

PROPOSED ALGORITHMS FOR BAYESIAN  
NETWORKS BASED ON FUZZY SET  
THEORY

By

JONG-NAM LEE

Bachelor of Science  
Seoul National University  
Seoul, Korea  
1977

Master of Business Administration  
State University of New York at Buffalo  
Buffalo, New York  
1981

Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
DOCTOR OF PHILOSOPHY  
July, 1989

Thesis  
1989D  
L478P  
cop. 2

PROPOSED ALGORITHMS FOR BAYESIAN  
NETWORKS BASED ON FUZZY SET  
THEORY

Thesis Approved:

*Marilyn G. Klethe*

Thesis Advisor

*Carl S. ...*

*Wayne A. ...*

*Blayne E. Mayfield*

*Norman N. Busham*

Dean of the Graduate College

## PREFACE

Bayesian networks, encountered in the study of artificial intelligence, are directed acyclic graphs in which nodes represent real life events. Arcs in these graphs signify the existence of direct causal influences between two such events. The strengths of these causal influences are quantified by conditional probabilities.

Pearl (1986) developed two algorithms used to update Bayesian networks in the face of subsequent information. These algorithms permit the use of Bayesian networks as methods of managing uncertainty in expert systems. Pearl applied probabilities in his proposed algorithms. The use of linguistic probabilities provides users of systems based on such probabilities a more meaningful method of system interaction than through the use of probabilities. Consequently, this study develops two algorithms similar to those proposed by Pearl, however, these employ fuzzy set theory rather than probability theory.

I wish to express my sincere gratitude to the individuals who assisted me through my graduate program. In particular, I wish to thank my committee chairperson, Dr. Meg Kletke, for her encouragement and advice. I am also grateful to the other committee member, Dr. Eui-Ho Suh, Dr. Blayne Mayfield, and Dr. Wayne Meinhart for serving on my

graduate committee.

My parents, Nak-Joon Lee and Jung-Suk Suh, encouraged and supported me all the way. Special thanks are due to my parents for their consistent support, moral encouragement, and understanding throughout my graduate program. My wife Myung-Sun Kim and my son Sung-Hwan provide moral support and have motivated me to keep the end goal constantly in sight. Thanks go to my wife and my son for their patience with me throughout my graduate program.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
Expert Systems.....	1
Expert System Structure.....	1
Knowledge Representation Schemes.....	3
Formal Logic.....	4
Semantic Networks.....	6
Frames.....	7
Script.....	8
Production Systems.....	9
Inference.....	10
Dealing with Uncertainty.....	13
Objectives of the Study.....	15
II. EXISTING APPROACHES TO UNCERTAINTY MANAGEMENT IN EXPERT SYSTEMS.....	19
Theory of Endorsement.....	19
Bayesian Approach.....	21
Principle of Maximum Entropy/ Minimum Cross-Entropy.....	25
Bayesian Networks.....	28
Metaprobability Theory.....	29
Certainty Factor Approach.....	30
Original CF Approach.....	32
Revised CF Approach.....	34
Heckerman's Interpretation of CF Approach.....	36
Dempster-Shafer Theory.....	38
Fuzzy Set Theory and Its Applications.....	46
Support Logic Programming.....	53
III. PROPOSED ALGORITHMS FOR BAYESIAN NETWORKS.....	58
Pearl's Algorithms.....	58
Pearl's Algorithm for a Tree Structure.....	59
Pearl's Algorithm for the Network Structure.....	61
Wise's Framework.....	64
Operations on Fuzzy Set and Fuzzy Number...	72

Chapter	Page
Existing Approaches to a Linguistic Approximation.....	77
Wenstop's Approach.....	77
Eshragh and Mamdani's Approach.....	79
Bonissone's Approach.....	83
Parametric Representation of Linguistic Probability.....	89
Proposed Approaches.....	90
Interpretation of Linguistic Probability.....	102
Fuzzy Set Interpretation.....	103
Fuzzy Number Interpretation.....	109
Proposed Simulation Model.....	112
Analysis of Results.....	126
Linguistic Approximation.....	131
Implementation.....	142
IV. SUMMARY AND CONCLUSIONS.....	144
LITERATURE CITED.....	147
APPENDIX-FLOW CHARTS OF THE PROGRAMS.....	157

## LIST OF TABLES

Table	Page
1. Example of a Frame .....	7
2. Example of a Script .....	9
3. Comparison of Existing Numeric Approaches .....	22
4. Listing of Indexes Representing the Importance .....	87
5. Probability Distribution using $C_z$ (= Zadeh's Measure) .....	120
6. Probability Distribution using $C_{DP}$ (= Dubois and Prade's Measure) .....	120
7. Probability Distribution using $C_s$ (= Sugeno's Measure) .....	121
8. Five Element Term Set: $L_1$ .....	133
9. Nine Element Term Set: $L_2$ .....	134
10. Thirteen Element Term Set: $L_3$ .....	135
11. Combination of the Hedges with a Term Set $L_1$ .....	136
12. Combination of the Hedges with a Term Set $L_2$ .....	137
13. Combination of the Hedges with a Term Set $L_3$ .....	138



## LIST OF FIGURES

Figure	Page
1. Structure of an Expert System .....	2
2. Example of a Semantic Network .....	6
3. Example of an Inference Diagram in a Rule-Based System .....	14
4. Example of the Network Structure .....	61
5. Example of Belief Propagation in Bayesian Networks .....	62
6. Basic Experimental Design for Comparisons .....	70
7. Example of a Fuzzy Set .....	73
8. Fuzzy Numbers A and B .....	75
9. Sum of Fuzzy Numbers A and B .....	75
10. Illustration of Wenstop's Approach .....	78
11. S-Type and II-Type .....	80
12. Positive and Negative Skewness .....	84
13. Parametric Representation $[a,b,\alpha,\beta]$ .....	87
14. Example of the Network Diagram .....	95
15. Flow Chart of a Simulation Model .....	158
16. Flow Chart of Proposed Approach to a Linguistic Approximation .....	160

## NOMENCLATURE

ES	expert system
ESs	expert systems
KB	knowledge base
AI	artificial intelligence
CWA	closed world assumption
OWA	open world assumption
iff	if and only if
CF	certainty factor
DST	Dempster-Shafer theory
FST	fuzzy set theory
SLOP	support logic programming
UIS	uncertain inference system
UIS'	uncertain inference systems
ME	principle of maximum entropy
MXE	principle of minimum cross entropy
TMS	truth maintenance system
FD	a frame of discernment
bpa	basic probability assignment

## CHAPTER I

### INTRODUCTION

#### Expert Systems

##### Expert System Structure

Computer-based models and systems for decision support are beginning to play an important role in decision making. An ES which is one of the popular computer-based systems for decision support, is a computer application that solves complicated problems that would otherwise require extensive human expertise. ESs have been used to do a wide variety of things, such as diagnosing and prescribing treatments for certain infectious diseases, configuring new computer installations, and many more.

To do so, an ES simulates the human reasoning process by applying specific knowledge and inferences. Internally, an ideal ES has the following characteristics: extensive specific knowledge from the domain of interest; application of search techniques; support for heuristic analysis; capacity to infer new knowledge from existing knowledge; symbolic processing; and an ability to explain its own reasoning [Rolston, 1988].

Heuristic rules are rules of thumb that suggest the procedures to be followed when invariant procedural rules

are not available. The very presence of heuristics contributes greatly to the power and flexibility of ESs and distinguishes ESs from more traditional software. ESs use a symbolic representation for the relationship between stored information in a knowledge base. Performance of inference and heuristic search in ESs heavily depends on the manipulation of symbols, e.g., strings of characters (i.e., "names"). Thus symbolic processing becomes an important issue in ESs.

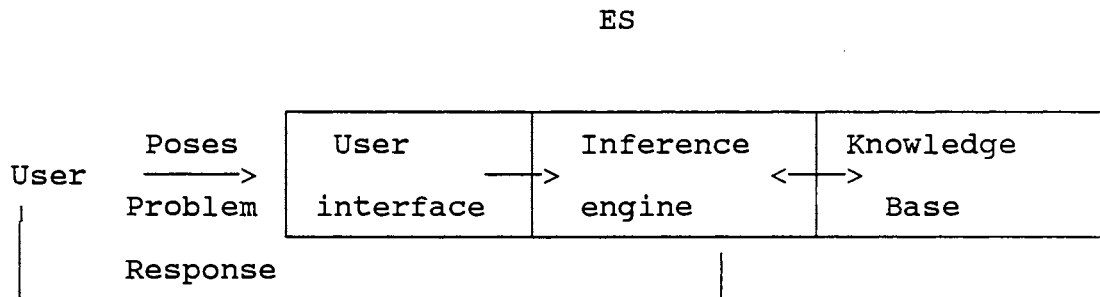


Figure 1. Structure of an ES

A structure of an ES is shown in Figure 1 [Holsapple and Winston, 1987]. When using an ES, a user interacts with the system via its user interface. Then the inference engine which is the problem-solving software, actually carries out the reasoning needed to solve a problem. In

doing so, it uses the knowledge stored in the knowledge base (KB). The KB contains a collection of rules, each of which captures the knowledge about how to reason in a specific problem area addressed by an ES. When the problem is solved, the inference engine reports the solution to the user with the explanation of its line of reasoning in reaching that solution.

A rule is one of the schemes to represent the knowledge, and other schemes include formal logic, frame, semantic net, and script. Formal logic, an outgrowth of early philosophical considerations, was one of the earliest forms of (formal) knowledge representation used in AI, while the nonformal ones are more flexible and widely used schemes for representing knowledge. The knowledge representation schemes is discussed below in detail.

### Knowledge Representation Schemes

A number of knowledge representation schemes for ES have been developed that range from nonformal representation schemes such as semantic net, frame, script, and production system to formal schemes such as a (first-order) predicate logic. A production system, the most commonly used scheme, uses rules for knowledge representation.

## Formal Logic

The (first-order) predicate logic consists of four major components: the "alphabets," a "formal language," a set of basic statements called "axioms," and a set of "inference rules." The "alphabets" consist of "constants," "variables," "functions," "predicates," "connectives," "quantifiers," and "delimiters" such as parentheses and commas.

A "constant" is used to represent a specific element from the domain, where BLUE representing a blue color is an example. A "variable" is used to represent a member of a set of domain elements without specifying a specific element, where "animal" can be a "variable" whose elements include "lion" and "tiger." A "function" describes an element by identifying it as a unique result of the application of a specified mapping between other elements in the domain. For example, "father(JOHN)" which could represent a unique individual who is a father of JOHN, uses "father" as a "function."

The "predicate" is used to represent relation within the domain such that its value is true if the elements in the domain are related in the specified way and false if they are not. BIGGER(TOM,BOB) which could represent the fact "Tom is bigger than Bob," is an example of the predicate. The "variable" can be used as the argument of "predicate." However, the first-order predicate logic does

not allow a "function" or "predicate" to be used as the argument of the "predicate."

The connective is used to combine predicates. There are two types of quantifiers, that is, universal quantifier and existential quantifier, where the universal quantifier is used to assert that a formula consisting of predicates is true for all values of the associated variable and the existential quantifier is used to assert that there exists at least one value such that the associated formula is true.

The "atomic formulas" are individual predicates together with arguments. The "literals" are atomic formulas and negated atomic formulas. The "well-formed formulas (WFFS)" are defined recursively: literals are WFFS; WFFS connected together by the connectives are WFFS; and WFFS surrounded by quantifiers are also WFFS [p.211, Winston,1984].

A "formal language" associated with first-order predicate logic is the set of all formulas that can be legally constructed from the "alphabets." A set of statements, e.g.,  $\text{Feathers}(\text{Squigs})$  and  $[\text{Feathers}(x) \rightarrow \text{Bird}(x)]$  for all  $x$ , are "axioms," where  $\text{Feathers}(\text{Squigs})$  could represent the fact that Squigs has feathers and  $[\text{Feathers}(x) \rightarrow \text{Bird}(x)]$  could represent a rule, i.e., if a creature has the feathers then a creature is a bird.

### Semantic Network

The concept of semantic networks was introduced by Ross Quillian [1968]. It focuses on the graphical representation of relations between elements in a domain, where its basic components are nodes and links. Node represents domain element, while the link represents a (binary) relation between elements. For example, the facts that a horse is a type of a mammal and a tail is a part of a horse can be represented by a semantic networks shown in Figure 2 [Rolston, 1988]:

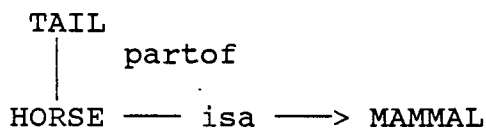


Figure 2. An example of a semantic network

### Frames

Minsky [1975] coined the term "frames" in an attempt to represent knowledge in the context of which many ordinary events or objects appear. A frame is a collection of semantic net nodes and slots that together describe a stereotyped object, act, or event, where each slot



represents a standard property or attribute of the element represented by it [Winston,1984]. A frame that provides a partial description of the class of objects called CAR is shown in Table 1.

TABLE 1  
AN EXAMPLE OF A FRAME

Frame: CAR
Specialization of: LAND VEHICLE
Model:
Range: (sedan, convertible, 2-door, station wagon)
Default: sedan
Body: steel
Windows: glass
Mobility: self-propelled
Mobility mechanism: has wheels
Tires: rubber
Fuel:
Range: (gasoline, diesel, propane)
Default: gasoline
Number of Seats:
Range: (1-9)
Default: none

Inheritance is a very important concept in a frame system. Any given class of objects can be included in several different frames that represent objects at different levels of specification. For example, the class "car" can

be included in the frames named "land vehicle." The "car" frame, for example, inherits the attribute of "mobility mechanism: has wheels" from the fact that it is a specialization of a "land vehicle" frame. The use of inheritance enables the reasoning process to be efficient primarily because we can avoid rediscovering old facts in new situations.

