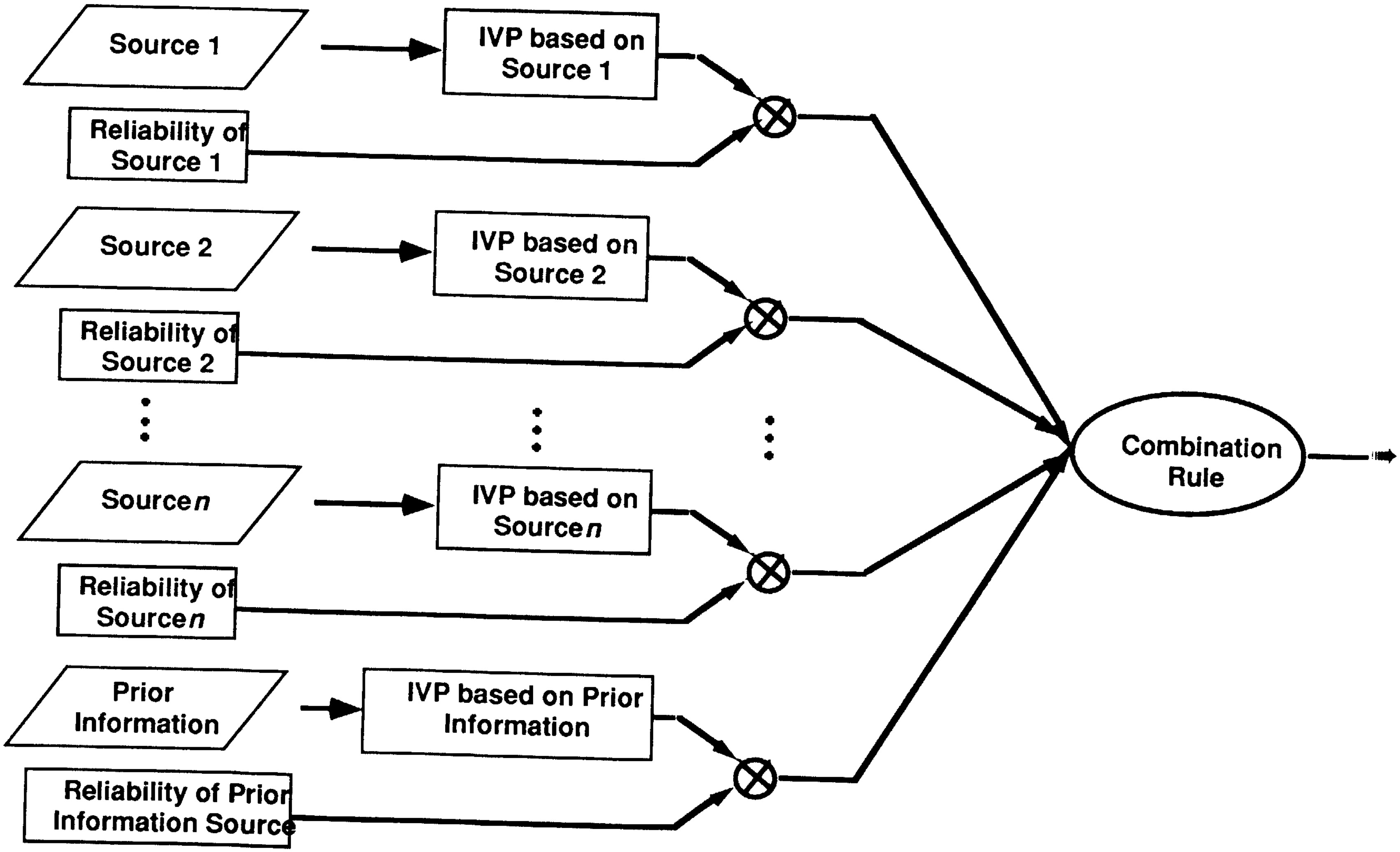


Figure 6.1 Block Diagram of Evidential Reasoning Method for Multisource Data Classification.

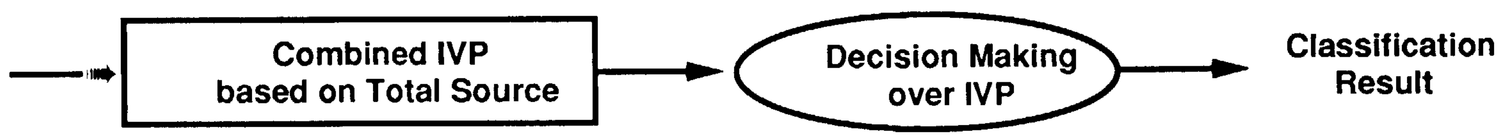


Figure 6.1, Continued.

compound vectors as data from a single source, the comparison of the MSD and the ML classifications will assess the advantages and the disadvantages of the multisource data analysis approach compared to a standard single source analysis approach used in remote sensing.

Table 6.1 Multisource Data Sets.

Name	Types of Data Sources
Anderson River Data	Airborne MSS, SAR, Elevation, Slope, Aspect
Indiana Agricultural Area Data	Airborne MSS
Finney County Data	HIRIS

6.2. Classification of Multispectral Data combined with Topographic Data

The Anderson River data set* used in the first experiment consists of 3 multispectral data sources (optical and radar) and 3 topographic data sources. Table 6.2 describes the types of data sources for the first experiment. The image of this data set consists of 256 lines by 256 columns and covers a forestry site around the Anderson River area in British Columbia, Canada. Source 1 is 11-band Airborne Multispectral Scanner data (A/B MSS). Sources 2 and 3 are Synthetic Aperture Radar (SAR) imagery in Shallow mode and Steep mode, respectively. The column "spectral band" for sources 2 and 3 describes the band and the transmit and receive type of SAR images. For example, XHV means that the image is obtained in X-band (λ=3cm) of the microwave region by horizontal polarization transmit and vertical polarization receive. Sources 4 – 6 provide digital terrain data obtained as follows:

---

* The SAR/MSS Anderson River data set was acquired, processed and loaned to Purdue University by the Canadian Center for Remote Sensing, Department of Energy, Mines and Resources, of the Government of Canada.

Table 6.2 Description of Anderson River Data Set.

Source Index	Data Type	Spectral Region	Input Channel	Spectral Band(μm)
1	A/B MSS	Visible	1	.38 - .42
			2	.42 - .45
			3	.45 - .50
			4	.50 - .55
			5	.55 - .60
			6	.60 - .65
			7	.65 - .69
		Near IR	8	.70 - .79
			9	.80 - .89
			10	.92 - 1.10
		Thermal	11	8 - 14
2	SAR	Shallow		XHV
				XHH
				LHV
				LHH
3	SAR	Steep		XHV
				XHH
				LHV
				LHH
4	Topo-graphic	Elevation		
5		Aspect		
6		Slope		

a) digital elevation model (DEM)

$$\text{gray level} = \{\text{elevation (in meters)} - 61.996\} \div 7.2266$$

b) digital aspect model (DAM)

$$\text{gray level} = \text{aspect (in degrees)} \div 2$$

c) digital slope model (DSM)

$$\text{gray level} = \text{slope (in degrees)}$$

Table 6.3 lists the information classes in the area, and Figure 6.2 shows the ground truth map. More than three quarters of the area is covered by mixed forestry. The information classes were defined based on a forestry map, and it has been observed that some of the classes are very difficult to classify accurately. In this experiment, 6 of the more separable classes were selected, and these are listed in Table 6.4. Figure 6.3 displays the test areas of the 6 classes over the enhanced A/B MSS image. Some of the field labels are not readable. However, they can be confirmed by the ground truth map in Figure 6.2. Figures 6.4 and 6.5 are Synthetic Aperture Radar imagery respectively in Shallow and Steep mode, and Figures 6.6 through 6.8 are the digital terrain imagery of the data set.

Table 6.5 is the global statistical correlation coefficient matrix among the data sources. Correlation coefficients between pairs of variables from different sources are generally quite low compared to those from the same source. When the data can be assumed to be normally distributed, their uncorrelatedness implies statistical independence. In the experiments, we treat the data sources (including the topographic data sources) which have relatively low correlation as "globally independent" in order to assume that they reasonably closely satisfy the "distinctness" of bodies of evidence required by Dempster's rule.

In the experiment with the Anderson River data set, 100 pixels per class were used for training data, which is between 4% and 8% of the total pixels of the classes in the test fields. The training samples are uniformly distributed over the test fields so that they may be considered as good representatives of the

Table 6.3 Information Classes in Anderson River Data Set.

Class Index	Cover Types	Tree Sizes	No. of Pixels	% of Total
1	Douglas Fir (DF) 1	> 40m	1946	2.97
2	DF 2	31 - 40m	13158	20.08
3	DF 3	21 - 30m	6576	10.03
4	DF 4	10 - 20m	1045	1.59
5	Bare Soil, Slides		110	0.17
6	DF+Other Species 1	> 40m	1973	3.01
7	DF+Other Species 2	31 - 40m	5761	8.79
8	DF+Other Species 3	21 - 30m	1309	2.00
9	DF+Lodgepole Pine 1	31 - 40m	510	0.78
10	DF+Lodgepole Pine 2	21 30m	5636	8.60
11	DF+Cedar 1	> 40m	2483	3.79
12	DF+Cedar 2	31 - 40m	2895	4.42
13	Lodgepole Pine	10 - 20m	113	0.17
14	Hemlock+Cedar	31 - 40m	3173	4.84
15	DF+Hemlock	31 - 40m	2961	4.52
16	Hemlock+DF 1	31 - 40m	825	1.26
17	Hemlock+DF 2	21 - 30m	456	0.70
18	Rock, Talus		1982	3.02
19	Forest Clearings		12624	19.26
Total			65536	100.0