### Script

A script, which is a specialization of the general concept of a frame, is a structure that is used to store the prototypes of expected sequences of events [Schank and Abelson, 1977]. Many different components including entry conditions, script results, props, roles, and scenes can be used to construct a script.

The entry conditions represent conditions that must exist for the script to be applicable and script results represent conditions that will be true after the events in the script have occurred. Props represent slots that represent objects that are involved in a script and roles represent slots that represent agents (e.g., people) that perform actions in a script. Scenes represent specific sequences of events that make up a script [Rolston, 1988]. A script that could represent the process of driving to a theater is shown in Table 2.

TABLE 2  
AN EXAMPLE OF A SCRIPT

Script: TRIP TO THEATER	
	Scene 1: START UP
	* owner finds keys
Props:	* owner unlocks car door
car	* owner starts car
keys	* owner places car in gear
car door	* owner releases parking brake
parking space	Scene 2: DRIVE
Roles:	* owner finds opening in traffic
owner	* owner enters traffic
valet	* owner drives to theater
Entry Conditions:	Scene 3: VALET CONTACT
Owner and car	* owner stops car
at start point	* owner exits car
Results:	* owner gives keys to valet
owner and car	Scene 4: VALET PARKING
at theater	* valet enters car
	* valet finds empty parking space
	* valet enters parking space
	* valet stops car
	* valet sets parking brake
	* valet exits car

### Production Systems

A production system, the most commonly used scheme in ESs, uses rules for the knowledge representation, where a production system consists of: a knowledge base; a rule base; and an inference mechanism [Newell and Simon, 1972]. One of the advantages of a production system consists in storing the knowledge in a uniform and modular form.

This makes each production rule essentially a separate, independent entity that makes adding, deleting, or modifying productions rules very easy. On the other hand, a production system has a disadvantage in that the independence of production rules makes it very difficult to force the execution of a specific sequence of events, even though a specific sequence may be desirable in a certain application.

### Inference

An inference is the process of deriving a conclusion in logic by either induction or deduction. The techniques which are used for inference in a number of knowledge representation schemes, are to be presented. First, inference based on a formal logic will be discussed. One obvious strategy to prove a theorem is to search forward from the axioms using rules of inference such as "modus ponens." "Modus ponens" states that "if  $P_1$  is true, and  $P_1$  being true implies that  $P_2$  is true, then  $P_2$  is true."

One of the greatest advantages of representation using a formal logic is that a syntactic inference is possible and is guaranteed to be valid. A syntactic inference is the inference performed by applying a set of well-defined rules of inference to a set of facts mechanically without a complete understanding of the meanings of facts. However, there is no guarantee that the syntactic inference will

always produce valuable results.

Another strategy is to prove a theorem by showing that the negation of a theorem cannot be TRUE. This strategy is called "proof by refutation" or "proof by contradiction." It adopts a rule of inference called "resolution."

"Resolution" states that "if there is an axiom of the form  $E_1 \vee E_2$ , and there is another axiom of the form  $\sim E_2 \vee E_3$ , then  $E_1 \vee E_3$  logically follows," where ' $\vee$ ' and ' $\sim$ ' denote "union" and "not," respectively.

Second, even if the reasoning based on semantic nets is generally straightforward, the inferences are not guaranteed to be always valid, primarily because it is based on the closed world assumption (CWA). Under the open world assumption (OWA) a theorem is assumed to be false if and only if (iff) it can be proven false, whereas under the CWA a theorem is assumed to be false if no proof of a theorem exists [Gallaire and Minker, 1978].

Third, a frame system has the advantage in that it allows us to reason, to some extent, under the condition where the information available is incomplete, and it allows us to infer facts that are not explicitly observed. This is made possible by a collection of default values, where default values are the expectations regarding an object if none is explicitly provided.

However, one of the difficulties with a frame representation is the problem of establishing the default

values for a frame accurately. This occurs primarily because there is no exact agreement among any group of observers as to the typical characteristics of any object.

Fourth, the reasoning on the basis of a script is straightforward with two steps, that is, the selection of an appropriate script and use of the scenes to infer the existence of unobserved events. However, its reasoning is not reliable for predicting future events on the basis of a scene. For example, in Table 2, the fact that the valet has found a parking space does not necessarily imply that he/she will continue to follow the scene and park the car.

Finally, there are two types of reasoning used in a production system: forward reasoning and backward reasoning. Forward reasoning (or forward chaining) examines each rule in a forward direction, looking first at its premise. When a rule's premise is found to match a theorem to be proved, the rule is fired and the actions in its conclusion are taken. On the other hand, a backward reasoning (or backward chaining) looks first at a rule's conclusion, rather than its premise. Because the match process can identify many matches in a large production system, the process of selecting a specific match to be executed, called conflict resolution, is necessary.

### Dealing With Uncertainty

Uncertainty is present in many real life problems and human experts need to cope with it in decision making. Thus, the ESs which are developed to solve real life problems must also be able to reason under uncertainty. The uncertainty might be present in the knowledge or in the collected data.

For example, the reasoning under uncertainty in a rule-based system should provide answers to the following questions [Berenji, 1987]:

1. How should we combine uncertainties in the premises of a rule (e.g., A, B in the Figure 3)?
  2. How should we propagate this uncertainty to the conclusion of the rule (e.g., combine with the strength of the conclusion, 0.7 in Figure 3)?
  3. If more than one rule results in the same conclusion (e.g., rule 1 and rule 2 in Figure 3), how should we combine them to get an aggregate measure for supporting (or refuting) the conclusion (e.g., conclusion C in Figure 3)?
- In Figure 3, there are two rules resulting in the same conclusion C, where rule 1 states that if premises A and B are true then C is true, and rule 2 states that if premise D is true then C is true.

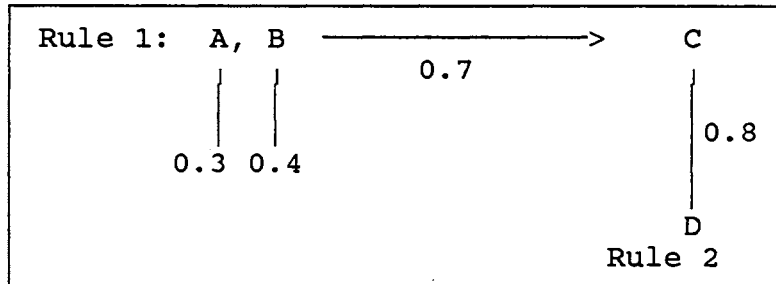


Figure 3. An example of an inference diagram in a rule-based system

The presence of uncertainty in reasoning systems is due to a variety of sources: the "reliability" of information, the inherent "imprecision" of the representation language in which the information is conveyed, the "incompleteness" of information, and the "aggregation" or "summarization" of information from multiple sources [Bonissone and Decker, 1986]. To address these problems, AI researchers have developed a variety of approaches for reasoning under uncertainty.

The existing approaches are divided into two classes: numeric and non-numeric approaches. The numeric approaches represent uncertainty as a precise quantity (scalar or interval) on a given scale. The typical numeric approaches that are currently available are the certainty factor (CF) approach, the Bayesian approach, Dempster-Shafer theory (DST), fuzzy set theory (FST) and support logic programming



(SLOP). On the other hand, the non-numeric approaches do not use a quantity as the representation of uncertainty. Instead, they are based on the idea of dealing with the reasons for believing and disbelieving the specific hypothesis(or event). The typical non-numeric approach is the theory of endorsement [Cohen; 1983a, 1983b].

There have been a great deal of debates between some AI researchers favoring the Bayesian approach and others favoring non-Bayesian numerical approaches such as CF approach, DST, FST, and SLOP. Their basic lines of arguments are as follows: AI researchers favoring the Bayesian approach insist that the Bayesian approach is good enough to handle the uncertainty, while others favoring non-Bayesian approaches insist otherwise.

The non-Bayesian numeric approaches that have drawn a greatest deal of attentions are DST and FST. We shall refer to the non-Bayesian (numerical) approaches to uncertainty management in ESs as uncertain inference systems (UIS').

#### Objectives of the Study

The objective of this study is to develop two algorithms for Bayesian networks based on FST. The first algorithm is for a tree structure, and the second algorithm is for the network. Bayesian (belief) networks are the directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence

of direct causal influences between the linked propositions, and the strengths of these influences are quantified by conditional probabilities [Pearl, 1986a].

Pearl [1982; Kim and Pearl, 1983; 1985a; 1986a] showed that if a network is singly-connected, the probabilities can be updated by a local propagation and the impact of new information is imparted to all propositions in time proportional to the longest path in the network. A singly-connected network is a network such that there exists only one (undirected) path between any pair of nodes.

A singly-connected network includes a tree structure in which each node has at most one parent node and a network structure in which each node is allowed to have multiple parent node(s). A local propagation denotes a propagation which is performed locally.

Our proposed algorithms employ the linguistic probabilities instead of the probabilities. While Pearl's algorithms are such that the belief at the root node is quantified by prior probability and the strengths of the influences between the linked propositions are quantified by conditional probabilities, the proposed algorithms use the linguistic probabilities to represent the belief at the root node and the strengths of the influences between the linked propositions.

Expressions such as "very likely," "quite unlikely," and "maybe" to characterize the degree of likelihood of a

statement are called the linguistic probabilities. The linguistic probabilities are employed in FST. AI researchers favoring FST argued that the probabilities require an unreasonable level of precision, whereas the linguistic values do not.

Wise [1986; Wise and Henrion, 1986] developed a framework for comparing UIS's to the probability theory supplemented by the principle of maximum entropy/minimum cross-entropy (ME/MXE). Wise's framework is justified mainly by the work by Cox [1946] who demonstrated that the axioms of probability are the necessary consequences of the intuitive properties of measures of belief.

Based on the argument that the second-order probability theory capable of handling any input data is not computationally feasible, Wise proposed the use of the (first-order) probability theory supplemented by ME/MXE, instead of the second-order probability theory. The second-order probability distribution deals with the probability measures on the space of first-order (i.e., ordinary) probability distribution.

The simulation model is adopted in the development of new algorithms, where the simulation model is developed based on Wise's framework. The simulation model is used in choosing the adequate operations performed on the linguistic probabilities. A part of the proposed algorithms, that is, a proposed scheme for a linguistic approximation is

implemented in C language.

CHAPTER II  
EXISTING APPROACHES TO UNCERTAINTY MANAGEMENT  
IN EXPERT SYSTEMS

Theory of Endorsement

The typical non-numeric approach is the theory of endorsement [Cohen and Grinberg, 1983a; Cohen and Grinberg, 1983b]. The theory of endorsement may be regarded as the extension of a truth maintenance system (TMS) developed by Doyle [1979]. TMS is a subsystem for performing a problem solving by recording and maintaining the reasons for beliefs. The basic line of arguments for the non-numeric approaches is that the reasoning programs must be able to make assumptions and subsequently revise their beliefs when discoveries contradict these assumptions.

The fundamental problems with numerical approaches are [Cohen and Grinberg, 1983a]: (1) Numerical approaches are not able to treat different kinds of evidence differently. For example, numerical approaches are not able to discriminate eyewitness evidence from circumstantial evidence, because evidence is nothing more than a proposition with an associated number; and (2) Numerical approaches are not able to treat the same evidence

differently in different contexts.

Suppose that a man was convicted on the basis of eyewitness testimony of a crime committed by another man of very similar appearance. Eyewitness testimony relies on some assumptions, one of them could be that individuals have distinctive appearances. If that assumption is violated, the evidence loses its reliability.

TMS stores the justification at each node, where the justification is a record of the nodes on which each node depends. However, TMS makes little differentiation between kinds of justifications, i.e., it is primarily interested in whether a node has the support, not in what kind of support it has. The theory of endorsement was developed hoping that this weakness of TMS can be remedied.

The crux of the theory of endorsement is to deal with the reasons for believing or disbelieving a hypothesis. Endorsements are records of reasons for believing or disbelieving a hypothesis: the reasons for believing being a positive endorsements and the reasons for disbelieving being the negative endorsements [Cohen and Grinberg, 1983a].

This theory has four components: endorsements, heuristics for ranking endorsements, heuristics for propagating endorsements over inference, and heuristics for discounting uncertainty. Endorsements can be propagated over inferences, but in a manner that is sensitive to the context of the inference. This theory was implemented in a program called SOLOMON which does a decision making for

portfolio investment.

Several drawbacks of the theory of endorsement have been pointed out by a number of researchers including Bhatnagar and Kanal [1986]. The main drawback is how to differentiate between two competing hypotheses with different bodies of endorsements and select one over the other. This problem can occur due to the fact that we cannot assign any strict rank ordering to endorsements.

One suggested way to overcome this problem is to weigh endorsements against each other individually by a pairwise ranking, which increases the computational complexity. Cheeseman [1985] who is one of the AI researchers strongly defending the probability theory against non-Bayesian approaches, showed that all the basic ideas of this model can be explained by a deduction in the theory of relative probabilities.

#### Bayesian Approach

Table 3 [Berenji, 1987] summarizes the advantages, disadvantages and applications of the existing numeric approaches to uncertainty management in ESSs. In AI community, there have been a number of debates between researchers favoring the Bayesian approach and those favoring non-Bayesian approaches.

Researchers including Zadeh [1986b] and Shafer [1976] argued that the point-valued probability is inadequate to represent the uncertainty. They also argued that another

TABLE 3  
COMPARISON OF EXISTING APPROACHES

Approach	Advantages	Disadvantages	Applications
Original CF approach	* Fast, simple * Domain experts prefer CF more	* Ad hoc method is not based on a strong theory * inconsistent	*MYCIN, EMYCIN-based and many other expert systems
Revised CF approach	* Improves CF model's CF calculation	* Use a weaker theory than the original CF method	*MYCIN, EMYCIN-based, and many other expert systems
Heckerman's interpretation of CF	* consistent with probability axioms	* is based on strong independence assumption * Seldom applicable	*New, not any applications reported so far
Bayesian probabilistic method	* A sound and formal theory * not too complex to use	* Not expressive enough to easily handle all forms of uncertainty	* PROSPECTOR and many others
Dempster-Shafer theory	* capability to explicitly represent ignorance	* Dempster's rule not applicable to conflicting evidences	* GISTER, and many other
Possibility theory and fuzzy logic	* Linguistic terms could be mathematically represented	* Membership functions are context dependent	* PRUF, REVEAL SPII, RUM, and many others
Support logic programming	* provides a logic based treatment of value terms	* independence assumptions required in most cases	* No application reported so far



problem with the Bayesian approach consists in how to estimate the probabilities provided that we do not have sufficient information. On the other hand, others including Cheeseman, Pearl, Cox, Fox, and Lemmer have strongly defended the Bayesian approach by showing the inadequacy of non-Bayesian approaches.

Zadeh [1986b] who introduced FST in 1965 defended FST by illustrating five examples which, he claims, do not lend themselves to solution by the probability theory. Some of the examples are as follows: (1) An urn contains  $n$  balls of various sizes. Several of the balls are large. What is the probability that a ball drawn at random is large; and (2) Given the proposition "most Swedes are tall," find the fraction of Swedes who are very tall.

Cox [1946] demonstrated that the axioms of probability are the necessary consequences of intuitive properties of measures of belief. That is, if a set of simple properties is assumed, the axioms of the probability theory must be accepted. Similarly, Horvitz et al. [1986] showed that the non-Bayesian approaches do not satisfy some of the intuitive properties of measures of belief. In response to Zadeh [1986b], Cheeseman [1986] showed how we can solve these problems FST claims the probability cannot solve, using the second-order probability theory.

Cheeseman [1985] listed the following misconceptions held by researchers claiming the inadequacy of the probability theory [Cheeseman, 1985]:

- (1) Probability is a frequency ratio.
- (2) Bayesian analysis requires a large amount of data.
- (3) Prior probabilities assume more information than given.
- (4) Numbers are not necessary.
- (5) More than one number is needed to represent uncertainty.
- (6) The Bayesian approach doesn't work.

Fourth misconception is the criticism of non-numeric approaches, and fifth misconception is the criticism of DST. DST employs interval-valued probability to represent the uncertainty.

In particular, he advocated the probability as a measure of belief rather than a frequency ratio, since a frequency interpretation of the probability severely restricts the domain of its applicability. In fact, even among the statisticians, there has been a disagreement on the definition of probability for a long time, where some statisticians favored the definition of probability as a measure of belief while the others favored the definition of probability as a frequency ratio.

The definition of probability as a frequency ratio is as follows [Cheeseman, 1985]:

"The probability of an event(hypothesis) is the ratio of the number of occurrences in which the event is true to the total number of such occurrences."

This definition requires a large total number of such occurrences which may not be possible to obtain in a number of domains, especially medicine. This is one of the

arguments against the probability theory used by researchers favoring non-Bayesian approaches.

On the other hand, the definition of probability as a measure of belief is as follows [Cheeseman, 1985]:

"The (conditional) probability of a proposition given a particular evidence is a real number between zero and one, that is a measure of an entity's belief in that proposition, given the evidence."

Bayes' theorem which is a core of the Bayesian approach, provides a method for updating the prior belief in a hypothesis  $H$ , represented as  $P(H)$ , in light of a new evidence  $E$  to obtain the posterior belief  $P(H|E)$ :

$$P(H_j|E) = \frac{P(H_j) * P(E|H_j)}{\sum_i P(H_i) * P(E|H_i)}$$

Bayes' theorem requires the probabilities  $P(H_i)$  and  $P(E|H_i)$  to be known or estimated in advance. This is one of the major problems with the Bayesian approach, in that it requires a huge amount of statistical data to determine these probabilities.

The Bayesian approach has been implemented in a number of ESs including an ES called PROSPECTOR to determine the major types of ore deposits present in a geological site.

#### Principle of Maximum Entropy/Minimum Cross Entropy (ME/MXE)

ME was first applied by Jaynes [1979] to the statistical mechanics problem of predicting the most likely

state of a system given the physical constraints (e.g., conservation of energy). Jaynes also used this method to provide the prior probabilities for the Bayesian analysis. This method was applied to the problem of finding the best approximation to a given probability distribution based on the knowledge of some of the joint probabilities. This method was employed by an ES called PROSPECTOR.

ME can be used in determining the prior probabilities. Suppose we know that a system has a set of possible states  $X_i$  with unknown probabilities  $P^*(X_i)$ , and you have the information about constraints on the distribution  $P^*$ . Suppose you need to choose a distribution  $P$  that is the best estimate of  $P^*$  given the available information. ME states that, of all the distributions that satisfy the constraints, we should choose the one  $P$  with the largest entropy, where entropy is represented by  $-\sum_i P(X_i) * \log(P(X_i))$  [Shore and Johnson, 1980].

It was argued that ME gives the most unbiased probability estimates given the available evidences [Cheeseman, 1983; Jaynes, 1979; shore and Johnson, 1980]. Cheeseman [1983] introduced a new method for computing the ME probability of an event of interest, given the specific evidence about the related events, and subject to any linear probability constraints. His method was designed to avoid the combinatorial computational time inherent in other methods for computing ME values, without imposing strong

restrictions on the constraints that can be used.

MXE is a generalization that applies in cases when a prior distribution  $P$  that estimates  $P^*$  is known in addition to the constraints. Cross-entropy is defined as a measure of how much information one would have to receive to change one distribution into another. MXE states that, of the distributions that satisfy the constraints, we should choose the distribution  $Q$  with the least cross-entropy, where cross-entropy is defined as  $\sum_i P(X_i) * \log[Q(X_i)/P(X_i)]$ . Minimizing cross-entropy is equivalent to maximizing entropy when the prior probabilities are uniformly distributed. Unlike the entropy maximization, cross-entropy minimization generalizes correctly for continuous probability distribution.

Shore and Johnson [1980] showed that ME/MXE are uniquely correct methods for inductive inference when new information is given in the form of expected values. The form of expected values includes  $\sum_i P^*(X_i) * f_k(X_i) = a$  or  $\sum_i P^*(X_i) * f_k(X_i) < b$ , where  $a$  and  $b$  are constants,  $X_i$  denotes the possible state,  $P^*(X_i)$  denotes the unknown probability, and  $f_k$  denotes the mapping (function) from  $X_i$  to the probability.

However, there have been a number of researchers including Grosf [1986b] criticizing ME in that ME minimizes "information" in the rather specialized sense of Shannon information measure.

## Bayesian Networks

Gordon and Shortliffe [1984, 1985] who developed a CF approach, argued that the application of DST to the domain of MYCIN is more adequate than a CF approach. CF approach is the first non-Bayesian approach developed, where MYCIN is known to be the first ES developed based on a CF approach. They [1984] favored DST because of the drawbacks of Bayesian approach and of a CF approach. CF approach and DST is presented in detail later.

Gordon and Shortliffe [1985] studied the application of DST to a reasoning in a tree-structured hierarchy of hypotheses. Pearl [1986c], who is one of the AI researchers favoring the Bayesian approach, showed that the Bayesian approach performs as well as DST in a tree-structured hierarchy of hypotheses.

Bayesian networks are the directed acyclic graphs in which the nodes represent propositions, the arcs represent the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by the conditional probabilities.

The underlying assumption of Bayesian networks is the conditional independence defined below. A and B are conditionally independent of C iff  $P(A \text{ and } B | C) = P(A | C) * P(B | C)$ . This underlying assumption has been criticized as the major weakness restricting its applicability to AI problems, primarily because this assumption is not

guaranteed to be true in the real world applications.

Pearl [1986a; Kim and Pearl, 1983; 1985a; 1986a] presented the algorithms dealing with propagating the impact of new information through a singly-connected network with the time complexity of  $O(M)$ , where  $M$  = the length of the longest path in the network, in such a way that when equilibrium is reached, each proposition will be assigned a measure of belief consistent with the axioms of the probability theory. A singly-connected network is a network in which there exists only one directed path between two nodes if there exists the causal relationship between two nodes.

#### Metaprobability Theory

Metaprobability theory is a higher order probability theory including the second-order probability theory. The second-order probability theory deals with probability measures on the space of first-order probability distribution which is defined over some domain state. A number of AI researchers including Zadeh criticized first-order (ordinary) probability theory mainly in that it does not represent ignorance.

On the other hand, Fung and Chong [1986] showed that the metaprobability theory can represent the ignorance. They concluded that the metaprobability theory is of practical use and it may perform better than DST in certain

applications. These conclusions were based on the experiment to compare the performance of metaprobability theory with that of DST in updating the beliefs with an evidence. For the readers interested in the details of their experiment, refer to [Fung and Chung, 1986].

Similarly, Wise, who developed a framework for comparing UIS's to the probability theory, argued that the second-order probability theory is capable of handling the generalized input including "imprecise" information which is not adequately dealt with by the first-order probability theory. An example of "imprecise" information could be "it will probably rain tomorrow." Cheeseman [1986] also made a similar argument.

However, the second-order probability theory may not be computationally feasible in certain applications, while it offers all the advantages of the Bayesian approach. In fact, simply forming the first-order distribution to model an expert's belief state may be only marginally feasible, let alone forming a second-order distribution over all possible first-order distributions.

#### Certainty Factor (CF) Approach

The CF approach is the first non-Bayesian approach proposed and developed by Buchanan and Shortliffe [1984a, 1984b, 1984c]. It was implemented in MYCIN which is known to be the first ES to diagnose bacterial infections and



prescribe treatment for them, and also implemented in EMYCIN which is an ES shell developed out of MYCIN. The work on MYCIN was the part of Stanford Heuristic Programming Project which began in 1960's. The programming language used in developing MYCIN is LISP which is a well-known symbolic AI language.

CF represents the measure of the belief update due to the new evidence. The value of CF lies between -1 and 1. The positive (negative) CF value indicates that the evidence confirms (disconfirms) the hypothesis to certain degrees.

Horvitz and Heckerman [1986] emphasized the difference between the measure of absolute belief and measure of belief update. Their criticism of CF consists in that many ESs employing CF approach described CF as the measure of absolute belief at the knowledge acquisition stage, although CF in fact represents the measure of change in belief.

There have been a number of other criticisms of CF approach: (1) Unlike the probability theory, CF approach is an ad hoc approach which is not based on a strong theory; and (2) In its combining schemes of aggregating several evidences, the assumption of maximum correlation between two evidences is implicitly used, although that assumption may not hold in some applications. On the other hand, CF approach is intuitively appealing to practitioners because of its simplicity, thus has been employed by a large number of ESs.

There are three variants of CF approach: the original CF approach, the revised CF approach, and Heckerman's interpretation of the CF approach. After the original CF approach was introduced by Buchanan and Shortliffe [1984b], they [1984c] revised the original CF approach to remedy its weaknesses. Heckerman's interpretation of the CF approach is a result of redefining the revised CF approach to remedy inconsistencies inherent in the revised CF approach.

#### Original CF approach

$CF(H,E)$ ,  $MB(H,E)$ , and  $MD(H,E)$  are defined below, where  $CF(H,E)$  denotes the net change in belief of a hypothesis  $H$  due to a new evidence  $E$ ,  $MB(H,E)$  is an increased belief of a hypothesis  $H$  due to a new evidence  $E$ , and  $MD(H,E)$  is an increased disbelief of a hypothesis  $H$  due to a new evidence  $E$ :

- $CF(H,E) = MB(H,E) - MD(H,E)$
- $MB(H,E) = \begin{cases} 1 & \text{if } P(H)=1 \\ [\max[(P(H|E), P(H)) - P(H)] / [1 - P(H)] & \text{otherwise} \end{cases}$
- $MD(H,E) = \begin{cases} 1 & \text{if } P(H)=0 \\ [P(H) - \min[P(H|E), P(H)]] / P(H) & \text{otherwise} \end{cases}$

Its characteristics are as follows:

- $0 \leq MB(H,E) \leq 1$ ,  $0 \leq MD(H,E) \leq 1$ , and  $-1 \leq CF(H,E) \leq 1$ ;
- $MB(H,E) = 1$ ,  $MD(H,E) = 0$ ,  $CF(H,E) = 1$  if  $H$  is shown to be certain;

- $MB(H,E) = 0$ ,  $MD(H,E) = 1$ ,  $CF(H,E) = -1$  if the negation of  $H$  is shown to be certain;
- $MB(H,E) = 0$  if  $H$  is not confirmed by  $E$ ;
- $MD(H,E) = 0$  if  $H$  is not disconfirmed by  $E$ ;
- $CF(H,E) = 0$  if  $E$  neither confirms nor disconfirms  $H$ .