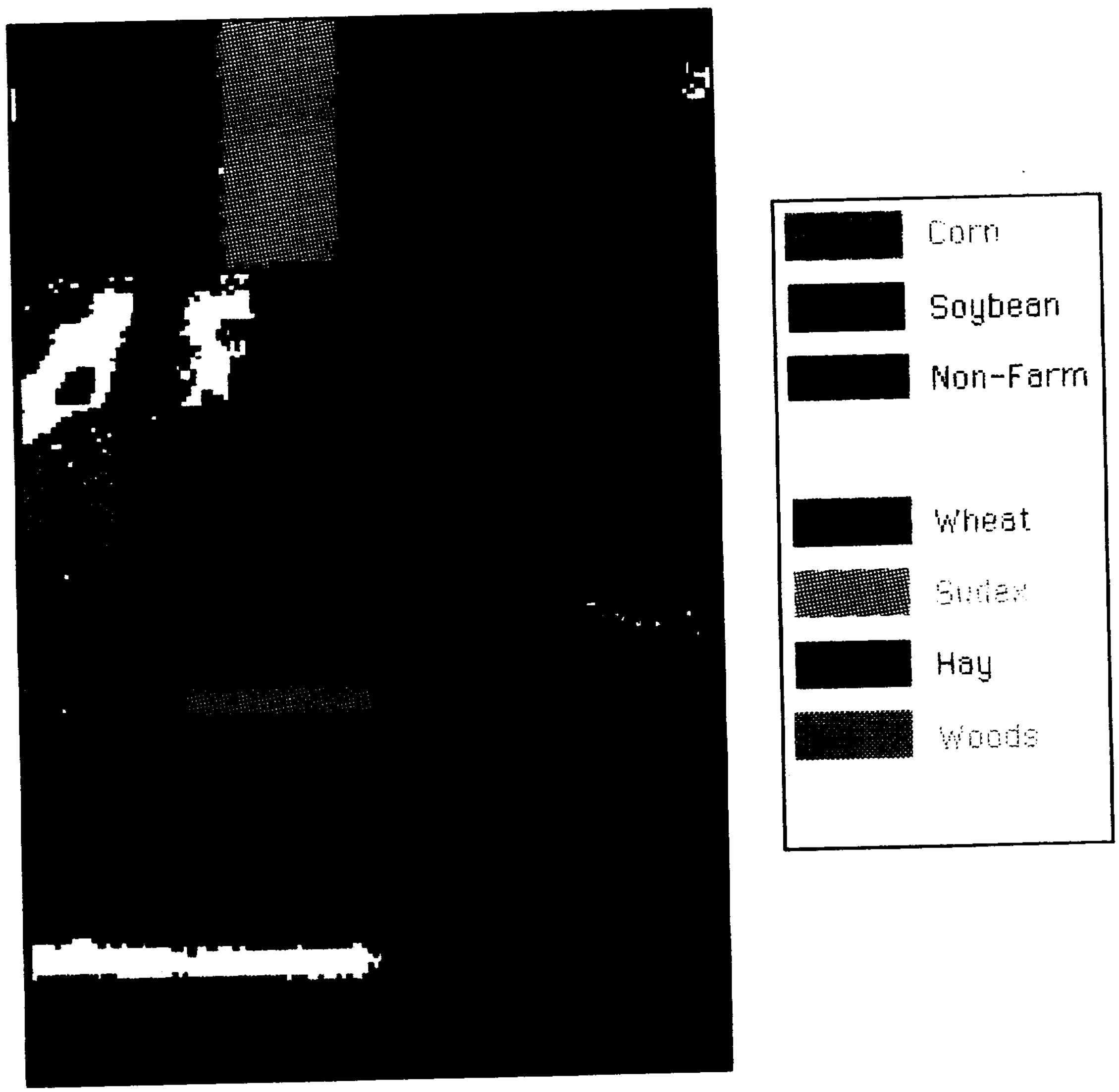


Figure 6.13 Ground Truth Map of Indiana Agricultural Area Data Set.

Table 6.4 Information Classes for Test of Anderson River Data Set.

Class Index	Cover Types	Tree Sizes	No. of Pixels	% of Total
2	Douglas Fir 2 (df2)	31 - 40m	2246	21.72
3	Douglas Fir 3 (df3)	21 - 30m	1501	14.52
7	DF+Other Species 2 (df+os2)	31 - 40m	1352	13.08
10	DF+Lodgepole Pine 2 (df+lp2)	21 - 30m	1589	15.37
14	Hemlock+Cedar (hc)	31 - 40m	1587	15.35
19	Forest Clearings (fc)		2064	19.96
Total			10339	100.0

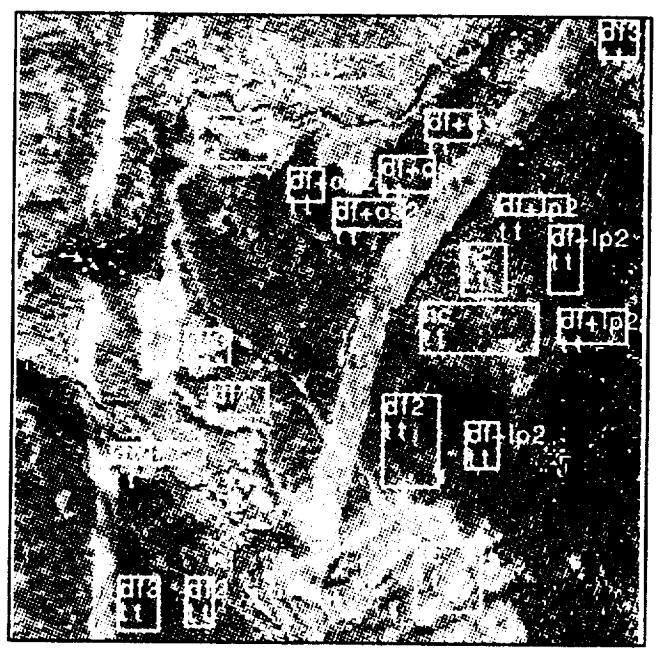
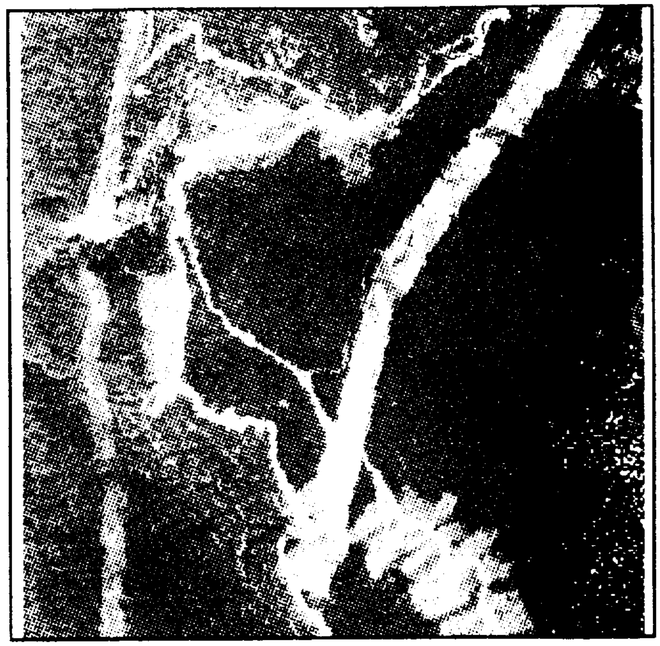


Figure 6.3 Test Areas over Histogram Equalized A/B MSS (Channel 10) Image of Anderson River Data Set.



Figures 6.4 Histogram Equalized SAR-Shallow mode (LHH) Image of Anderson River Data Set

Figures 6.5 Histogram Equalized SAR-Steep mode (LHH) Image of Anderson River Data Set

Figure 6.6 Digital Elevation Image of Anderson River Data Set.

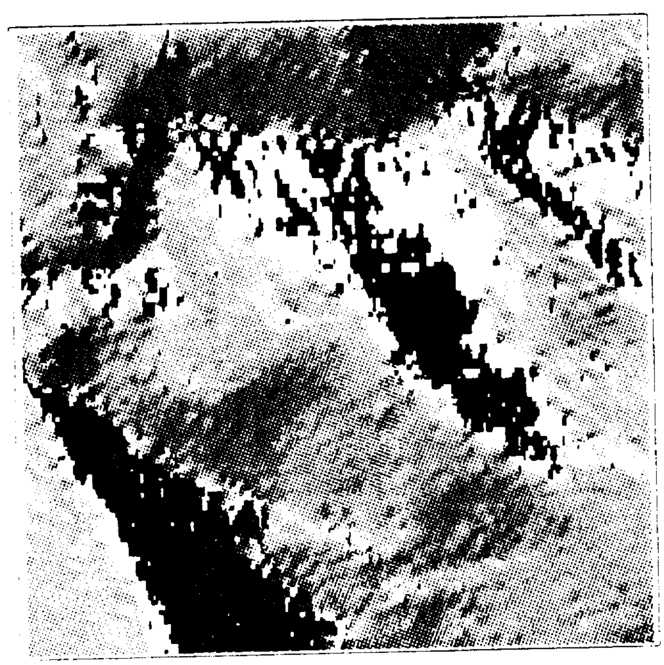


Figure 6.7 Digital Aspect Image of Anderson River Data Set.

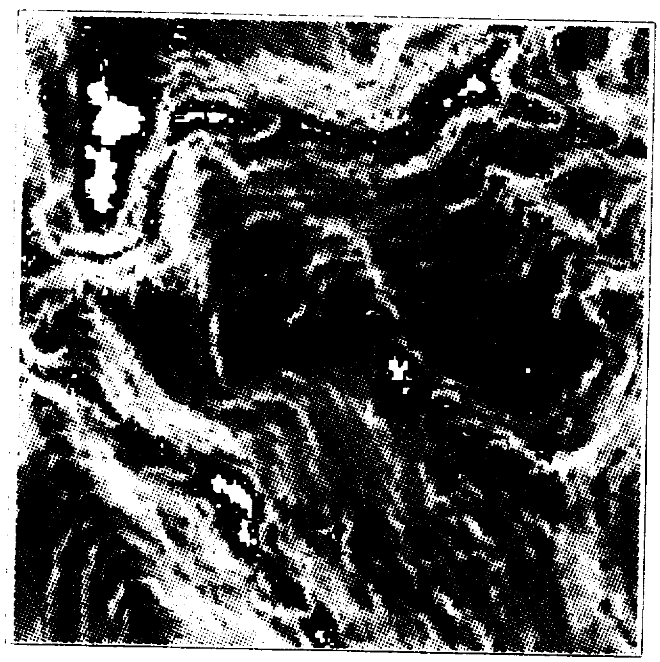


Figure 6.8 Digital Slope Image of Anderson River Data Set.

Table 6.5 Statistical Correlation Coefficient Matrix of Anderson River Data Set.

		A/B MSS										
		1	2	3	4	5	6	7	8	9	10	11
A/B MSS	1	1.000	0.815	0.753	0.709	0.670	0.633	0.626	0.573	0.459	0.520	0.593
	2		1.000	0.956	0.933	0.905	0.882	0.875	0.686	0.505	0.563	0.747
	3			1.000	0.975	0.961	0.955	0.951	0.677	0.465	0.516	0.792
	4				1.000	0.996	0.984	0.981	0.744	0.530	0.570	0.765
	5					1.000	0.992	0.990	0.742	0.526	0.562	0.761
	6						1.000	0.998	0.672	0.442	0.477	0.760
	7							1.000	0.684	0.454	0.490	0.773
	8								1.000	0.926	0.956	0.617
	9									1.000	0.959	0.464
	10										1.000	0.532
	11											1.000

Table 6.5, Continued.