The combining functions of CF approach are shown below, where  $H$ ,  $H_1$ , and  $H_2$  denote the hypotheses, and  $E$ ,  $E_1$ , and  $E_2$  denote the evidences [Buchanan and Shortliffe, 1984b].

$$MB(H, E_1 \text{ and } E_2) = \begin{cases} 0 & \text{if } MD(H, E_1 \text{ and } E_2) = 1 \\ MB(H, E_1) + [MB(H, E_2) * [1 - MB(H, E_1)]] & \text{otherwise} \end{cases}$$

$$MD(H, E_1 \text{ and } E_2) = \begin{cases} 0 & \text{if } MB(H, E_1 \text{ and } E_2) = 1 \\ MD(H, E_1) + [MD(H, E_2) * [1 - MD(H, E_1)]] & \text{otherwise} \end{cases}$$

- $MB(H_1 \text{ and } H_2, E) = \min(MB[H_1, E], MB[H_2, E])$
- $MD(H_1 \text{ and } H_2, E) = \max(MD[H_1, E], MD[H_2, E])$
- $MB(H_1 \text{ or } H_2, E) = \max(MB[H_1, E], MB[H_2, E])$
- $MD(H_1 \text{ or } H_2, E) = \min(MD[H_1, E], MD[H_2, E])$
- If the truth or falsity of a piece of evidence  $E_1$  is not known with certainty, but a CF (based on prior evidence  $E$ ) denoted by  $CF[E_1, E]$  is known reflecting the degree of belief in  $E_1$ , then if  $MB^*[H, E_1]$  and  $MD^*[H, E_1]$  are the degrees of belief and disbelief in  $H$  when  $E_1$  is known to be true with certainty then the actual degrees of belief and disbelief are given by:

$$MB[H, E_1] = MB^*[H, E_1] * \max(0, CF[E_1, E])$$

$$MD[H, E_1] = MD^*[H, E_1] * \max(0, CF[E_1, E]).$$

Buchanan and Shortliffe [1984b] admitted the problem with the combining scheme to aggregate two evidences associated with a hypothesis H, in that it is built around the intuitive grounds rather than the theoretical grounds. They concluded that CF approach have not avoided many of the problems inherent with the Bayesian approach, but it was an approach such that judgmental knowledge can be efficiently represented and utilized for the modeling of medical decision making, especially in contexts where (a) statistical data are lacking and (b) conditional independence can be assumed.

On the other hand, Adams [1984] showed that a substantial part of the CF model is equivalent to the probability theory with the assumption of independence. His work implies that although the CF approach is the result of attempts to develop an alternative approach to the probability theory, the CF model can be reduced to a special case of the probability theory. To a certain extent, this result may strengthen the argument for the Bayesian approach.

#### Revised CF Approach

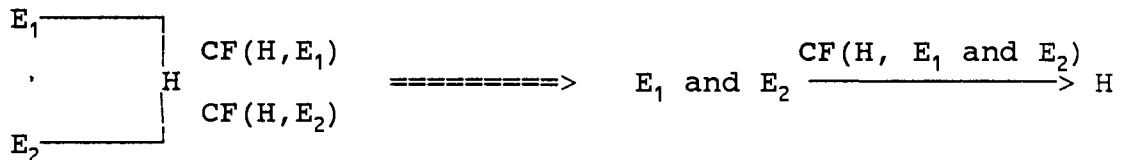
The original CF approach suffers from the following drawbacks [Buchanan and Shortliffe, 1984c]: (1) the potential for a single piece of negative (positive) evidence

to overwhelm several pieces of positive (negative) evidence; and (2) the computational expense of storing both MB's and MD's, rather than cumulative CF's. The revised CF approach was developed to remedy these drawbacks [Buchanan and Shortliffe, 1984c].

The first drawback of the original CF approach can be illustrated in the following example. Consider eight rules all supporting a hypothesis of interest with CF's in the range 0.4 to 0.8. Suppose that the combination of CF's of these eight rules supporting a hypothesis of interest results in a CF of 0.99. Also suppose that there is a single disconfirming rule with CF= 0.8. Then the net support for a hypothesis of interest would be  $CF = MB - MD = 0.999 - 0.8 = 0.1999$ . This result is counterintuitive and also occasionally led MYCIN to reach incorrect inferences, especially where the absolute value of final CF less than, say, 0.2 is eliminated from further consideration. In MYCIN, in order to make inferences efficient, a final belief below the established threshold, e.g., 0.2, is eliminated from further consideration.

In the revised CF approach, the definition of CF is unchanged for any single piece of evidence and that the combining function is unchanged when both CF's are of the same sign. The change occurs only when two CF's of opposite sign are combined. The definition of revised CF is:  $CF = (MB - MD) / [1 - \min(MB, MD)]$

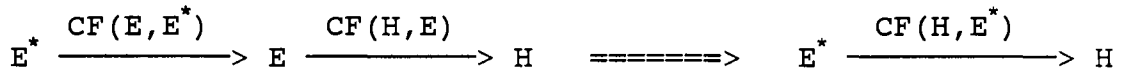
The propagation of uncertainty is accomplished by the repeated applications of two combination schemes, that is, parallel combination and sequential combination [Heckerman, 1986]. The parallel combination scheme is as follows:



Let  $X = \text{CF}(H, E_1)$ ,  $Y = \text{CF}(H, E_2)$ , and  $Z = \text{CF}(H, E_1 \text{ and } E_2)$ .

$$Z = \begin{cases}
 X + Y - X * Y & X, Y \geq 0 \\
 (X + Y) / [1 - \min(|X|, |Y|)] & X, Y \text{ of opposite sign} \\
 X + Y + X * Y & X, Y < 0
 \end{cases}$$

The sequential combination scheme is as follows:



Let  $W = \text{CF}(E, E^*)$ ,  $X = \text{CF}(H, E)$ ,  $Y = \text{CF}(H, \text{not } E)$ , and  $Z = \text{CF}(H, E^*)$ .

$$Z = \begin{cases}
 W * X & W \geq 0 \\
 -W * Y & W < 0
 \end{cases}$$

### Heckerman's Interpretation of CF Approach

Heckerman [1986] redefined CF in order to eliminate the inconsistency between the definition of the revised CF and its combining schemes. The redefined CF is called Heckerman's interpretation of the CF approach. Heckerman argued that it is inappropriate to regard the combining schemes as approximated combination rules for the CF in that

there are inconsistencies between the definition of CF and the combining schemes.

Consider a hypothesis  $H$  and an evidence  $E$ . Let  $P(H)$  and  $P(H|E)$  denote the prior, and posterior belief (or probability) in  $H$ , respectively. Since  $CF(H,E)$  is a measure of the change in belief in  $H$  given  $E$ , it is reasonable to expect that there is some function  $f$  such that  $P(H|E) = f(CF(H,E), P(H))$ . After a number of manipulations on Bayes' theorem, we can obtain the following equation:

$$O(H|E,e) = \frac{P(E|H,e)}{P(E|\bar{H},e)} O(H|e) \quad (1)$$

where  $O(X) = P(X) / (1 - P(X))$ ,  $H$  = a hypothesis,  $e$  = prior evidence, and  $E$  = new evidence.

The ratio in equation (1) is called a "likelihood ratio." An equation (1) can be written as  $O(H|E,e) = L(H,E,e) * O(H|e)$ , where  $L(H,E,e)$  denotes a likelihood ratio. The  $L(H,E,e)$  represent a belief update, thus can be considered as a potential probabilistic interpretation for CF. The only difficulty with  $L$  is that it ranges from 0 to infinity ( $\infty$ ) rather than from -1 to 1. We can resolve this problem easily by setting  $CF(H,E) = F(L(H,E,e))$  where  $F$  is some function which maps  $L$  into  $[-1,1]$ .

One possible choice for the function  $F$  is as follows:

$$F_1(X) = \begin{cases} (X-1)/X & X \geq 1 \\ X-1 & X < 1 \end{cases}$$

This function generates the following probabilistic

interpretation for CF, after Bayes' theorem is applied:

$$CF_1 = \begin{cases} [P(H|E) - P(H)] / [P(H|E) * (1 - P(H))] & \text{for } P(H|E) > P(H) \\ [P(H|E) - P(H)] / [P(H) * (1 - P(H|E))] & \text{for } P(H) > P(H|E) \end{cases}$$

It can be shown that any monotonically increasing function  $F$  which satisfies  $F(1/X) = -F(X)$  and  $F(\infty) = 1$  generates a valid probabilistic interpretation for a CF. A function  $F$  is called a monotonically increasing function if and only if for  $x_1 \leq x_2$ ,  $F(x_1) \leq F(x_2)$ . This implies that this redefinition accommodates an unlimited number of probabilistic interpretations for the CF. Heckerman's work consolidated an argument for the probability theory by demonstrating a clear relationship between the CF and the probability theory.

#### Dempster-Shafer Theory (DST)

DST was introduced by Shafer [1976] as an extension of the work of Arthur Dempster [1967] in the probability theory. As discussed earlier, even Gordon and Shortliffe [1984] who developed the CF approach defended DST as an alternative approach to the Bayesian approach and the CF approach. Furthermore, they suggested that DST is more appropriate than the Bayesian approach and the CF approach in especially medical domain due to its ability to model the narrowing of the hypothesis set with the accumulation of evidences.



Suppose a physician is considering a case of cholestatic jaundice for which there is a diagnostic hypothesis set of hepatitis(hep), cirrhosis(cirr), gallstone(gall) and pancreatic cancer(pan). In DST, this set is called "a frame of discernment", denoted FD. Each hypothesis in FD corresponds to a one-element subset called a "singleton." The hypotheses in FD are assumed mutually exclusive and exhaustive. Subsets  $A_1, A_2, \dots, A_n$  are defined to be mutually exclusive if  $A_i \cap A_j = \emptyset$  for every  $i \neq j$ . Subsets  $A_1, A_2, \dots, A_n$  are defined to be mutually exhaustive if  $A_1 \cup A_2 \cup \dots \cup A_n = W$ , where  $W$  denotes the whole set.

DST uses a number in the range  $[0,1]$  to indicate the belief in a hypothesis given a piece of evidence. The impact of each distinct piece of evidence on the subsets of FD is represented by a function called a "basic probability assignment (bpa)" which is a generalization of the traditional probability density function.

The "bpa," denoted "m," assigns a number in the range  $[0,1]$  to all subsets of FD such that the numbers sum to 1, whereas the traditional probability density function assigns a number in the range  $[0,1]$  to every "singleton" of FD such that the numbers sum to 1.

A belief function denoted  $Bel$ , assigns to the subset  $A$  of FD the sum of the beliefs committed exactly to every subset of  $A$ . The quantity  $1 - Bel(A^c)$  expresses the plausibility of  $A$ , i.e., the extent to which the evidence

allows one to fail to doubt A, where 'c' denotes the complement. Thus, the information contained in Bel concerning a given subset A may be expressed by the interval  $[\text{Bel}(A), 1-\text{Bel}(A^c)]$ . We can argue that DST allows the belief to be expressed as the interval, contrary to the probability theory in which a single-valued probability is assigned to each event.

In the Bayesian approach,  $\text{Bel}(A) + \text{Bel}(A^c) = 1$ , thus the width of the interval  $[\text{Bel}(A), 1-\text{Bel}(A^c)]$  becomes 0. On the other hand, in DST, the width of the interval is usually not 0 and can be regarded as a measure of the belief that is committed to neither the hypothesis A nor the negation of the hypothesis A, i.e., a measure of ignorance. The fact that DST allows the explicit representation of ignorance unlike the Bayesian approach is regarded as one of the major advantages of DST over the Bayesian approach.

The scheme to combine multiple evidences is presented below [Buchanan and Shortliffe, 1984c]. Given two belief functions, based on two observations, but with the same frame of discernment, Dempster's (combination) rule computes a new belief function, where  $\text{mass}_1$ ,  $\text{mass}_2$  and  $\text{mass}_3$  represent bpa's.

$$A, B, C \in \text{FD}, \text{mass}_3(C) = (1-K)^{-1} \sum_{A \cap B = C} \text{mass}_1(A) * \text{mass}_2(B)$$

$$\text{where } K = \sum_{A \cap B = \phi} \text{mass}_1(A) * \text{mass}_2(B) < 1$$

After Dempster's combining scheme is applied, it is

possible for the empty set to have the positive bpa. Dempster's rule states that this problem can be remedied by the normalization of assigned values. The normalization procedure reassigns the bpa's which are assigned to null sets, to the non-empty sets. However, the normalization procedure used in Dempster's rule has been criticized by several AI researchers including Zadeh [1986a] in that it causes inconsistency.

Zadeh [1986a] viewed DST as the application of a retrieval technique to the second-order relations in the context of a relational database. The second-order relation is defined as the relation in which the data entries are relations in first normal form. A relation is said to be in first normal form iff it satisfies the constraint that it contains atomic values only.

Consider a relation EMP3:(Name, Age(car)), where the query is to determine the fraction of employees who have the cars whose ages are between 2 and 4. Note that a relation EMP3 is the second-order relation. If the normalization is performed, then we get the conclusion that all employees have cars that are two to four years old. However, if the normalization is not performed, the conclusion we get is that 2 employees out of 5 have a car that is two to four years old. Apparently, the conclusion obtained after the normalization is misleading due to the fact that the normalization eliminates the null values from consideration.

Hunter [1987] also pointed out the problem with a normalization of Dempster's rule by comparing it to a probabilistic logic.

EMP3:	<u>Name</u>	<u>Age(car)</u>	
	1	[3,4]	
	2	-	
	3	[2,3]	Note: '-' denotes the null value
	4	-	indicating the employee has no
	5	-	car.

Garvey et al. [1981] developed the inference rules for DST. His inference rules are as presented below, where the statement above the line in each rule allows the statement below the line to be inferred.

$$(1) \quad \frac{}{\text{FD: } [1,1]}$$

$$(2) \quad \frac{A \in \text{FD}}{\text{A: } [0,1]}$$

$$(3) \quad \frac{\begin{array}{l} \text{A: } [S_1(A), P_1(A)] \\ \text{A: } [S_2(A), P_2(A)] \end{array}}{\text{A: } [S(A), P(A)], \text{ where } S(A) = \max[S_1(A), S_2(A)] \\ P(A) = \min[P_1(A), P_2(A)]}$$

$$(4) \quad \frac{\text{A: } [S(A), P(A)]}{\sim\text{A: } [S(\sim A), P(\sim A)], \text{ where } S(\sim A) = 1 - P(A), P(\sim A) = 1 - S(A), \\ \text{and } \sim = \text{not}}$$

- (5) 
$$\begin{array}{l} \text{A: } [S(A), P(A)] \\ \text{B: } [S(B), P(B)] \end{array}$$

---

$$\text{A U B: } [S(A \cup B), P(A \cup B)],$$
$$\text{where } S(A \cup B) = \max[S(A), S(B)]$$
$$P(A \cup B) = \min[1, P(A) + P(B)]$$
- (6) 
$$\begin{array}{l} \text{A U B: } [S(A \cup B), P(A \cup B)] \\ \text{A: } [S(A), P(A)] \end{array}$$

---

$$\text{B: } [S(B), P(B)], \text{ where } S(B) = \max[0, S(A \cup B) - P(A)]$$
$$P(B) = P(A \cup B)$$
- (7) 
$$\begin{array}{l} \text{A: } [S(A), P(A)] \\ \text{B: } [S(B), P(B)] \end{array}$$

---

$$\text{A } \cap \text{ B: } [S(A \cap B), P(A \cap B)]$$
$$\text{where } S(A \cap B) = \max[0, S(A) + S(B) - 1]$$
$$P(A \cap B) = \min[P(A), P(B)]$$
- (8) 
$$\begin{array}{l} \text{A } \cap \text{ B: } [S(A \cap B), P(A \cap B)] \\ \text{A: } [S(A), P(A)] \end{array}$$

---

$$\text{B: } [S(B), P(B)]$$
$$\text{where } S(B) = S(A \cap B)$$
$$P(B) = \min[1, 1 + P(A \cap B) - S(A)].$$

The advantages of DST are: (1) DST is able to model the narrowing of the hypothesis set with the accumulation of evidence; (2) DST allows the explicit representation of ignorance which is committed neither to hypothesis nor to the negation of hypothesis; (3) DST is based on a relatively firm mathematical foundation, especially compared to an ad hoc approach such as the CF approach; and (4) If the arcs of semantic nets have the associated confidences which is represented as the interval-valued probabilities, then

Dempster's rule can be used to combine them for non-monotonic reasoning system [Ginsberg, 1984].

The monotonic and non-monotonic reasoning system are presented below. A formal logic leads to a monotonic reasoning system. In a monotonic reasoning system, once the truth value of any predicate becomes "true", it remains "true." However, these characteristics of a monotonic reasoning system have the limitation in that it cannot be applied in the real world for the reasons including one that available information is frequently incomplete, at any given decision point. In dealing with these difficulties, human problem solvers often use the beliefs that are subject to change given further information.

Contrary to a monotonic reasoning system, a non-monotonic reasoning system tracks a set of tentative beliefs and revises those beliefs when new knowledge is observed. TMS [Doyle, 1979] and the theory of endorsement [Cohen and Grinberg, 1983a] can be regarded as the non-monotonic reasoning systems.

The disadvantages of DST are summarized as follows:

(1) Its underlying assumption that a frame of discernment is assumed to be mutually exclusive and exhaustive, leads to an exponential-time requirements which makes it intractable computationally; (2) Dempster's rule requires that the bodies of evidence to be combined be independent and from the same frame of discernment; (3) Dempster's rule is not

applicable to the conflicting evidences; and (4) The normalization of Dempster's rule leads to counter-intuitive results.

Suppose that the number of elements comprising a frame of discernment is 20, which may not even be of a reasonable size in real world applications. The number of subsets becomes 1048576 ( $=2^{20}$ ) to which the bpa's are assigned. If we consider the fact that when the evidences are combined, bpa's assigned to all subsets need to be updated, we can see a huge computational time requirement.

Thus, there have been a number attempts to reduce this exponential time complexity: Barnett's algorithm [1981], and Gordon and Shortliffe's algorithm [1985]. First, Barnett [1981] showed that an exponential-time requirement of DST can be reduced to simply a polynomial time if DST is applied to single hypotheses and to their negations, and if evidences are combined in a specified order. However, the condition that the evidences are applied to single hypotheses and to their negations may be too strong to hold in the real world applications.

Second, Gordon and Shortliffe [1985] developed a new method, namely, a variant of Barnett's method that achieves the computational efficiency while permitting the management of evidential reasoning in a tree-structured hierarchical hypothesis space. Their algorithm is based on the pruning of a tree which reduces the computational complexity.

## Fuzzy Set Theory (FST) and Its Applications

FST was introduced by Zadeh [1965] in 1965. It is a mathematical generalization of the (ordinary) set theory. Unlike the conventional set theory which requires full membership or non-membership for the elements of a universe of discourse in a set, a fuzzy set allow partial membership.

Because the ordinary set  $A=\{1,2,3,4\}$  does not contain the element 6, its membership value is false (or 0). However, in a fuzzy set "tall"=  $\{1/6'10", 0.9/6'1", 0.6/5'10"\}$ , where the first element of a fuzzy set pair denotes the membership value and its second element denotes the height, for example, the membership of the height 6'1" is 0.9.

There have been a number of studies to compare FST to the Bayesian approach in terms of performance. First, Stalling [1977] applied both FST and the Bayesian approach to a syntactic pattern recognition of handwritten capitals, and concluded that the Bayesian approach offers the computational and philosophical advantages over FST. Second, Maier and Sherif [1985] demonstrated that FST is applicable to a wide range of industrial controller problems and a simple fuzzy control algorithm performs nearly as well as the probability-based control algorithm.

Finally, Tribus [1980] applied FST and the probability theory to the problem of literature search, and concluded that there is no significant difference to favor one over



the other. We can conclude that the experimental results are mixed or inconclusive.

AI researchers favoring FST argue that the Bayesian approach is not appropriate to handle especially "imprecise" information, e.g., "it will probably rain tomorrow." On the other hand, AI researchers favoring the Bayesian approach argued that FST is context-dependent and since the second-order probability theory can handle "imprecise" information, FST is unnecessary. For instance, the fuzzy set "young" =  $\{1/18, 0.9/20, 0.6/30\}$ , could be defined in the context of college students, whereas it could be defined differently in the context of elementary school students.

We could argue that the choice between FST and the Bayesian approach may be dependent upon the advantages of approaches in terms of the psychological accessibility of probabilistic information in different formats. This implies that FST and the Bayesian approach perform better under different sets of conditions.

A fuzzy variable is defined as follows [Zadeh, 1975a]: A fuzzy variable is characterized by a triple  $(X, U, R(X:u))$  in which  $X$  is the name of the variable;  $U$  is the universe of discourse (finite or infinite set);  $u$  is a generic name for the elements of  $U$ ; and  $R(X:u)$  is a fuzzy subset of  $U$  which represent a fuzzy restriction on the values of  $u$  imposed by  $X$ . Some examples of a fuzzy variable are "young," "old," and "not young and not old." Like the conventional (or

nonfuzzy) variable, the marginal restriction (analogous to the marginal distribution), conditional restriction (analogous to the conditional distribution), noninteraction (analogous to independence) and intersection (analogous to dependence) are defined for a fuzzy variable.

A possibility theory was developed by Zadeh [1978a] as an extension of FST. It focuses primarily on imprecision which is intrinsic in the natural languages. He argued that a language is possibilistic rather than probabilistic. A possibility distribution function associated with  $X$  is defined to be numerically equal to the membership function of a fuzzy variable. Suppose that a fuzzy set "small integer" be defined as  $\{1/1, 1/2, 0.8/3, 0.6/4, 0.4/5\}$ . The proposition "X is a small integer" associates with  $X$  a possibility distribution  $\{1/1, 1/2, 0.8/3, 0.6/4, 0.4/5\}$ .

There is a distinctive difference between a possibility distribution and the probability distribution. If we consider the statement "Hans ate  $X$  eggs for breakfast," [Zadeh, 1978a] then a possibility distribution with  $X$  is interpreted as the degree of ease with which Hans can eat  $X$  eggs, whereas the probability distribution with  $X$  is the probability that Hans ate  $X$  eggs. However, there is a weak relationship between the probability distribution and a possibility distribution. The relationships between the probability distribution and a possibility distribution are to be discussed in detail in Chapter 3.

Because a possibility distribution is analogous to the probability distribution to some extent, the concepts analogous to a multivariate probability distribution, a marginal probability distribution, and the conditional probability distribution are defined in a possibility theory. They are called n-ary possibility distribution, a marginal possibility distribution, and the conditional possibility distribution, respectively.

A linguistic variable is defined as a variable of a higher order than a fuzzy variable, in the sense that a linguistic variable takes fuzzy variables as its values [Zadeh, 1975a]. For example, a linguistic variable "age" can take the values such as "young," "very young," and "old" which are fuzzy variables. A variable taking the linguistic truth values as its values is called a linguistic truth variable, where the expressions such as "very true," "quite true," and "completely false" to characterize the degree of truth of a statement are called linguistic truth values [Zadeh, 1975a].

A linguistic (truth) value is defined as a fuzzy subset of the interval  $[0,1]$ . For instance, a linguistic truth value "true" could be defined as  $\{0.5/0.7, 0.7/0.8, 0.9/0.9, 1/1\}$ . A linguistic truth variable is a special case of a linguistic variable in that a linguistic truth variable takes only linguistic truth values as its values. Zadeh [1975a] defined a number of operations on linguistic truth

values including a negation, a conjunction, a disjunction, imply, modus ponens, and a generalized modus ponens.

Modus ponens is the basic rule of inference in the traditional logic such that we can infer the truth of a proposition B from the truth of A and the implication  $A \Rightarrow B$ . Similarly, given "u is more or less small" and "IF u is small THEN v is large," a generalized modus ponens could lead to the conclusion "v is more or less large."

Like the conventional probability theory, a linguistic probability (analogous to the conventional probability), a linguistic random variable (analogous to a random variable in the probability theory), and a linguistic probability distribution (analogous to the conventional probability distribution) are defined for a linguistic variable.

Treating truth as a linguistic variable leads to a fuzzy (linguistic) logic that provides a basis for approximate reasoning. That is, approximate reasoning is a mode of reasoning in which the truth-values and the rules of inference are fuzzy rather than precise. For example, given two facts, "most students are undergraduates" and "most undergraduates are young," the answer to the question "how many students are young?" can be obtained from a fuzzy logic.

Since FST was introduced by Zadeh in 1965, it has been applied to a variety of areas including decision making problems in management, engineering, and even in

mathematics. Mamdani [1976] surveyed the field of applications of a fuzzy logic in the synthesis of controllers for dynamic plants and concluded that the application performs well as expected. Kuang [1986a,1986b,1986c] showed the application of FST to hydraulic systems diagnostics and troubleshooting. Adamo [1980a] introduced a fuzzy decision tree method that is an extension to a decision tree in which the involved data (probabilities, costs, profits, losses) are represented as linguistic values.

Zimmermann [1986] and Prade [1980] discussed the application of FST to the mathematical programming models. In general, the mathematical programming models can be written as follows:

$$\begin{aligned} & \text{Maximize } f(X) \\ & \text{s.t. } g_1(X) \geq 0 \end{aligned}$$

For example, LP model can be written as follows:

$$\begin{aligned} & \text{Maximize } Z = C^T * X \\ & \text{s.t. } A * X \leq b \\ & \quad X \geq 0 \end{aligned}$$

where  $C$  and  $X$  are row-vectors of dimension  $n$ ,  $b$  is a column-vector of dimension  $m$ , and  $A$  is an  $m \times n$  matrix.

Unlike the conventional mathematical programming models in which the elements of  $A, b, C$  are crisp numbers, a fuzzy mathematical programming model permits the elements of  $A, b, C$  to be fuzzy numbers. Crisp number is the term to denote the ordinary (non-fuzzy) number and used to distinguish it from

a fuzzy number. A fuzzy number is defined as a fuzzy subset of real numbers, where "small," "approximately 8," "very close to 5" are its examples. Zimmermann discussed the fuzzy linear and nonlinear programming problems and algorithms to determine the optimal solution.

The algorithms for fuzzy mathematical programming problems are primarily built around the algorithms for (ordinary) mathematical programming problems. Prade [1980] argued that the adaption of an ordinary algorithm to a fuzzy mathematical programming problem is not always straightforward. On the other hand, although the direct application of FST can solve the problem, it is not generally computationally attractive. Prade also discussed PERT, assignment problem, a traveling salesman problem, and a transportation problem.

There have been a number of FST applications to mathematics. Trillas and Riera [1978] introduced the general types of entropies for fuzzy sets. Kim and Roush [1980] developed a fuzzy matrix theory. A fuzzy matrix is defined as a matrix whose entries are the values in  $[0,1]$ , where its example is shown below.

$$A = \begin{bmatrix} 0.8 & 0.1 \\ 0.2 & 1 \end{bmatrix}$$

Several languages based on FST including L.P.L. language (linguistic oriented programming language) and PRUF

(Possibilistic Relational Universal Fuzzy), have been developed. PRUF is a meaning representation language for natural language processing. Some of the characteristics of PRUF are [Zadeh, 1978b]: (1) A basic assumption underlying PRUF is that the imprecision that is intrinsic in natural languages is, possibilistic rather than probabilistic; (2) The logic underlying PRUF is a fuzzy logic; and (3) The quantifiers in PRUF are allowed to be linguistic such as "most," "many," "few."

Fuzzy quantifiers denote the collection of quantifiers in natural languages whose representative elements are: "most," "many," "quite a few," and "frequently." [Kandel, 1986]. Zadeh [1983b] introduced the computational approach to fuzzy quantifiers, in which quantifiers are treated as fuzzy numbers. L.P.L. language developed by Adamo [1980b,1980c] has the characteristics similar to those of PRUF.

#### Support Logic Programming (SLOP)

FST has a major advantage over the Bayesian approach in that it is capable of handling the "imprecise" information, while DST has a number of advantages, one of which is that it allows the interval-valued belief instead of a single-valued belief (or probability). Baldwin [1986] developed a programming language called SLOP which is similar to Prolog except that it employs the main features of FST and DST.