		SAR SHALLOW				SAR STEEP				TOPOGRAPHIC		
		LHH	LHV	XHH	XHV	LHH	LHV	XHH	XHV	Aspect	Eleva	Slope
SAR SHAL	LHH	1.000	0.323	0.447	0.316	0.086	0.097	0.147	0.143	0.114	-.027	-.006
	LHV		1.000	0.312	0.426	0.161	0.164	0.187	0.208	0.106	-.033	0.027
	XHH			1.000	0.326	0.007	0.085	0.105	0.104	0.033	-.177	0.022
	XHV				1.000	0.161	0.166	0.201	0.216	0.082	-.062	0.046
SAR STEEP	LHH					1.000	0.348	0.472	0.378	0.094	0.101	0.131
	LHV						1.000	0.338	0.558	0.150	-.054	0.064
	XHH							1.000	0.391	0.139	0.131	0.124
	XHV								1.000	0.175	0.027	0.072
TOPO	Aspect									1.000	0.127	-.117
	Eleva										1.000	-.023
	Slope											1.000

Table 6.5, Continued.

		SAR SHALLOW				SAR STEEP				TOPOGRAPHIC		
		LHH	LHV	XHH	XHV	LHH	LHV	XHH	XHV	Aspect	Eleva	Slope
A/B MSS	1	0.074	0.094	0.102	0.088	-.123	0.008	-.193	-.035	-.076	-.589	-.039
	2	0.082	0.105	0.107	0.097	-.117	0.041	-.190	-.005	-.063	-.546	-.055
	3	0.075	0.103	0.088	0.087	-.099	0.061	-.169	0.017	-.041	-.424	-.061
	4	0.074	0.102	0.082	0.087	-.081	0.076	-.140	0.038	-.031	-.333	-.071
	5	0.069	0.099	0.070	0.082	-.072	0.081	-.128	0.045	-.024	-.271	-.074
	6	0.060	0.089	0.052	0.070	-.065	0.078	-.122	0.044	-.013	-.217	-.066
	7	0.062	0.093	0.051	0.073	-.065	0.079	-.121	0.047	-.009	-.205	-.067
	8	0.103	0.147	0.127	0.139	-.074	0.096	-.101	0.074	-.034	-.327	-.107
	9	0.099	0.145	0.135	0.141	-.066	0.086	-.079	0.069	-.036	-.320	-.100
	10	0.108	0.158	0.136	0.154	-.076	0.083	-.100	0.068	-.042	-.365	-.106
	11	0.092	0.131	0.089	0.110	-.084	0.047	-.152	0.014	-.072	-.341	-.066

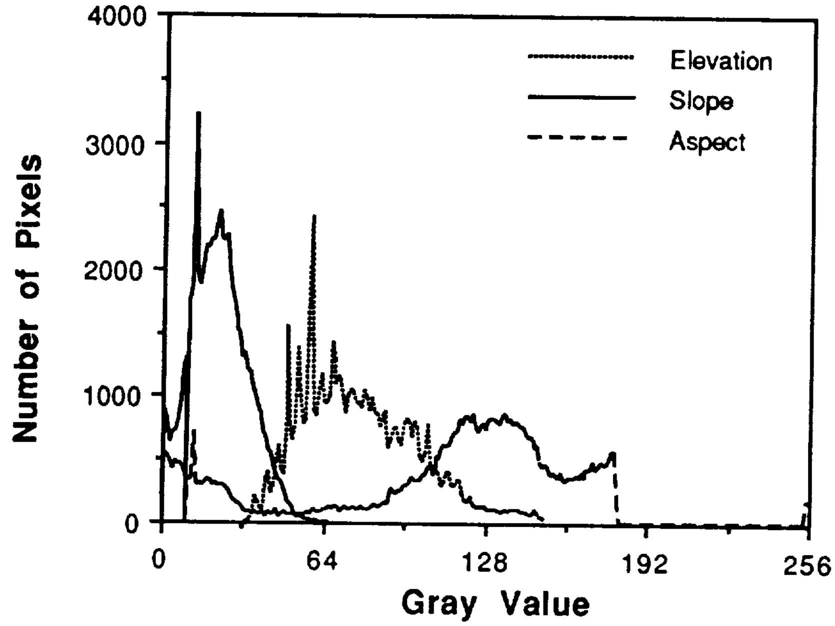


Figure 6.9 Histogram of Anderson River Topographic Data (Total Area).

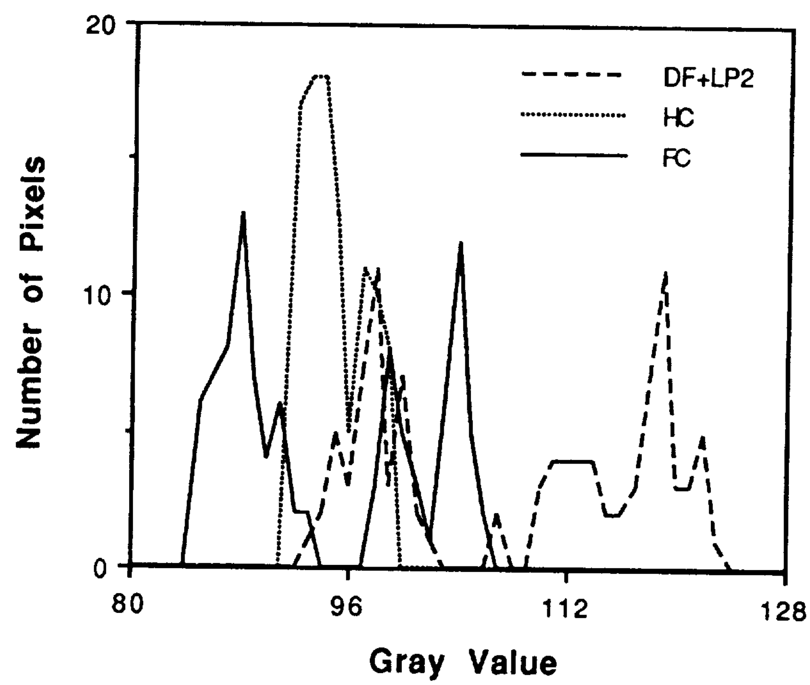


Figure 6.10 Classwise Histogram of Training Samples of a Subset of the Classes in theAnderson River Elevation Data.

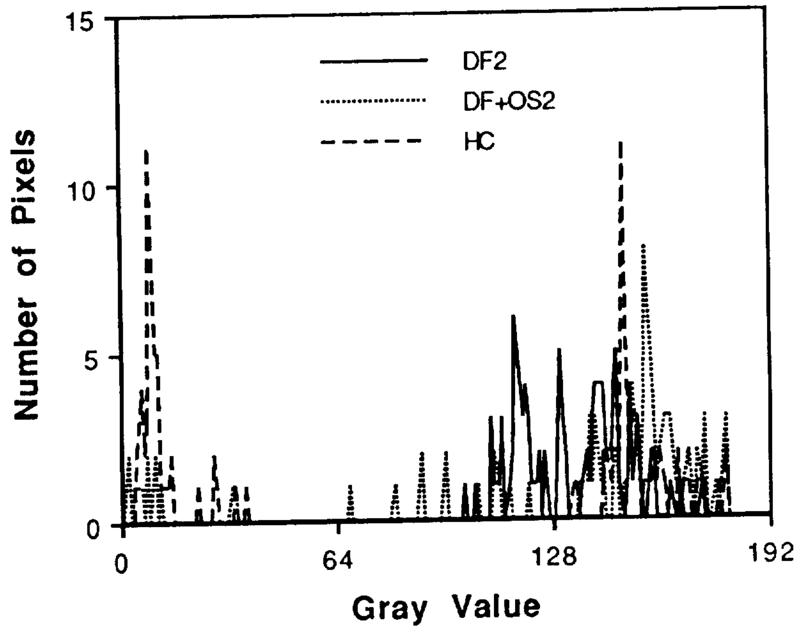


Figure 6.11 Classwise Histogram of Training Samples of a Subset of the Classes in the Anderson River Aspect Data.

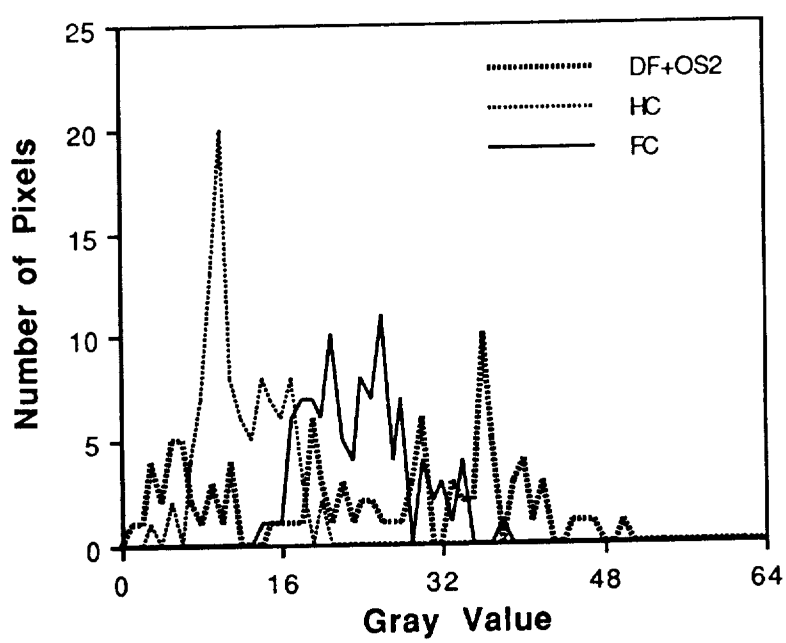


Figure 6.12 Classwise Histogram of Training Samples of a Subset of the Classes in the Anderson River Slope Data.

total samples. As we can observe in Figures 6.9 through 6.12, some of the classes defined in Table 6.4 cannot be assumed to be normally distributed in the topographic data. Thus, it was decided to adopt a nonparametric approach such as the "Nearest Neighbor" (NN) method [Fukunaga (1972)] in computing probability measures while the optical and radar data sources were assumed to have Gaussian probability density functions. Table 6.6 compares the overall classification accuracies obtained by the ML method with the Gaussian assumption and k-NN method for the individual topographic data sources. The results show that the topographic data are information-bearing in the sense of classification and suggest that the topographic data sources, especially Elevation, should be included in the classification. Although the k-NN method results in various classification accuracies for different k's, it always gives higher accuracies than the ML method especially for the training data. In the MSD classification, interval-valued belief functions for the bodies of statistical evidence provided by these topographic data sources were constructed from the likelihood functions obtained by the 2-NN method.

Table 6.6 Overall Classification Accuracy (%) obtained by ML Method and k-NN Method for Topographic Data Sources.