Prolog is one of the well-known AI languages based on first order logic.

SLOP generalizes a logic programming to the case where uncertainties, either of a probability or fuzzy nature, could be modeled. SLOP is a Prolog-like programming system in which uncertainties associated with facts and rules are represented by a pair of support factors  $[S_n.S_p]$ , where  $S_n$  and  $S_p$  are called the necessary support and the possible support, respectively. The necessary support can be viewed as a lower bound, whereas a possible support can be viewed as an upper bound.

The following is an example of SLOP:

```
design(X,good) :- eng_report(X,satisfactory),
reliability(X,high): [0.9,1].
```

This rule states that if the engineering report about design X is satisfactory and its reliability is high, then the design is considered to be good with the necessary support of .9 and a possible support of 1.

Unlike Prolog, fuzzy predicates such as "good" and "high" are allowed. Also, in contrast to Prolog in which the degree of belief attached to a rule always equals 1, the degree of belief associated with a rule is expressed as the interval-valued number like DST.

Furthermore, in contrast to an ordinary logic and PROLOG, SLOP does not rely upon a CWA with regard to the knowledge representation. In other words, it is not assumed



that facts which are not in the data are necessarily false. A fact not present in the knowledge base has the necessary support of 0 and a possible support of 1. The general forms of SLOP are as follows:

$A :- B_1, B_2, \dots, B_n : [S_1, S_2]$ , where  $S_1 =$  necessary support and  $S_2 =$  possible support

$A : [S_1, S_2]$ .  $[S_1, S_2]$  is equivalent to the interval  $[\text{Bel}(A), 1 - \text{Bel}(\sim A)]$  of DST, where  $S_2 - S_1$  is the measure of the ignorance in support of the rule.

The interpretation of a negation is as follows:

If  $P :- Q : [S_1, S_2]$ , then  $\sim P :- Q : [S_1^*, S_2^*]$ , where  $S_1^* = 1 - S_2$  and  $S_2^* = 1 - S_1$ .

Similarly, if  $P : [S_1, S_2]$ , then  $\sim P : [S_1^*, S_2^*]$ , where  $S_1^* = 1 - S_2$  and  $S_2^* = 1 - S_1$ .

The SLOP calculus allows different models for combining evidences, that is, a multiplication model and min model:

(Case 1)  $X : [S_1(X), S_2(X)]$ ,  $Y : [S_1(Y), S_2(Y)]$ ,  $X$  and  $Y : [S_1(X \text{ and } Y), S_2(X \text{ and } Y)]$ , and  $X$  or  $Y : [S_1(X \text{ or } Y), S_2(X \text{ or } Y)]$

(1) multiplication model:  $S_1(X \text{ and } Y) = S_1(X) * S_1(Y)$

$$S_2(X \text{ and } Y) = S_2(X) * S_2(Y)$$

$$S_1(X \text{ or } Y) = S_1(X) + S_1(Y) - S_1(X) * S_1(Y)$$

$$S_2(X \text{ or } Y) = S_2(X) + S_2(Y) - S_2(X) * S_2(Y)$$

(2) min model:  $S_1(X \text{ and } Y) = S_1(X) \wedge S_1(Y)$

$$S_1(X \text{ or } Y) = S_1(X) \vee S_1(Y)$$

$$S_1((\text{not } A) \text{ or } (\text{not } B)) = S_1(\text{not } A) \vee S_1(\text{not } B)$$

$$S_2(X \text{ and } Y) = S_2(X) \wedge S_2(Y)$$

$S_2(X \text{ or } Y) = S_2(X) \vee S_2(Y)$ , where '^' denotes the minimum and 'v' denotes the maximum

(Case 2)  $P : \neg Q : [S_1(P|Q), S_2(P|Q)]$ ,  $Q : [S_1(Q), S_2(Q)]$  and  $P : [S_1(P), S_2(P)]$

(1) multiplication model:  $S_1(P) = S_1(P|Q) * S_1(Q)$

$$S_2(P) = 1 - [1 - S_2(P|Q) * S_1(Q)]$$

(2) min model: no general formulae available

(Case 3)  $P : [S_1(P), S_2(P)]$ ,  $Q : [S_1(Q), S_2(Q)]$  and  $P \Rightarrow Q :$

$[S_1(P \Rightarrow Q), S_2(P \Rightarrow Q)]$

(1) multiplication model:  $S_1(P \Rightarrow Q) = S_1(P|Q) * S_1(Q)$

$$S_2(P \Rightarrow Q) = 1 - [1 - S_2(P|Q) * S_1(Q)]$$

(2) min model: no general formulae available, where " $\Rightarrow$ "  $\equiv$  imply

(Case 4)  $P : [S_1, U_1]$  and  $P : [S_2, U_2]$  and  $P : [S, U]$

(1) multiplication model:  $S = (S_1 * U_2 + S_2 * U_1 - S_1 * S_2) / K$

$$U = [(1 - U_1) * (1 - S_2) + (1 - U_2) * (U_1 - S_1)] / K$$

$$K = 1 - S_2 * (1 - U_1) - S_1 * (1 - U_2)$$

(2) min model:

(if there is a conflict):  $S = S_1 \vee S_2$

$$U = 1 - [(1 - S_1(\text{not } P)) \vee (1 - S_2(\text{not } P))]$$

(if there is not conflict): identical to Case 1

SLOP could be used in conjunction with FRIL (Fuzzy

Relational Inference Language). FRIL, developed by Baldwin and Zhou [1984], is a query language similar to ordinary query languages such as SQL and INGRESS except that FRIL can access fuzzy base relations. A relation "Person\_height\_1" shown below is a fuzzy base relation.

The typical characteristics of a fuzzy relation are:  
 (1) A fuzzy variable such as "tall" or "not tall" or "very tall" is allowed as the legal value of attribute, e.g., in "Person\_height\_1" and (2) the additional attribute "membership" is needed in "Person\_height\_1."

Person_height_1:	<u>Name</u>	<u>Height</u>	<u>membership</u>
	Adrian	\$Tall	1
	Bill	not \$Tall	1
	Lofti	5-10	1
	Laurie	\$Very tall	1

## CHAPTER III

### PROPOSED ALGORITHMS FOR BAYESIAN NETWORKS

#### Pearl's Algorithms

As defined earlier, Bayesian networks are directed acyclic graphs with each node representing a proposition (or variable), each arc signifying the existence of direct causal influence between the linked propositions, and the strengths of these influences quantified by conditional probabilities [Pearl, 1986a]. Pearl[1982; Kim and Pearl, 1983; 1985a; 1986a] developed the algorithms (for the singly-connected network) to impart the impact of new information to all nodes by local propagation in time proportional to the longest path in the network.

The singly-connected network in which there exists only one (undirected) path between any pair of nodes, includes a tree structure in which each node is allowed to have at most one parent node and a network structure in which each node is allowed to have multiple parent nodes. In this section, Pearl's algorithm for a tree structure is described first, followed by the description of an algorithm for the network structure.

### Pearl's Algorithm for a Tree Structure

In a tree structure, each node  $B$  stores two parameters,  $\pi(B)$  and  $\delta(B)$ , where  $\pi(B)$  is the support attributed to node  $B$  by its parent node and  $\delta(B)$  is the support node  $B$  receives from its child node(s).  $\pi(B)$  shall be referred to as "ancestor belief" and  $\delta(B)$  shall be referred to as "descendant belief."  $\pi(B)$  and  $\delta(B)$  are defined as follows:  $\pi(B) = P(B|D_B^+)$  and  $\delta(B) = P(D_B^-|B)$ , where  $D_B^-$  and  $D_B^+$  denote the data contained in a tree rooted at  $B$  and the data contained in the rest of a tree, respectively.

The belief of node  $B$  is obtained by combining these two supports via the product  $Bel(B) = \alpha * \pi(B) * \delta(B)$ , where  $\alpha$  and  $Bel$  denote a normalizing constant and the belief, respectively. A normalizing constant is necessary, when  $Bel$  becomes a vector, to make the sum of the elements of the vector to be 1.

Several types of nodes which require special treatments are identified by Pearl [1986a]:

- (1) Anticipatory node (a leaf node that has not been instantiated yet): For such a node  $x$ ,  $\delta(x) = (1 \ 1 \ 1 \ \dots 1)$ .
- (2) Data node (a node with instantiated value): Given that a node  $x$  has a number of outcomes, if the  $j$ th outcome of node  $x$  is observed to be true, we set  $\pi(x) = \delta(x) = (0 \ 0 \ 0 \ 1 \ 0 \ \dots 0)$  with 1 at the  $j$ th position. Suppose a node  $x$  represents the killer and there are four suspects. If we

have a witness evidence to convince that the 4th suspect is the killer, then  $\pi(x) = \delta(x) = (0 \ 0 \ 0 \ 1)$ .

(3) Dummy node (a node representing judgmental evidence bearing on another node): for a dummy node B, we do not specify  $\pi(B)$  or  $\delta(B)$  but, instead, define  $\delta_B(A)$  as  $K * P(B|A)$ , where K is constant. We define  $\delta_E(B)$  as the message sent from a node E (a child node of node B) to a node B, whereas  $\pi_B(A)$  denotes the message sent from a node A (the parent node of node B) to a node B.

(4) Root node: At the root node x, we set  $\pi(x) =$  prior probability of the root node.

Given  $\pi(x)$  and  $\delta(x)$  stored with each node x, the major problem is to determine how the influence of new information will spread through a tree, namely, how the parameters  $\pi(x)$  and  $\delta(x)$  of a node x can be determined from those of its neighbor nodes. The propagation scheme [Pearl, 1986a] consists of the following four steps:

Step 1: When a node B is activated to update its parameters, it simultaneously checks the  $\pi_B(A)$  message communicated by its parent node A and the messages  $\delta_1(B), \delta_2(B) \dots$  communicated by each of its child nodes.

Step 2:  $\delta(B)$  is computed as the product of  $\delta_1(B), \delta_2(B), \dots$ :  

$$\delta(B) = \delta_1(B) * \delta_2(B) * \dots = \prod_k \delta_k(B).$$

Step 3:  $\pi(B)$  is computed using  $\pi(B) = \beta * P(B|A) * \pi_B(A)$ , where  $\beta$  is a normalizing constant.

Step 4: Using the messages received together with the updated values of  $\delta(B)$  and  $\pi(B)$ , each node B computes the new messages sent to its child node(s) and its parent node.

Step 5 (Bottom-up propagation): The new message  $\delta_B(A)$  that a node B sends to its parent node A is computed by  $\delta_B(A) = \delta(B) * P(B|A)$ .

Step 6 (Top-down propagation): The new message  $\pi_E(B)$  that a node B sends to its kth child node E is computed by  $\pi_E(B) = \alpha * \pi(B) * \prod_{m \neq k} \delta_m(B)$ , or alternatively,

$$\pi_E(B) = \alpha' \frac{\text{Bel}(B)}{\delta_E(B)}$$

where  $\alpha$  and  $\alpha'$  are normalizing constants, and '!' denotes "not."

Figure 5 shows six successive stages of belief propagation through a binary tree [Pearl, 1986a].

#### Pearl's Algorithm for the Network Structure

Pearl's algorithm for the network structure is presented based on Figure 4 [Pearl, 1986a]:

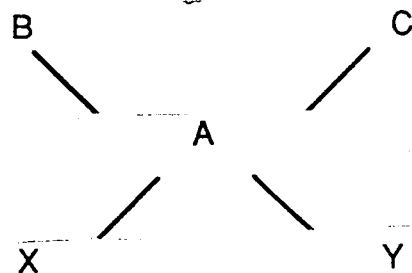


Figure 4. Example of the network structure

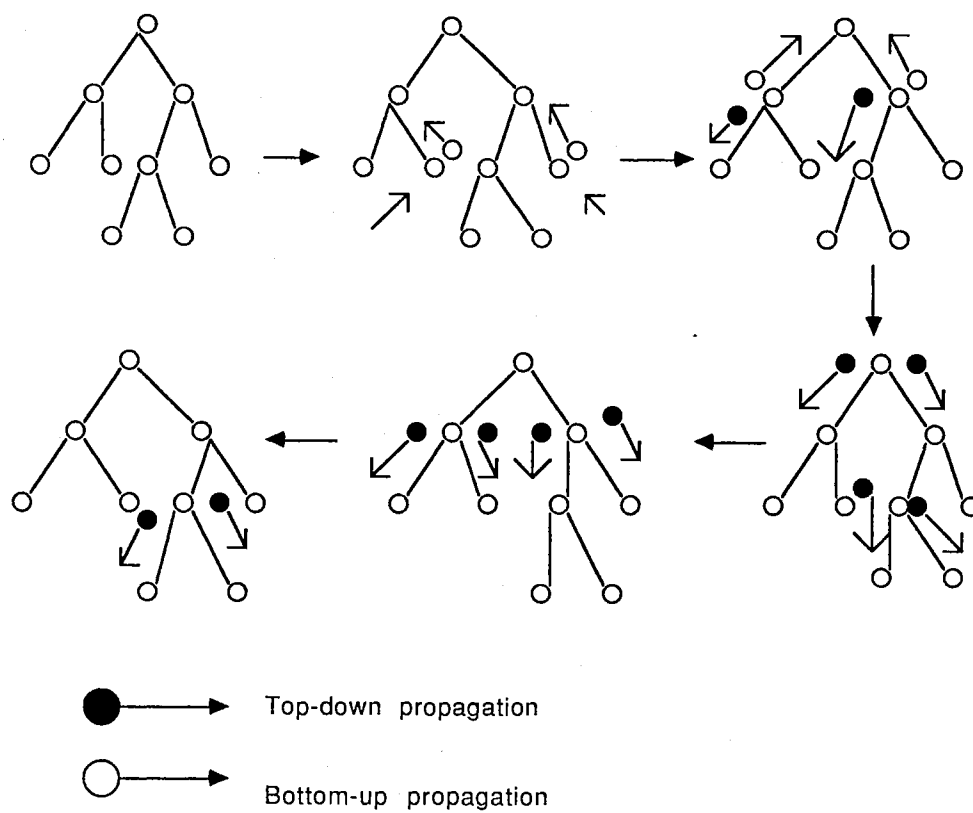


Figure 5. Illustration of Belief Propagation in Bayesian Networks



At each arc two parameters  $\pi$  and  $\delta$  are stored, where  $\pi$  denotes the "ancestor belief" contributed by its parent nodes and  $\delta$  denotes the "descendant belief" contributed by its child nodes. Each node contains the conditional probability to represent the strength of causal influence between the linked nodes. In Figure 4, the arc linking a node B to a node A stores  $\pi_A(B)$  and  $\delta_A(B)$ , where they represent the message sent from node B to node A and the message sent from node A to node B, respectively. The node A stores the conditional probability  $P(A|B \text{ and } C)$ .