Samples	Method	Elevation	Aspect	Slope
Training	ML	45.83	30.33	29.17
	1-NN	67.00	50.00	48.67
	2-NN	66.67	47.63	46.50
	5-NN	65.50	44.50	45.83
Testing	ML	42.64	32.06	30.72
	1-NN	45.33	35.63	34.51
	2-NN	46.79	38.59	37.38
	5-NN	45.03	35.29	36.19

Table 6.7 Average Measures of Conflict between Pairs of Sources using Partially Consonant Belief Function for Training Samples.

	SAR Shallow	SAR Steep	Aspect	Elevation	Slope
A/B MSS	.362	.402	.390	.343	.417
SAR Shallow		.328	.384	.391	.424
SAR Steep			.402	.387	.433
Aspect				.397	.395
Elevation					.410

Table 6.8 Average Measures of Conflict between Pairs of Sources using Partially Consonant Belief Function for All Samples.

	SAR Shallow	SAR Steep	Aspect	Elevation	Slope
A/B MSS	.375	.411	.402	.352	.421
SAR Shallow		.336	.407	.384	.429
SAR Steep			.413	.401	.446
Aspect				.399	.382
Elevation					.413

In order to rank the sources by their reliability, the average J-M distance and the average Transformed Divergence of each source were calculated and compared with the overall classification accuracy obtained by the ML method over the training samples (Table 3.2). We also computed the average measures of conflict between pairs of the sources using the consonant belief function (Tables 3.3, 3.4) and the partially consonant belief function (Tables 6.7, 6.8). Assuming that A/B MSS is the most reliable in the sense of classification, all the measures agree that Elevation and SAR-Shallow are the 2nd and the 3rd, respectively. They do not agree at all for the remaining sources. In the multisource data classification with this data set, the remaining sources have been considered as equally reliable.

For the purpose of comparison, the ML classification based on the stacked vector approach was carried out for various sets of the data sources, adding one source at a time to the A/B MSS data in the order Elevation, SAR-Shallow, SAR-Steep, Aspect, and Slope. Then the MSD classification was performed using different combinations of interval-valued belief functions and decision rules. Tables 6.9 and 6.10 compare the results for the training samples and the test samples, respectively. Even though the compounded data in the ML classification were treated as having Gaussian distributions, the ML and the MSD methods produced similar results for the training samples. This is not surprising because the ML method uses conventional additive probabilities assuming that the knowledge concerning the actual unknown probabilities is complete, which is reasonable as far as the training samples are concerned.

In the MSD classification using the partially consonant belief function (PCBF), the information classes were partitioned as {df2, df3, df+lp2} and {df+os2, hc, fc}. This partition was made on the basis of the classwise separability measures of the individual sources so that the average separability between the partitions is maximized.

Comparing the performance of the two belief functions, the consonant belief function (CBF) was better for the training samples while PCBF was better for the test samples. It is not known at this point whether CBF or PCBF is better. As far as the decision rules are concerned, the maximum plausibility (MP) rule was superior to the other rules, the maximum support (MS) rule and the maximum average probability (MA) rule. It is also not known in general which

Table 6.9 Results of ML Classification and MSD Classification over Training Samples of Anderson River Data.

	Decision Rule	Sources					
		1	1, 4	1, 2, 4	1 – 4	1 – 5	1 – 6
ML		82.50	88.67	91.67	92.00	92.83	93.50
CBF	MP	–	89.83	92.00	92.50	93.17	94.33
	MS	–	88.67	91.17	91.33	92.33	93.67
	MA	–	88.50	91.00	91.67	91.67	93.50
PCBF	MP	–	88.67	91.50	92.17	92.67	93.83
	MS	–	86.83	89.67	91.33	91.00	92.17
	MA	–	87.50	90.17	91.83	91.67	92.83

Table 6.10 Results of ML Classification and MSD Classification over Test Samples of Anderson River Data.

	Decision Rule	Sources					
		1	1, 4	1, 2, 4	1 – 4	1 – 5	1 – 6
ML		74.16	77.77	79.13	78.93	79.80	81.01
CBF	MP	–	80.60	82.39	82.69	83.02	84.54
	MS	–	78.45	81.42	81.67	82.24	83.65
	MA	–	78.21	80.95	82.05	81.88	83.16
PCBF	MP	–	80.86	82.76	83.15	84.27	85.95
	MS	–	78.94	81.31	81.64	83.05	84.16
	MA	–	78.49	81.67	82.25	83.78	84.44

rule is the best. Further research is needed to determine whether guidelines can be devised for selection of the belief function and decision rule.

The MSD classification for all the sources was iteratively performed with various degrees of source reliability. In this case, the MP rule was used as a decision rule because it produced the best results in the classification of multiple data sources with equal reliabilities. Tables 6.11 and 6.12 show the overall classification results over the training samples and the test samples, respectively. The results show not only that the classification accuracy may increase as the reliabilities of the additional data sources are varied but also that it can be degraded if the additional data sources are discounted too much. It is also observed that the variations in the accuracy by PCBF are relatively smaller than those by CBF. The reason is because the width of a partially consonant interval-valued probability is usually less than the width of a

Table 6.11 Results of MSD Classification over Training Samples of Anderson River Data with Various Degrees of Source Reliability.

	Source Reliability						Overall (%)
	1	2	3	4	5	6	
CBF	1.0	1.0	1.0	1.0	1.0	1.0	94.33
	1.0	0.8	0.8	0.8	0.8	0.8	95.17
	1.0	0.8	0.6	0.8	0.6	0.6	95.83
	1.0	0.7	0.5	0.7	0.5	0.5	95.00
	1.0	0.6	0.4	0.8	0.4	0.4	93.83
PCBF	1.0	1.0	1.0	1.0	1.0	1.0	93.83
	1.0	0.8	0.8	0.8	0.8	0.8	95.00
	1.0	0.8	0.6	0.8	0.6	0.6	95.17
	1.0	0.7	0.5	0.7	0.5	0.5	93.67
	1.0	0.6	0.4	0.8	0.4	0.4	91.67

consonant interval-valued probability, which makes PCBF less sensitive to the changes in source reliability.

Overall, the MSD classification using evidential reasoning was able to produce higher accuracy than the ML classification. The increase in the classification accuracy obtained by the MSD classification should be primarily attributed to the ER method's capability of adequately representing bodies of statistical evidence by interval-valued probabilities. Furthermore, the MSD classification was capable of incorporating various degrees of source reliability into the process by treating the multiple sources separately. It was also possible in this particular experiment to utilize non-parametric information using the k-NN method together with parametric information. This is another advantage of the MSD classification by treating the multiple sources separately.

Table 6.12 Results of MSD Classification over Test Samples of Anderson River Data with Various Degrees of Source Reliability.

	Source Reliability						Overall (%)
	1	2	3	4	5	6	
CBF	1.0	1.0	1.0	1.0	1.0	1.0	84.54
	1.0	0.8	0.8	0.8	0.8	0.8	85.40
	1.0	0.8	0.6	0.8	0.6	0.6	85.69
	1.0	0.7	0.5	0.7	0.5	0.5	84.25
	1.0	0.6	0.4	0.8	0.4	0.4	83.04
PCBF	1.0	1.0	1.0	1.0	1.0	1.0	85.95
	1.0	0.8	0.8	0.8	0.8	0.8	86.09
	1.0	0.8	0.6	0.8	0.6	0.6	86.74
	1.0	0.7	0.5	0.7	0.5	0.5	85.27
	1.0	0.6	0.4	0.8	0.4	0.4	83.21

6.3. Classification of Single-Source Multispectral Data

In the previous section, the proposed method was applied to the classification of multisource data obtained by various sensors. The data set used in this section is 12-band Airborne MSS data whose flightline ID is "CRN BLT LO FL21" taken on August 21, 1971. Table 6.13 describes the spectral regions and bands of the 12 input channels comprising the MSS data. The size of the image is 220 lines by 140 columns, and the image covers an agricultural area in Indiana. Figure 6.13 is the ground truth map of this area, which is digitized and geometrically registered with the MSS data imagery.

Although the registration has been made very carefully, the ground truth map contains geometric registration errors. The error is more noticeable along the boundaries between different ground types. If the whole area were used for test, incorrect classifications evaluated on the basis of the ground truth map would result not only from bad performance of a classifier but also from the geometric registration error. In order to avoid this confusion, test areas were chosen. Figure 6.14 shows the test areas on the MSS image (Channels 1, 4, 9). There were 9 information classes for the test, and Table 6.14 lists them with their actual number of pixels counted from the ground truth map.

This experiment was designed to observe how the proposed method overcomes the Hughes phenomenon when the number of training samples is so small. The strategy underlying the method is to decompose the relatively large body of evidence into smaller, more manageable pieces, to assess plausibilities based on each piece, and to combine the assessments by a combination rule.

The set of multiple data sources was formed as shown in Table 6.15 by dividing the 12-band MSS data based on the global statistical correlation (Table 6.16) which coincides with the spectral regions. As expected, the correlation between pairs of bands from different spectral regions (except the thermal region) are relatively low compared to those within each spectral region. Even though the thermal band was relatively highly correlated with the visible bands, we chose to treat it as though it were a distinct source. The consequence of having done so is apparent in the experimental results.

Table 6.13 Description of Airborne MSS Data of Indiana Agricultural Area Data Set.

Spectral Region	Input Channel	Spectral Band(μm)
Visible	1	0.46 - 0.49
	2	0.48 - 0.51
	3	0.50 - 0.54
	4	0.52 - 0.57
	5	0.54 - 0.60
	6	0.58 - 0.65
	7	0.61 - 0.70
Near Infrared	8	0.72 - 0.92
	9	1.00 - 1.40
Middle Infrared	10	1.50 - 1.80
	11	2.00 - 2.60
Thermal	12	9.30 -11.70

Table 6.14 Information Classes in Indiana Agricultural Area Data Set.

Class Index	Cover Types	No. of Test Samples	% of Total
1	Corn	3489	26.11
2	Soybean	6454	48.31
3	Non-Farm	593	4.44
4	Oat	398	2.98
5	Wheat	602	4.51
6	Sudex	936	7.01
7	Hay	412	3.08
8	Wood	361	2.70
9	Pasture	115	0.86
Total		13360	100.0

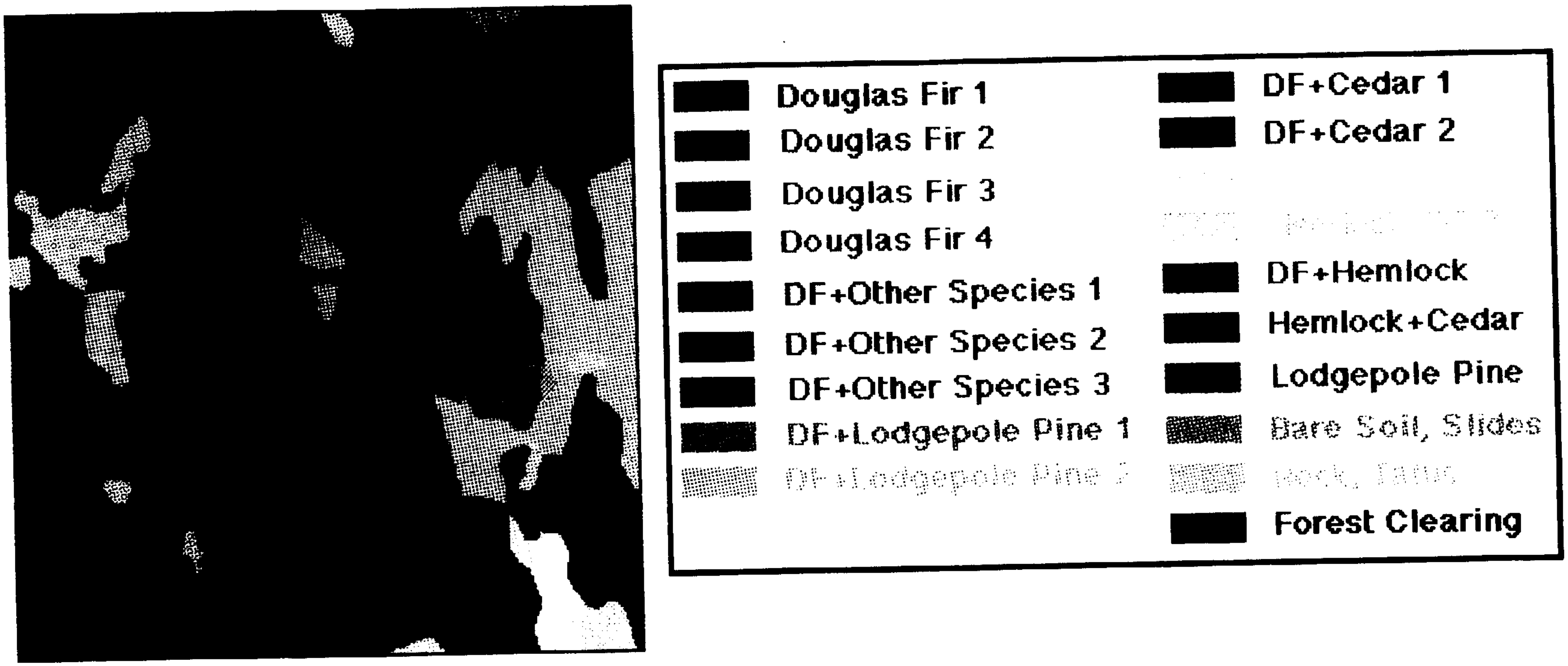


Figure 6.2 Ground Truth Map of Anderson River Data Set.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

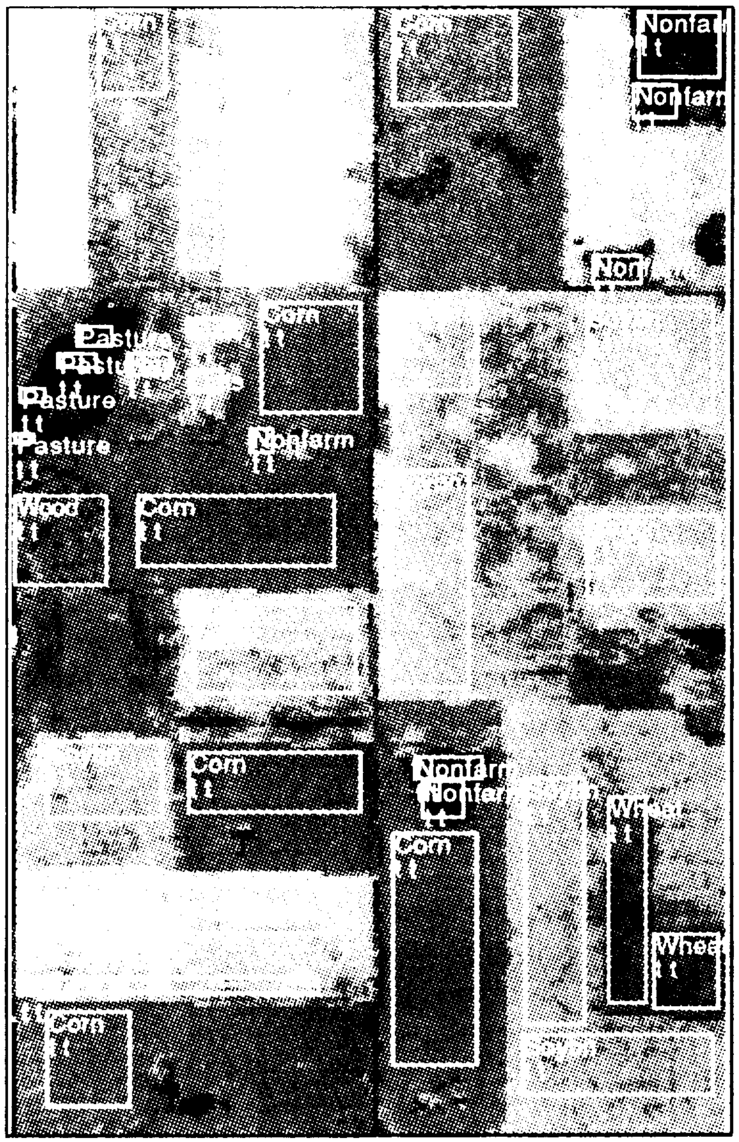


Figure 6.14 Test Areas over A/B MSS (Channel 8) Image of Indiana Agricultural Area Data.

For each class, from 15 to 30 samples uniformly distributed over the test fields were selected for training. First, the ML classification was performed with various sets of the input bands. Tables 6.17 and 6.18 are the results over the training samples and the test samples, respectively. The overall classification accuracy is the percentage ratio of the number of the correctly classified pixels to the total number of pixels while the average classification accuracy is the arithmetic mean of the classwise accuracies.

Then the proposed method was applied to subsets of the input channels, treating them as multiple sources. Tables 6.19 and 6.20 show the results of this MSD classification over the training samples and the test samples, respectively. In this case, the consonant belief function and the maximum plausibility rule were adopted, and the "multiple sources" were assumed equally reliable.

In the ML classification, both the overall and the average accuracies increased as the number of features was increased for the training samples; but this was not true for the test samples. In the MSD classification utilizing all input channels, although both accuracies were below 100% for the training samples, they were comparable to or higher than the accuracies produced by the ML method. The results exhibit two interesting features. First, the classification accuracy for the MSD classifications decreases as the set of bands is more finely subdivided. This is because more information in inter-channel statistical correlation is lost as the data set is more finely subdivided. Second, there is a

Table 6.15 Divided Sources of Indiana Agricultural Area Data Set.

Source Index	Spectral Region	Input Channels
1	Visible	1 to 7
2a	Near Infrared	8 9
2b	Middle Infrared	10 11
2c	Near & Middle Infrared	8 to 11
2d	Thermal	12
2	Infrared	8 to 12

Table 6.16 Statistical Correlation Coefficient Matrix of Indiana Agricultural Area Data Set.

Band	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	.864	.899	.750	.836	.839	.909	-.312	.284	.413	.556	.726
2		1.000	.893	.696	.868	.913	.939	-.355	-.267	.371	.577	.767
3			1.000	.885	.954	.892	.904	-.195	-.174	.442	.547	.691
4				1.000	.919	.794	.733	.047	-.003	.491	.489	.542
5					1.000	.908	.885	-.160	-.140	.448	.548	.692
6						1.000	.936	-.353	-.310	.393	.577	.805
7							1.000	-.445	-.358	.409	.607	.830
8								1.000	.858	.350	.076	-.520
9									1.000	.517	.254	-.415
10										1.000	.861	.378
11											1.000	.623
12												1.000

Table 6.17 Results of ML Classification over Training Samples for Various Sets of Input Bands.

	Percent Agreement with Ground Truth Map										
	Class Index (No. of Pixels per Class)									Accuracy	
Input Bands	1 (30)	2 (30)	3 (15)	4 (15)	5 (15)	6 (18)	7 (15)	8 (15)	9 (15)	Overall	Average
1 to 12	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
1 to 7	96.67	96.67	100.00	80.00	93.33	83.33	93.33	86.67	93.33	92.26	91.48
8 to 12	100.00	96.67	100.00	66.67	100.00	94.44	93.33	60.00	100.00	91.67	90.12
8 to 11	96.67	76.67	100.00	46.67	100.00	72.22	93.33	66.67	93.33	83.33	82.84
8, 9	96.67	83.33	86.67	0.00	6.67	77.78	73.33	20.00	86.67	64.88	59.01
10, 11	100.00	83.33	100.00	20.00	20.00	0.00	53.33	40.00	93.33	61.90	56.67
12	83.33	86.67	93.33	0.00	40.00	11.11	0.00	0.00	0.00	43.45	34.94

Table 6.18 Results of ML Classification over Test Samples for Various Sets of Input Bands.

	Percent Agreement with Ground Truth Map										
	Class Index (No. of Pixels per Class)									Accuracy	
Input Bands	1 (3489)	2 (6454)	3 (593)	4 (398)	5 (602)	6 (936)	7 (412)	8 (361)	9 (115)	Overall	Average
1 to 12	99.08	97.92	87.02	42.71	68.94	90.81	19.90	66.20	91.30	90.97	73.77
1 to 7	89.45	72.89	91.57	41.21	80.56	67.09	38.59	43.49	71.30	75.17	66.23
8 to 12	96.70	91.56	99.16	40.70	95.51	71.47	74.21	54.85	97.39	89.02	80.18
8 to 11	96.10	73.27	97.64	33.92	91.86	63.25	72.33	54.29	94.78	78.92	75.27
8, 9	90.51	82.00	86.68	0.00	10.80	66.35	54.13	15.24	95.65	75.13	55.70
10, 11	93.24	60.75	93.76	10.80	20.26	4.70	56.55	26.87	95.65	62.72	51.40
12	81.11	84.13	90.21	0.00	34.72	37.07	0.00	0.00	0.00	69.99	36.36

Table 6.19 Results of MSD Classification over Training Samples.

Input Sources	Percent Agreement with Ground Truth Map: Class Index (No. of Pixels per Class) 1 (30)	2 (30)	3 (15)	4 (15)	5 (15)	6 (18)	7 (15)	8 (15)	9 (15)	Accuracy: Overall	Average
1, 2	100.00	100.00	100.00	86.67	100.00	94.44	100.00	100.00	100.00	98.21	97.90
1, 2c, 2d	100.00	100.00	100.00	80.00	100.00	94.44	100.00	100.00	100.00	97.62	97.16
1,2a,2b,2d	100.00	96.67	100.00	73.33	100.00	88.89	100.00	100.00	93.33	95.24	94.69

Table 6.20 Results of MSD Classification over Test Samples.