As can be seen in Figure 4, the link  $A \longrightarrow Y$  partitions the graph into two parts: an upper subgraph,  $G_{AY}^+$ , and a lower subgraph  $G_{AY}^-$ .  $G_{AY}^-$  is a subtree rooted at node Y, while  $G_{AY}^+$  is the rest of the network.  $D_{AY}^+$  is defined as the data contained in  $G_{AY}^+$ , while  $D_{AY}^-$  is defined as the data contained in  $G_{AY}^-$ . The overall strength of belief in node A is calculated as follows:  $Bel(A) = \alpha * P(D_{AX}^- | A) * P(D_{AY}^- | A) * [\Sigma(P(A|B, C) * P(B|D_{BA}^+) * P(C|D_{CA}^+))] = \alpha * \delta_X(A) * \delta_Y(A) * \Sigma(P(A|B, C) * \pi_A(B) * \pi_A(C))$ .

The belief of a node B can be calculated as follows:  $BEL(B) = \alpha * \pi_A(B) * \delta_A(B)$ , where  $\alpha$  = a normalizing constant. Given the parameters  $\pi$  and  $\delta$  stored with each link, the influence of new information is spread through the network via the messages to its parent nodes and child nodes. The schemes of updating the messages are shown below:

$$\delta_A(B) = \alpha * \Sigma [\pi_A(C) * \Sigma(\delta_X(A) * \delta_Y(A) * P(A|B \text{ and } C))]; \text{ and}$$

$$\pi_X(A) = \alpha * \delta_Y(A) * [\Sigma P(A|B \text{ and } C) * \pi_A(B) * \pi_A(C)].$$

### Wise's Framework

Before Wise's framework is discussed, a number of studies attempting to determine the relationship between the non-Bayesian approaches and the probability theory, or among the non-Bayesian approaches are presented. Two related studies are presented below. First, Grosf [1986a] compared Heckerman's interpretation of CF approach and DST with the Bayesian approach, and Heckerman's interpretation of CF approach with DST.

Let  $O(H) = P(H) / [1 - P(H)] = P(H) / P(\sim H)$ ,  $O(H|E) = P(H|E) / [1 - P(H|E)] = P(H|E) / P(\sim H|E)$ ,  $L(H,E) = O(H|E) / O(H)$ , where  $H$  and  $E$  denote the hypothesis and an evidence, respectively. As discussed earlier in Chapter 2,  $L(H,E)$  is called a "likelihood ratio" and represents a belief update, thus is a basis of Heckerman's interpretation of CF approach. Grosf showed that the mapping between Heckerman's interpretation of CF approach and the "likelihood ratio" is:  $C = [L-1]/[L+1]$  or  $L = [1+C]/[1-C]$ , where  $C$  and  $L$  denote CF and  $L(H,E)$ , respectively.

Because Heckerman's interpretation of CF approach is developed based on the "likelihood" concept, this mapping is not surprising. Furthermore, the resulting mapping becomes obvious from the fact that  $C$  is defined as  $(L(H,E)-1) / (L(H,E)+1) = \{P(H|E) - P(H)\} / \{P(H) * [1 - P(H|E)] + P(H|E) * [1 - P(H)]\}$ .  $C$  is one of the valid interpretations of

Heckerman's CF approach which can have an infinite number of valid interpretations.

The mapping between point-valued DST (special case of DST) and a "likelihood ratio"  $L$  is:  $B = L/(L+1)$  or  $L = B/(1-B)$ , where  $B$  denotes the belief function of DST. This may imply that DST is richer and more powerful than the probability theory, due to the relationship between a point-valued DST and  $L$ . However, the conclusion is not definite.

The mapping between Heckerman's interpretation of CF approach and a point-valued DST is:  $B = [1+C]/2$  or  $C = 2B-1$ . This mapping may indicate that DST is richer and more powerful than Heckerman's interpretation of CF approach. However, we cannot derive definite conclusion.

Second, there have been a claim that DST is a generalization of the probability theory, because DST permits the representation of the interval-valued probability, whereas the probability theory allows a point-valued probability. Black [1987] and Kyburg [1987] showed that this claim is not true.

Kyburg [1987] showed that the belief function models allow a subset of closed convex probability distribution, i.e., not all closed convex sets of the probability distributions are represented in DST, contradicting the studies claiming that DST is a generalization of the probability theory. The Black's work which is an extension of Kyburg's work, showed many convex sets of the probability

distributions generates the same belief function, which consolidates the claim that DST is not a generalization of the probability theory. Black also compared Bayes' rule to Dempster's combining rule, and concluded that Bayes's rule performs better than Dempster's combining rule.

These studies trying to determine the relationships among non-Bayesian approaches, and the relationships between non-Bayesian approaches and the probability theory, do not produce the results which are consistent enough to derive the definite conclusions. Thus, Several researchers including Wise [1986] and Grosf [1986b] attempted to develop the frameworks which can be used in comparing among UIS' or UIS' to the probability theory.

Wise's framework is presented below, primarily because the proposed simulation model is developed based on Wise's framework. The proposed simulation model is adopted in determining the type of operations performed on the linguistic probability.

Buchanan and Shortliffe [1984b] who developed the CF approach, tried to justify the development of CF approach by comparing the results of CF approach to those of the probability theory. Of course, the conclusion was that CF approach performed as well as the probability theory. Consider the facts that the CF approach is the first non-Bayesian approach, and that Buchanan and Shortliffe compared the performance of CF approach to those of the probability

theory to justify their work. From these facts, we can argue that the probability theory could be a normative approach.

This argument also applied to FST. Zadeh [1986b] defended FST by showing five examples which cannot be dealt with by the ordinary (first-order) probability theory. DST is only non-Bayesian approach such that little attempts have been made to compare it to the probability theory. Thus, We can make the following conclusion: even though some non-Bayesian approaches have the advantages over the probability theory in terms of the power of expressiveness or the amount of information needed, the probability theory is the only normative approach.

This conclusion could lead to an argument that the probability theory can be used as the standard in comparing UIS' with each other. Wise's framework was developed primarily based on this argument.

As mentioned earlier, the ordinary (first-order) probability theory has the limitation in its applicability, in that it is not powerful enough to handle the "generalized" input, e.g., "it will probably rain tomorrow." On the other hand, although the second-order probability theory offers all the advantages of the Bayesian approach and is shown to be powerful to handle any kind of input by the researchers including Cheeseman [1986], it has the following major drawbacks.

First, the second-order probability distribution may not be obtainable in the real world applications given the amount of available information. In fact, simply forming a first-order probability distribution to model an expert's belief state is only marginally feasible, let alone a second-order distribution over all possible first-order probability distribution. Second, for the sake of argument, suppose that the second-order probability distribution can be obtainable, although it is rarely true in the real world. However, its exponential time complexity inhibits its use.

Thus, Wise proposed the (first-order) probability theory supplemented by ME/MXE as a substitute for the second-order probability theory, based on the argument that the probability theory supplemented by ME/MXE produces the results that are good approximations to those produced by the second-order probability theory. That is, Wise showed that ME/MXE can be used to estimate the prior probabilities and to update prior probabilities with the generalized data, as a good approximation to the second-order probability theory.

The basic goal of Wise's work is to identify the conditions producing significant differences in the output of UIS' using the experiments. The emphasis is on comparing the outputs, not the simplicity, explicability, or ease of construction of the UIS itself. The general outline of his framework is shown in the Figure 6. The left-hand column is

where ME/MXE inference is performed, while the inference of UIS being explored is done in the right-hand column.

In Figure 6,  $P_0$  denotes the prior probability,  $P_1$  denotes the posterior probability.  $R^*$  denotes the converted rules,  $D^*$  denotes the converted data.  $P_1^*$  denotes the output obtained from UIS, and  $P_1'$  denotes the transformed probability from the output obtained from UIS'. The prior probabilities are estimated using ME, provided that a collection of rules. The posterior probabilities are estimated using MXE provided that the prior probabilities and a collection of specific facts.

In Figure 6, the "conversion" denotes the transformation of belief measure of the non-Bayesian approaches to the probability, or the transformation of the probability to belief measure of non-Bayesian approaches. In the case of UIS, a rule set is converted into an appropriate form for the specific UIS and a fact set is also converted into the suitable form.

Wise [1986] introduced the conversion scheme between UIS' and the probability theory. His conversion scheme between FST and the probability theory is: probability = fuzzy membership function. Suppose that  $X = \{0.1/1, 0.3/3, 0.5/5, 0.7/7, 0.9/9\}$ . In his conversion scheme,  $P(X=1) = 0.1$ ,  $P(X=3) = 0.3$ ,  $P(X=5) = 0.5$ ,  $P(X=7) = 0.7$ , and  $P(X=9) = 0.9$ . The proposed conversion scheme between CF and the probability theory is: (1)  $CF = 1 \Rightarrow P(X) = 1, P(X') = 0$ ; (2)

CF= 0 =>  $P(X) = P(X')$ ; and (3) CF= -1 =>  $P(X) = 0, P(X') = 1$ .  
 Values between these three points are computed by a  
 piecewise linear interpolation.

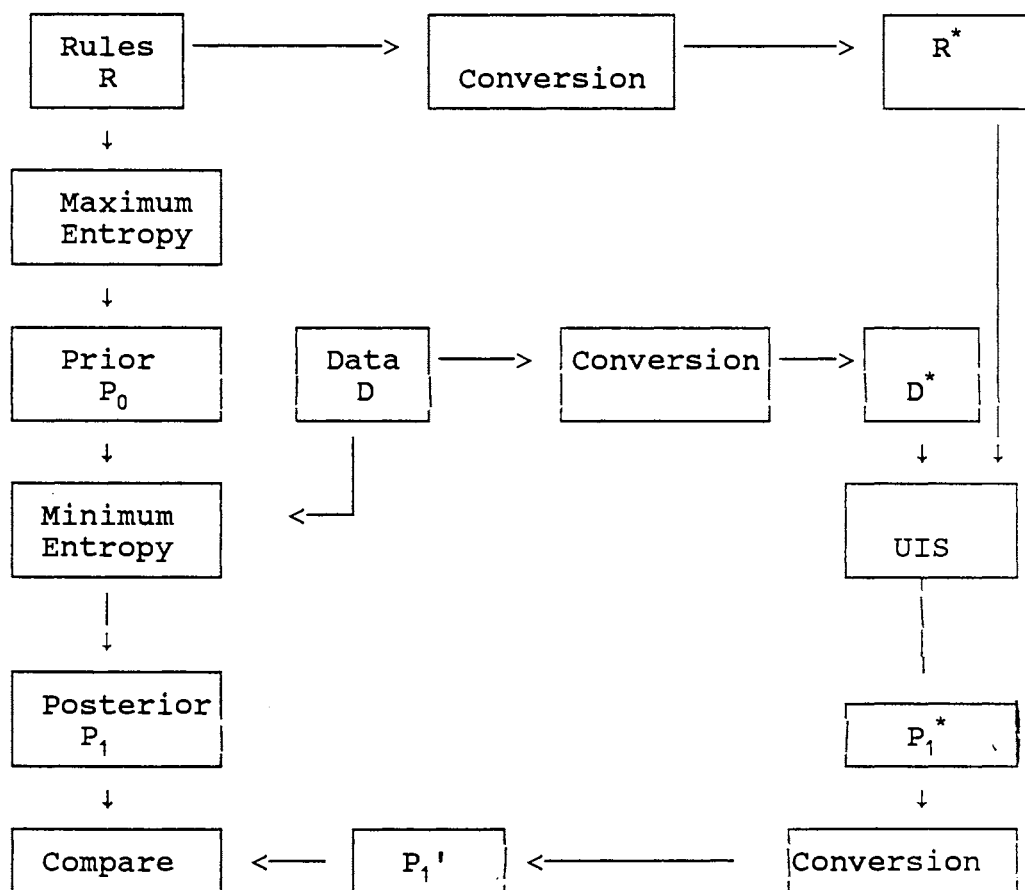


Figure 6. Basic Experiment Design for Comparisons



After the belief measure of non-Bayesian approach is transformed into the probability, the measures of error such as mean absolute error, mean squared error, normalized mean absolute error, and normalized mean squared error, are computed for each non-Bayesian approach. Obviously the non-Bayesian approach with the minimum error is regarded as the best non-Bayesian approach.

A question can arise as to the use of normalized mean absolute error and normalized mean squared error. The reasons for their use are presented below [Wise and Henrion, 1986]. A difficulty can arise in comparing performance on different cases in that they are likely to allow different ranges of error.

For example, if we randomly guess at the probability of 0.5, it is impossible to be off by more than 0.5, but if we guess the probability of an almost certain event (probability 0 or 1), then it is possible to be off by almost 1. Thus, an error near 0.5 is almost the worst possible in the former case, but is about average for the second.