Input Sources	Percent Agreement with Ground Truth Map: Class Index (No. of Pixels per Class) 1 (3489)	2 (6454)	3 (593)	4 (398)	5 (602)	6 (936)	7 (412)	8 (361)	9 (115)	Accuracy: Overall	Average
1, 2	97.70	95.51	96.12	55.78	96.68	84.51	70.87	82.27	97.39	93.08	86.31
1, 2c, 2d	96.85	91.78	95.62	47.74	96.51	76.39	63.11	82.27	93.91	89.97	82.69
1,2a,2b,2d	96.96	91.74	95.28	38.44	93.36	75.64	57.28	85.04	95.05	89.41	81.01

considerable increase in the average classification accuracy of the MSD classification for the test samples as compared to the ML classification accuracy. It is expected because the MSD classification classifies pixels based on the assessment of multiple sources instead of a single source. This is a major advantage of the MSD classification over any single source data classification. While the ML classification based on the stacked vector approach combines the features in the raw data level and buries their relative reliabilities in the statistical correlation information, the MSD classification combines the multiple groups of the features after assessing the individual groups with explicit consideration of their relative reliabilities.

In order to demonstrate the Hughes phenomenon, the ML classification over the test samples was performed with various numbers of the best features as determined by feature selection using both the J-M distance and the Transformed Divergence. The result of the feature selection was, from best to worst: 8, 12, 11, 10, 9, 7, 6, 4, 5, 3, 2, and 1. As shown in Figure 6.15, the ML method gave the highest accuracies at 8 features (8, 12, 11, 10, 9, 7, 6, 4).

However, the MSD classification based on the proposed method was able to utilize all features when applied to a "multisource" data set consisting of two "sources": one having the 8 best features and the other having the remaining 4 features. The first 4 lines in Table 6.21 are the results of classification with various degrees of reliability applied to the second source.

Another set of multisource data was formed by dividing the features into two groups each of which has roughly equally good features. The classification result from applying the proposed method to this data set is shown in the last line of Table 6.21. In this particular case, although the dependencies between sources were ignored, the accuracies were the highest. This is due to the reinforcing characteristic of Dempster's rule, which means that the combined body of evidence provides stronger support than any individual body of evidence.

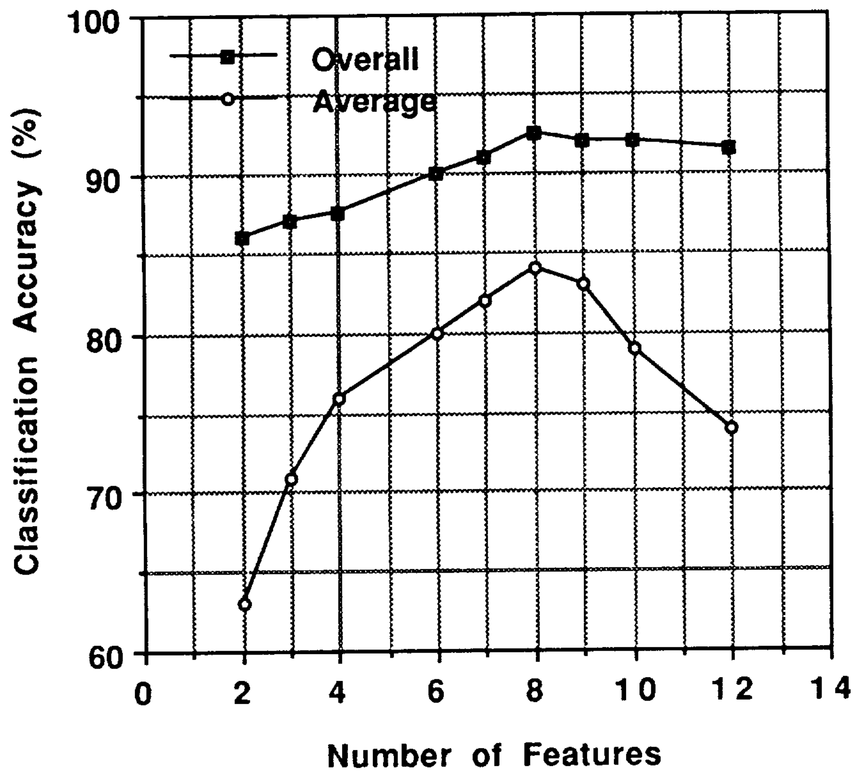


Figure 6.15 Results of ML Classification over Test Samples for Various Numbers of Features

Table 6.21 Results of MSD Classification for Data Set formed by Feature Selection.

Bands in Source 1 (Source Reliability)	Bands in Source 2 (Source Reliability)	Overall	Average
8 12 11 10 9 7 6 4 (1.0)	5 3 2 1 (1.0)	95.27	89.29
(1.0)	(0.9)	96.07	90.42
(1.0)	(0.8)	96.65	90.07
(1.0)	(0.7)	96.81	89.96
8 11 9 6 5 2 (1.0)	12 10 7 4 3 1 (1.0)	96.89	91.13

6.4. Classification of HIRIS Data

The High Resolution Imaging Spectrometer(HIRIS) is an Earth Observing System (EOS) sensor developed for high spatial and high spectral resolution. It can provide more information in the 0.4 to 2.5μm spectral region than any other earth-observing sensor. Table 6.22 compares some of the attributes of HIRIS and early Earth satellite observing sensors. [Goetz and Herring (1989)]

The high dimensionality of HIRIS data causes several difficulties in classifying the data. In addition to the high computational cost of classifying such data, a huge amount of training samples is necessary in order to have accurate estimation of the statistical parameters using all 192 channels. Furthermore, unless these parameters can be accurately estimated, it is even impossible to use statistical feature selection techniques to reduce the dimensionality.

In this section, the proposed method is applied to the classification of HIRIS data by decomposing the data into smaller pieces, i.e., subsets of

Table 6.22 Comparisons of MSS, Thematic Mapper (TM) and HIRIS.

	MSS	TM	HIRIS
No. of Spectral Bands	4	7	192
IFOV(ground)	79m	30/120m	30m
Dynamic Range	6/7 bits	8 bits	12 bits
Swath Width	185km	185km	30km
Data Rate	7.63Mbits/sec	67.4Mbits/sec	300Mbits/sec
Spectral Region	0.5 - 1.1μm	0.45-0.90μm 1.55-1.75μm 2.08-2.35μm 10.4-12.5μm	0.4-2.5μm
Spectral Resolution	0.1-0.3μm	0.6-2.27μm	0.01μm

spectral bands. The data set used in this experiment is simulated HIRIS data obtained by RSSIM [Kerekes and Landgrebe (1989)]. RSSIM is a simulation tool for the study of multispectral remotely sensed images and associated system parameters. It creates realistic multispectral images based on detailed models of the ground surface, the atmosphere, and the sensor. Table 6.23 provides a description of the simulated HIRIS data set.

Figure 6.16 is a visual representation of the global statistical correlation coefficient matrix of the data. The image is produced by converting the absolute values of coefficients to gray values between 0 and 255. Based on the correlation image, the 201 bands were divided into 3 groups in such a way that intra-correlation is maximized and inter-correlation is minimized. Table 6.24 describes the multisource data set after division. Note that the spectral regions of the input channels in Source 3 coincide with the water absorption bands.

With 225 training samples (a third of the total samples) for each class, the ML classification and the multisource data classification using the consonant belief function and the maximum plausibility decision rule were performed over the total samples for various sets of the sources, and the results are listed in Tables 6.25 and 6.26. In the multisource data classification for Source 1 and Source 2, first the sources were given the equal reliability and then Source 2 was discounted with degree of reliability 0.9 to show the effect of varying degrees of reliability on the classification accuracy.

Table 6.23 Description of Simulated HIRIS Data Set.

Name	Finney County Data Set
Data Type	201-band HIRIS data simulated by RSSIM
Spectral Region	0.4 - 2.4μm
Spectral Resolution	0.01μm
Image Size	45 lines $\times$ 45 columns (2025 samples)
Information Classes	Winter Wheat, Summer Fallow, Unknown

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

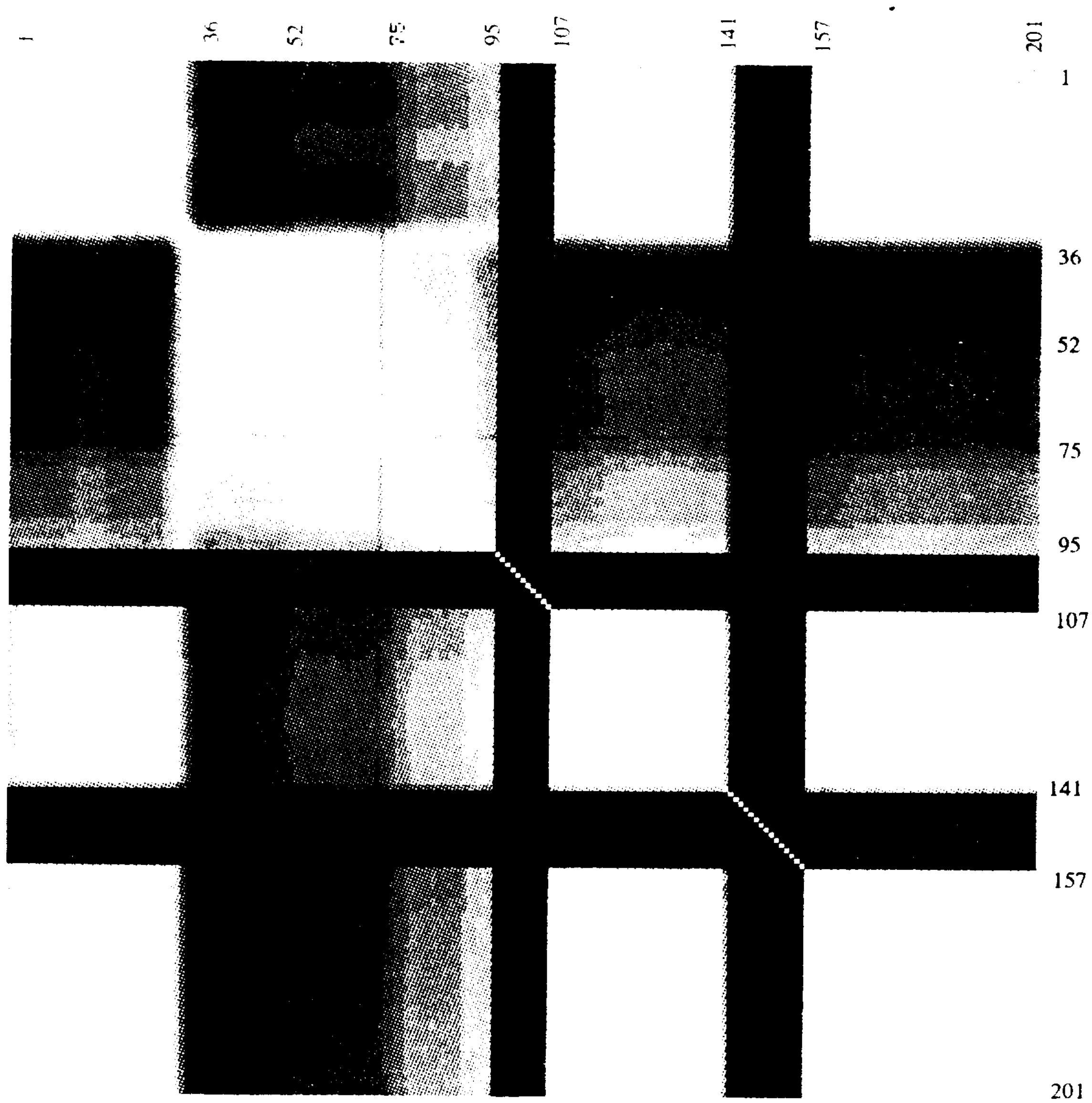


Figure 6.16 Global Statistical Correlation Coefficient Image of Finney County Data Set.

The results of the ML method apparently show effects of the Hughes phenomenon; the accuracy goes down as the dimensionality of the source increases while the number of training samples is fixed. In particular, the accuracy decreases by a considerable amount when all features are used. Presence of the Hughes phenomenon causes the ML method to be particularly sensitive to a bad source, Source 3 in this case. Meanwhile, the proposed MSD classification method always shows robust performance and gives consistent results.

To explore how to handle a situation in which the training samples were too limited to permit use of all available features, both methods were run again with 68 training samples (10% of the total samples), and the results are shown in Table 6.27. In this case, the features were selected with a uniform spectral interval from Source 1 and Source 2, excluding the features in Source 3. The table shows the number of features actually used for the subdivided sources. Four cases were run, each with a different spectral interval, resulting in a total of 51, 40, 31, and 20 features, respectively. The proposed method performed better in all four cases than did the ML method.

Table 6.24 Divided Sources of HIRIS Data Set.

Source Index	**Input Channels**	**No. of Features**
Source 1	1- 35, 107 - 141, 157 - 201	115
Source 2	36 - 95	60
Source 3	96 - 106 (1.35 - 1.45μm) 142 - 156 (1.81 - 1.95μm)	26

Table 6.25 Results of ML Classification with 225 Training Samples.

Source	S1	S2	S3	S1, S2	All
Classification Accuracy (%)	75.75	75.60	45.83	74.56	65.14

Table 6.26 Results of Multisource Data Classification with 225 Training Samples.

Reliability of			Classification Accuracy (%)
S1	S2	S3	
1.0	1.0	1.0	77.63
1.0	1.0	not used	77.83
1.0	0.9	not used	78.32

Table 6.27 Results of Classifications with 68 Training Samples.

	Classification Accuracy (%)							
Sources	S1	S2	S1	S2	S1	S2	S1	S2
# Features	33	18	27	13	21	10	14	6
ML	77.43		82.40		82.86		81.82	
MSDC	82.22		84.10		85.04		81.90	

6.5. Discussion

In this chapter, the Evidential Reasoning (ER) multisource data classification method presented in Chapters 3, 4, and 5 has been applied to the ground-cover classification of various multisource data sets. Once it is determined which belief function and decision rule will be used, the implementation of the method is as easy as implementing a typical ML method.

The first experiment, with the multisource data set consisting of 3 multi-channel data sources and 3 topographic data sources, was intended to assess the ability of the ER method in capturing and utilizing the information obtained from the topographic data sources as well as the multispectral data sources. In this particular experiment, some of the classes could not be assumed to be normally distributed in the topographic data. Thus, in the MSD classification based on the ER method, the nonparametric Nearest Neighbor method was adopted to compute the likelihood functions of test samples, which were then used to construct the IV belief functions for the bodies of evidence provided by the topographic data sources. By treating the multiple data sources separately, the proposed method was able not only to utilize nonparametric information together with parametric information but also to incorporate various degrees of source reliability into the process. The method provides more than one choice for representation of statistical evidence and a decision rule; these choices give a lot of flexibility to the multisource data analysis. At this point in the research it is not known exactly which choices should be made in general; the choices must depend on our knowledge concerning the context of the specific problem, such as the hierarchical structure of information classes and the amount and reliability of available information.

The ER method was also applied to the classification of two single-source data sets: 12-band A/B MSS data, and 201-band simulated HIRIS data. Both experiments were designed to observe how effectively the proposed method utilizes the available features and overcomes the Hughes phenomenon when the number of training samples is small. From single-source data a multisource data set was formed by decomposing the high-dimensional data into smaller and more manageable pieces based on the global statistical correlation information.

In the experimental results for the 12-band A/B MSS data, two observations were made: (1) the classification accuracy of the MSD classifications decreased as the set of bands was more finely subdivided, and (2) the average classification accuracy of the MSD classification increased significantly compared to the ML classification accuracy. According to the first observation, inter-channel statistical correlation must be kept within the subdivided sources (consistent with the independence assumption of Dempster's combination rule). Similar results were observed when the MSD classification was performed for the set of features subdivided based on feature selection. Although dependencies between sources were ignored, the classification accuracy was increased due to the reinforcing characteristic of Dempster's rule.

The experimental results for the 201-band simulated HIRIS data showed that the MSD classification provided robust and consistent performance despite the existence of an inconsistent source when training samples were very limited. The information obtained from an inadequate number of training samples is considered to be inexact and incomplete. The results have demonstrated the ability of the ER method to capture uncertain information based on inexact and incomplete bodies of evidence, and consequently to utilize features as effectively as possible.

CHAPTER 7

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 Conclusions

The problem of drawing inferences using subjective probability measures is not a trivial one, especially when it involves multiple information sources associated with various degrees of relative reliability. In this report we have investigated how interval-valued probabilities can be used to represent and integrate evidential information obtained from various data sources.

IV probability is a generalization of the conventional point-valued probability. It has been known as a more adequate scheme than the conventional additive probabilities for representing partial information provided by inexact and incomplete sources. Chapter 2 reviewed various systems of IV probabilities and introduced an axiomatic approach to IV probabilities. In the axiomatic approach the upper and the lower probabilities are given by a pair of set-theoretic functions.

One of the basic problems in applying IV probabilities to a real-world problem is how to infer the upper and the lower probability functions given a body of evidence. Chapter 3 investigated formal methods of constructing IV probability functions when the given body of evidence is based on the outcomes of statistical experiments governed by a probability model. This report has mainly focused on the two IV belief functions, the consonant and the partially consonant belief functions, which are based on the Likelihood Principle. Even though they require certain assumptions which are not difficult to satisfy in practice, they have mathematically simple and readily usable formulas. In order to include the relative reliabilities of sources in a multisource data analysis, the attempts to represent quantitatively the degree of reliability by the average Jeffries-Matusita distance, the average Transformed Divergence, and the average measures of conflict between pairs of sources were made. These measures were used to rank the multiple sources according to the relative reliabilities of the sources.

In the analysis of multiple data sources, a combination rule is an essential tool in order to base inferences and decisions on all available information. Chapter 4 formally stated desirable properties of combination rules and investigated the inferencing mechanisms of the subjective Bayesian updating rules and Dempster's rule for combining multiple bodies of evidence. It was also noted that Dempster's rule is a generalized form of Bayesian inference, which is characteristically reinforcing and robust to small variations in probability measures to be combined. The robustness of Dempster's rule was analyzed in the aspect of its differential behavior according to slight changes of initial belief measures.

Chapter 5 presented an account of basic elements in the decision theory for pattern recognition based on IV probabilities and developed the absolute rule and the Bayes-like decision rule for evidential intervals on the basis of the general interval-valued expectation. A problem with these rules is that there may happen ambiguous situations where decisions cannot be made. The minimum average expected loss rule was proposed to resolve such ambiguous situations. Further, the minimum upper expected loss rule and the minimum lower expected loss rule were proposed as alternatives to the previous two rules.

Overall concepts of interval-valued probabilities have been implemented and evaluated as a new method for classification of multisource data in remote sensing. As described in Chapter 6, the proposed method was applied to three separate sets of multisource data, one consisting of three multi-channel data sources and three topographic data sources, and two consisting of single-source multispectral data. The purpose of applying the method to the single-source data sets was to utilize as many features as effectively as possible (when training samples are limited) by decomposing a large number of channels into smaller and more manageable subsets based on the global statistical correlation.

In the method each data source is considered as a body of evidence providing partial information. When the body of evidence is represented by IV probabilities, the width of the interval represents the uncertainty associated with the corresponding source. The method combines the individual bodies of evidence into the total body of evidence. By treating the data sources separately, the method is not only able to utilize both parametric and nonparametric information but also able to incorporate various degrees of source reliability in the multisource data analysis.

The experimental results showed that compared to the conventional ML classification, the proposed method gave higher and more robust classification accuracies for test samples even when a far less reliable source was included in the data set. The increase in average classification accuracy was noteworthy. The results also showed that the classification accuracies could be increased by varying the degree of reliability assigned to each source as well as by choosing an appropriate decision rule.

The most important feature of the method is the capability of plausible reasoning under uncertainty in pattern recognition, especially where multiple data sources are not 100% reliable or provide conflicting information. The method of classification for multisource data based on IV probabilities can also be used to good advantage when there are only small numbers of training samples and reliable estimation of statistical information requires dividing the high-dimensional data into lower-dimensional subsets.

7.2 Suggestions for Further Research

The Evidential Reasoning method developed in this work could be further improved in the following respects:

(1) Computational complexity: It is apparent that the processing time will increase as the number of sources increases. Furthermore, since Dempster's rule computes the IV probability of a subset $A \subset \Omega$ as the sum of the basic probability assignments of A and all the subsets of A, the computational complexity grows exponentially with the number of elements in Ω. A possible way to reduce the computation is to restrict the number of focal elements to be considered. In a remote sensing context, this is possible by designing the

classes hierarchically.

(2) Generalization of the minimum average loss rule: Although the minimum upper expected loss rule (maximum plausibility rule) produced the best results in the experiments, it is considered to be due to the belief functions used. In general, the minimum average loss rule is considered to be more reliable than any other rule because it includes both the upper and the lower probabilities. This rule may be generalized by considering the IV expected loss as a convex set of measures.

LIST OF REFERENCES

Benediktsson, J.A., and P.H. Swain (1989a), "A method of statistical multisource classification with a mechanism to weight the influence of the data sources", *Proc. '89 IGARSS*, Vancouver, B.C., Canada, July 10-14, pp. 517-520.

Benediktsson, J.A., P.H. Swain, and O.K. Ersoy (1989b), "Neural network approaches versus statistical methods in classification of multisource remote sensing data", *Proc. '89 IGARSS*, Vancouver, B.C., Canada, July 10-14, pp. 489-492.

Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis,* 2nd Ed., Springer-Verlag.

Blonda, P., et al. (1989), "Classification of multitemporal remotely sensed images based on a fuzzy logic technique", *Proc. '89 IGARSS*, Vancouver, B.C., Canada, July 10-14, pp. 834-837.

Charniak, E. (1983), "The Bayesian basis of common sense medical diagnosis", *AAAI-83*, pp. 70-73.

Cheng, Y., and K.S. Fu (1985), "Conceptual clustering in knowledge organization", *IEEE Trans. PAMI*, **PAMI-7**, pp. 592-598.

Cheng, Y., and R.L. Kashyap (1986), "Construction and interpretations of formulas for combining evidence", *Proc. 8th Int. Conf. on Pattern Recognition*, Paris, France, Oct. 27-31, pp. 1226-1230.

Choquet, G. (1953), "Theory of capacities", *Ann. Inst. Fourier*, vol. **5**, pp. 131-295.

Cohen, P. (1985), *Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach*, Pitman Publishing, Inc.

Dempster, A.P. (1968), "A generalization of Bayesian inference", *J. of the Royal Stat. Soc., Ser. B*, vol. **30**, pp. 205-247.

Dempster, A.P. (1967), "Upper and lower probabilities induced by a multivalued mapping", *Ann. Math. Stat.*, vol. **38**, pp. 325-339.

Devijver, P.A., and J. Kittler (1982), *Pattern Recognition: A Statistical Approach*, Prentice Hall.

Duda, R., J. Gaschnig, and P. Hart (1979), "Model design in the PROSPECTOR consultant system for mineral exploration", in *Expert Systems in the Micro-electronic Age* by D. Michie (Eds.), Edinburg Univ. Press, Edinburg, pp. 157-167.

Duda, R., P. Hart, and N. Nilsson (1976), "Subjective Bayesian methods for rule-based inference systems", *AFIPS Conf. Proc.*, **45**, pp. 1075-1082.

Dubois, D., and H. Prade (1985), "Combination and propagation of uncertainty with belief functions", *Proc. 7th Int. Joint Conf. on AI.*

Dubois, D., and H. Prade (1980), *Fuzzy sets and systems : theory and applications*, Academic Press.

Eom, K.B. (1986), "Robust image models with application", Ph.D. Thesis, Purdue University, West Lafayette, IN.

Fleming, M.D., and R.M. Hoffer (1979), "Machine processing of Landsat MSS data and DMA topographic data for forest cover type mapping", *LARS Technical Report* 062879, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907.

Franklin, J., T.L. Logan, C.E. Woodcock, and A.H. Strahler (1986), "Coniferous forest classification and inventory using Landsat and digital terrain data", *IEEE Trans. Geos. and Remote Sensing*, vol. **GE-24**, no. 1, pp. 139-149.

French, S. (1985), "Group consensus probability distributions:A critical survey", *Bayesian Statistics 2* by J.M. Bernardo, DeGroot, Lindley, and Smith (Eds.), Elsevier Science Publishers, North-Holland, pp. 183-202.

Fukunaga, K. (1972), *Intro. to Statistical Pattern Recognition*, Academic Press.

Garvey, T.D. (1987), "Evidential reasoning for geographic evaluation for helicopter route planning", *IEEE Trans. Geos. and Remote Sensing*, vol. **GE-25**, no. 3, pp. 294-304.

Garvey, T.D., J.D. Lowrance, and M.A. Fischler (1981), "An inference technique for integrating knowledge from disparate sources", *Proc. 7th Joint Conf. on AI*, pp. 319-325.

Genest, C. (1986), "Combination of subjective opinion:An application and its relation to the general theory", *Reliability and Quality* by A.P. Basu (Ed.), Elsevier Science Publishers, North-Holland, pp. 141-155.

Ginsberg, M.L. (1984), "Non-monotonic reasoning using Dempster's rule", *AAAI-84*, pp. 126-129.

Goetz, A.F.H., and M. Herring (1989), "The High Resolution Imaging Spectrometer (HIRIS) for Eos", *IEEE Trans. Geos. and Remote Sensing*, vol. **GE-27**, no. 2, pp. 136-144.

Good, I.J. (1962), "Subjective probability as the measure of a non-measurable set", in *Logic, Methodology, and Philosophy of Science*, by E. Nagel, Suppes and Tarski (Eds.), Stanford Univ. Press, pp. 319-329.

Healey, R., P. Dowie, A. Mowle, and J. Holbrook (1988), "Integrating remote sensing data into a geographic information system: A foundation for rural land use strategies - Nature conservancy council project", *Proc. '88 IGARSS*, Edinburgh, Scotland, Sep. 13-16, pp. 111-112.

Hoffer, R.M., and staff (1975), "Computer-aided analysis of Skylab multispectral scanner data-mountainous terrain for land use, forestry, water resources, and geologic applications", *LARS Information Note* 121275, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47907.

Huber, P.J. (1981), *Robust Statistics*, John Wiley & Sons, New York.

Huber, P.J. (1973), "The use of Choquet capacities in statistics", *Int. Stat. Inst. Bulletin* **45**, Book 4, pp. 181-188.

Huber, P.J., and V. Strassen (1973), "Minimax tests and the Neyman-Pearson Lemma for capacities", *Ann. Stat.*, vol. **1**, pp. 251-263.

Hughes, G.F. (1968), "On the mean accuracy of statistical pattern recognizers", *IEEE Trans. Information Theory*, vol. **IT-14**, no. 1, pp. 55-63.

Hutchinson, C.F. (1982), "Techniques for combining Landsat and ancillary data for digital classification improvement", *Photogrammetric Engineering and Remote Sensing*, vol. **48**, no. 1, pp. 123-130.

Johnson, R.W. (1986) "Independence and Bayesian updating methods", *Artificial Intelligence*, vol. **29**, pp. 217-222.

Jones, A.R., J.J. Settle, and B.K. Wyatt (1988), "Use of digital terrain data in the interpretation of SPOT-1 HRV multispectral imagery", *Int. J. Remote Sensing*, vol. **9**, no. 4, pp. 669-682.

Kerekes, J.P., and D.A. Landgrebe (1989), "Modeling, Simulation, and Analysis of Optical Remote Sensing Systems", TR-EE 89-49, School of Electrical Engineering, Purdue University, West Lafayette, IN.

Kim, H., P.H. Swain, and J.A. Benediktsson (1986), "A method for combining multispectral and ancillary data in remote sensing and geographic information processing", *Proc. 1986 IEEE Inter. Conf. on S.M.C.*, vol. **2**, pp. 1543-1547.

Koopman, B.O. (1940), "The axioms and algebra of intuitive probability", Ann. Math.,vol. **41**, pp. 269-292.

Lee, T., J.A. Richards, and P.H. Swain (1987), "Probabilistic and evidential approaches to multisource data analysis", *IEEE Trans. Geos. and Remote Sensing*, vol. **GE-25**, pp. 283-293.

Lowrance, J.D., and T.D. Garvey (1982), "Evidential reasoning: a developing concept", *Proc. IEEE Int. Conf. on Cybernetics and Society*, pp. 6-9.

McDermott, D., and J. Doyle (1979), "Non-Monotonic Logic I", *Artificial Intelligence*, vol. **13**, pp. 41-72.

Moon, W.M. (1989), "Integration of remote sensing and geophysical/geological data using Dempster-Shafer approach", *Proc. '89 IGARSS*, Vancouver, B.C., Canada, July 10-14, pp. 838-841.

Papoulis, A. (1984), *Probability, Random Variables, and Stochastic Processes*, 2nd Ed., McGraw-Hill, Inc..

Pednault, E., S. Zucker, and L. Muresan (1981), "On the independence assumption underlying subjective Bayesian updating", *Artificial Intelligence*, vol. **16**, pp. 213-222.

Quarmby, N.A., J.L. Cushnie, and J. Smith (1988), "The use of remote sensing in conjunction with geographic information systems for local planning", *Proc. '88 IGARSS*, Edinburgh, Scotland, Sep. 13-16, pp. 89-92.

Shafer, G. (1976a), *A Mathematical Theory of Evidence*, Princeton University Press.

Shafer, G. (1976b), "A theory of statistical evidence", in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science by* W.L. Harper and Hooker (Eds.), vol. **2**, pp. 365-436, Reidel, Dordrecht.

Shortliffe, E.H. (1976), *Computer-Based Medical Consultations : MYCIN*, American Elsevier.

Smith, C.A.B. (1961), "Consistency in statistical inference and decision", *J. Royal Stat. Soc., Ser. B*, vol. **23**, pp. 1-25.

Strahler, A.H., and N.A. Bryant (1978), "Improving forest cover classification accuracy from Landsat by incorporating topographic information", *Proc. 12th Int. Symp. Remote Sensing of Environment*, Environmental Research Inst. of Michigan, April, pp. 927-942.

Suppes, P. (1974), "The measurement of belief", *J. Roy. Stat. Soc., Ser. B, vol.* **36**, pp. 160-191.

Suppes, P., and M. Zanotti (1977), "On using random relations to generate upper and lower probabilities", *Synthese*, vol. **36**, pp. 427-440.

Swain, P.H. (1985), "Approaches to computer reasoning in remote sensing and geographic information processing: a survey", *Proc. Int. Geoscience and Remote Sensing Symposium*, Univ. of Massachusetts, Amherst, MA.

Swain, P.H., and S.M. Davis, Eds. (1978), *Remote Sensing : The Quantitative Approach*, McGraw-Hill, New York.

Swain, P.H., J.A. Richards, and T. Lee (1985), "Multisource data analysis in remote sensing and geographic information processing", *Proc. 11th Int. Symposium on Machine Processing of Remotely Sensed Data*, pp. 211-217, Purdue Univ., W. Lafayette, IN.

Walley, P., and T.L. Fine (1982), "Towards a frequentist theory of upper and lower probability", *Ann. Stat.*, vol. **10**, pp. 741-761.

Winkler, R.L. (1968), "The consensus of subjective probability distributions", *Management Sci.*, vol. **15**, no. 2, October, pp. B61-B75.

Wolfenson, M. and T.L. Fine (1982), "Bayes-like decision making with upper and lower probabilities", *J. Amer. Stat. Assoc.*, vol. **77**, pp. 80-88.

Zadeh, L.A. (1979), "On the validity of Dempster's rule of combination of evidence", *Memorandum No. UCB/ERL M79/24.*

Zadeh, L.A. (1965), "Fuzzy sets", *Information and Control*, vol. **8**, pp. 338-353.

Zhang, M. and S. Chen (1987), "Evidential reasoning in image understanding", *Uncertainty in Artificial Intelligence*, Workshop, Univ. of Washington, Seattle, July 10-12, pp. 340-346.

purdue university