The worst possible error of the probability is  $\max(P, 1-P)$ , while the error of random guess probability can be defined as the expected error, if the estimates of the probabilities are uniformly distributed over the estimated domain. For FST, the domain is the close interval  $[0,1]$ , while for CF approach, the domain is the closed interval  $[-$

-1,1].

For FST, the expected mean absolute error,  $\mu(|\Phi|)$ , and the expected mean squared error,  $\mu(\Phi^2)$  are:  $\mu(|\Phi|) = 1/2 - P(1-P)$ ; and  $\mu(\Phi^2) = 1/3 - P(1-P)$ . Similarly,  $\mu(|\Phi|)$ 's and  $\mu(\Phi^2)$ 's for the CF approach and DST can also be derived.

The normalized absolute error is defined as follows:  $\xi = 1$ , iff  $\Phi = 0$ ;  $\xi = 0$ , iff  $\Phi = \mu(|\Phi|)$ ; and  $\xi = -1$ , iff  $\Phi = \max(P, 1-P)$ . Similarly, the normalized squared error,  $\eta$  is defined as:  $\eta = 1$ , iff  $\Phi = 0$ ;  $\eta = 0$ , iff  $\Phi = \mu(\Phi^2)$ ; and  $\eta = -1$ , iff  $\Phi = \max(P^2, (1-P)^2)$ . Thus, the normalized measures rescale the errors to give 1 for zero error, 0 when it is as good as random guessing, and -1 for the worst possible error, with a linear interpolation in between.

#### Operations on Fuzzy Set and Fuzzy Number

The operations performed on fuzzy set and a fuzzy number are presented here mainly because one of the problems which the proposed approaches deal with is the determination of the type of operations performed on the linguistic probability. The basic operations performed on fuzzy sets are as follows: equal, contained, union, intersection, complement, product, bounded sum, bounded difference, left-square, convex combination, and Cartesian product.

A fuzzy number is a fuzzy subset of real numbers. Its example is "very close to 5" that could be defined as  $\{0.6/1, 0.7/2, 0.8/3, 0.9/4, 1/5, 0.9/6\}$ . The operations

performed on the fuzzy numbers are: inverse, scalar, multiplication, exponential, absolute value, extended addition, extended multiplication, extended subtraction, extended division, extended max, extended min, and extended power function [Dubois and Prade, 1980a].

Dubois and Prade [1980a] introduced a general algorithm for the computation of operations on fuzzy numbers. Any continuous fuzzy set can be decomposed into the union of convex fuzzy sets whose membership functions are either strictly increasing or decreasing or constant. In Figure 7, the increasing set is  $\{T_1, T_2\}$ , the decreasing set is  $\{T_5\}$ , and the constant set is  $\{T_3, T_4, T_6\}$ , where a fuzzy set shown in Figure 7 represents a fuzzy number.

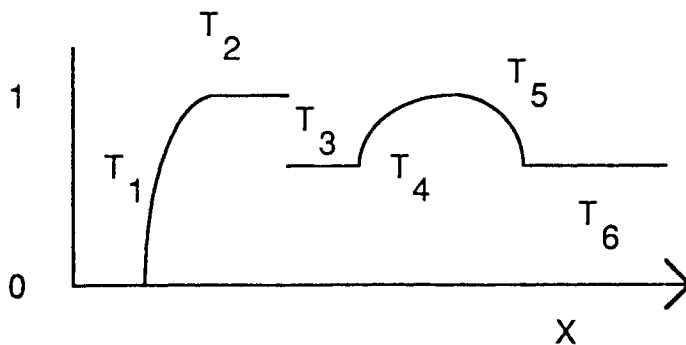


Figure 7. An Example of a Fuzzy Set

The general algorithm consists of the following four

steps:

Step 1 (Flattening): The  $n$  fuzzy sets are changed into fuzzy sets all having the same height. The height of a fuzzy set is the maximum membership value.

Step 2 (Decomposition of each fuzzy set into two sets, namely, the set of nondecreasing parts and the set of nonincreasing parts): The constant parts between two nondecreasing (nonincreasing) ones belong to the nondecreasing (nonincreasing) set. The constant parts, which are between parts of different kinds, belong to both.

Step 3 (Operation  $*^\circ$ ): The operation  $*^\circ$  is performed for parts belonging to the same kind of sets.

Step 4 (Union): The union of fuzzy sets obtained in step 3 is the final result.

The operation  $*^\circ$  represents any operation performed on fuzzy numbers.

Suppose that we want to add a fuzzy number  $A$  to a fuzzy number  $B$  shown in Figure 8. A fuzzy number  $A$  consists of three parts, i.e.,  $A_1$ ,  $A_2$ , and  $A_3$ , and a fuzzy number  $B$  consists of three parts, i.e.,  $B_1$ ,  $B_2$ , and  $B_3$ . In a fuzzy number  $A$ , the nondecreasing set is  $\{A_1, A_2\}$  and the nonincreasing set is  $\{A_2, A_3\}$ , whereas in a fuzzy number  $B$ , the nondecreasing set is  $\{B_1, B_2\}$  and the nonincreasing set is  $\{B_2, B_3\}$ .  $A_2$  and  $B_2$  can belong to both the nondecreasing set and nonincreasing set, because it lies between the parts of the different kind of sets.

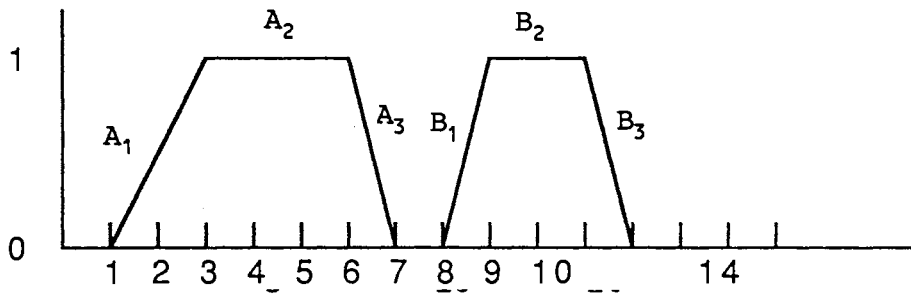


Figure 8. Fuzzy Number A and B

Let  $+^\circ$  denote the addition of fuzzy numbers. Also let  $C_{ij}$  represent the sum of two parts  $A_i$  and  $B_j$  denoted  $A_i +^\circ B_j$ . The following  $C_{ij}$ 's are calculated:  $C_{11} = A_1 +^\circ B_1$ ;  $C_{12} = A_1 +^\circ B_2$ ;  $C_{21} = A_2 +^\circ B_1$ ;  $C_{22} = A_2 +^\circ B_2$ ;  $C_{23} = A_2 +^\circ B_3$ ;  $C_{32} = A_3 +^\circ B_2$ ; and  $C_{33} = A_3 +^\circ B_3$ . The result of the addition of these fuzzy numbers is shown in Figure 9.

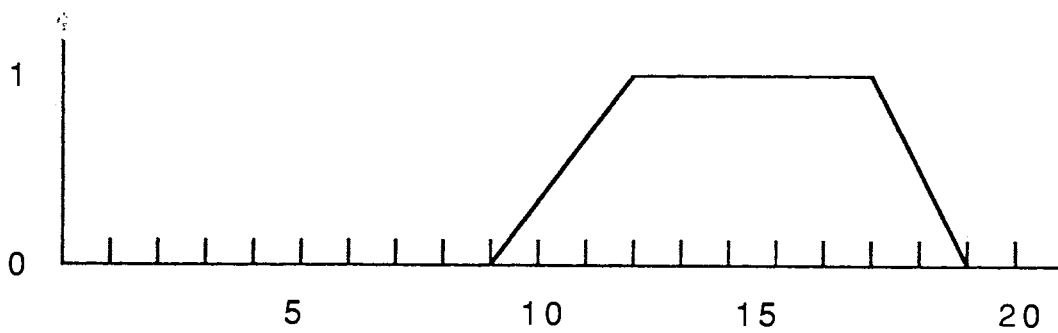


Figure 9. Sum of Fuzzy Numbers A and B

Suppose that two fuzzy numbers  $M_1$  and  $M_2$  are as follows with  $w_1 < w_2 < w_3$ :  $M_1 = w_1/p_1 + w_2/p_2 + w_3/p_3 + w_2/p_4 + w_1/p_5$  and  $M_2 = w_1/q_1 + w_2/q_2 + w_3/q_3 + w_2/q_4 + w_1/q_5$ . Then the result of any operation performed on two fuzzy numbers  $M_1$  and  $M_2$  is:  $M_1 *^\circ M_2 = w_1/(p_1 * q_1) + w_2/(p_2 * q_2) + w_3/(p_3 * q_3) + w_2/(p_4 * q_4) + w_1/(p_5 * q_5)$ , where  $*$  denotes any arithmetic operation on (ordinary) numbers.

The results of some operations on fuzzy numbers  $M_1$  and  $M_2$  are provided below:

(1) inverse of  $M_1$ :

$$M_1^{-1} = w_1/(1/p_1) + w_2/(1/p_2) + w_3/(1/p_3) + w_2/(1/p_4) + w_1/(1/p_5)$$

(2) scalar multiplication:

$$2M_1 = w_1/(p_1/2) + w_2/(p_2/2) + w_3/(p_3/2) + w_2/(p_4/2) + w_1/(p_5/2)$$

(3) extended addition of  $M_1$  and  $M_2$ :

$$M_1 +^\circ M_2 = w_1/(p_1+q_1) + w_2/(p_2+q_2) + w_3/(p_3+q_3) + w_2/(p_4+q_4) + w_1/(p_5+q_5)$$

(4) extended multiplication of  $M_1$  and  $M_2$ :

$$M_1 \times^\circ M_2 = w_1/(p_1 * q_1) + w_2/(p_2 * q_2) + w_3/(p_3 * q_3) + w_2/(p_4 * q_4) + w_1/(p_5 * q_5)$$

(5) extended subtraction of  $M_1$  from  $M_2$ :

$$M_2 \dots M_1 = w_1/(q_1-p_1) + w_2/(q_2-p_2) + w_3/(q_3-p_3) + w_2/(q_4-p_4) + w_1/(q_5-p_5)$$

(6) extended division of  $M_1$  by  $M_2$ :

$$M_1 \div^\circ M_2 = w_1/(p_1/q_1) + w_2/(p_2/q_2) + w_3/(p_3/q_3) + w_2/(p_4/q_4) + w_1/(p_5/q_5)$$

## Existing Approaches to Linguistic Approximation

Several approaches have been developed for a linguistic approximation. A linguistic approximation is the process of finding a label whose meaning is the same or the closest to the meaning of an unlabelled fuzzy set generated by the computational model. For example, a fuzzy set  $\{0.3/0.2, 0.4/0.4, 0.5/0.6, 0.8/0.9\}$  could be labelled as "more or less true" or "very true."

Note that one of the fundamental problems encountered by adopting the linguistic probability instead of ordinary probability is a linguistic approximation. Wenstop [1980], Eshragh and Mamdani [1979], Bonissone and Decker [1986], and Bonissone [1979] have developed the procedures for a linguistic approximation. These existing approaches are presented first followed by the discussion of a proposed approach.

### Wenstop's Approach

Wenstop's procedure was the first approach introduced for a linguistic approximation. Wenstop considered two parameters of a fuzzy set, that is, its imprecision and its location, where the imprecision of a fuzzy set is defined as the sum of membership values and the location is defined as the mean. Fifty-six linguistic values (or labels) were chosen which lie approximately evenly spread out in a location-imprecision coordinate system. Some of the

linguistic values are shown in Figure 10.

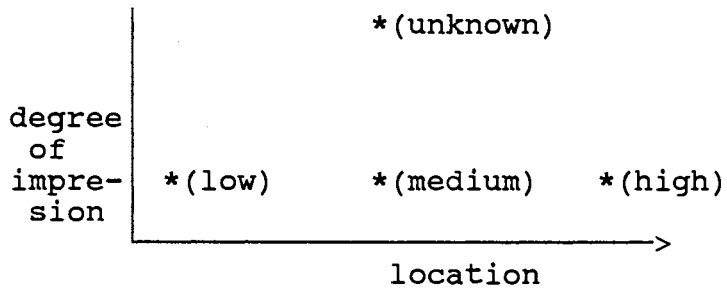


Figure 10. An illustration of Wenstop's approach

The selection of these two parameters was based not on a theoretical consideration but on an intuitive appeal. The label which has the shortest distance to  $X$  is chosen as the best label of a fuzzy set  $X$ . His approach requires the following condition to be satisfied: fuzzy sets should be regular, i.e., normal, unimodal and rather steep sided.

A fuzzy set  $X$  is called normal if its height is 1; otherwise it is subnormal. The height of a fuzzy set  $X$  is the supremum of  $f(X)$ , where  $f(X)$  denotes its membership function. For example, a fuzzy set defined as  $\{0.1/1, 0.5/5, 0.9/10\}$  is subnormal, because its height is 0.9, whereas a fuzzy set defined as  $\{0.1/1, 0.5/5, 1/1\}$  is normal. If a fuzzy set has a unique value of element at



which the membership function attains its maximum, a fuzzy set is called unimodal. If the fuzzy set input is not regular, then the appropriate procedures are applied to it before the procedure for a linguistic approximation is applied. He implemented this system in APL language.

Wenstop's approach has several drawbacks. First, the number of the linguistic values and their labels are predetermined, thus it lacks in the flexibility from the users' point of view. Second, the selection of two parameters does not have a strong basis, rather it is based on a intuitive appeal. Finally, the required condition that a fuzzy set is regular may be too strong to be satisfied in real world applications, thus restricts its applicability. Furthermore, a procedure developed to convert unregular fuzzy set into a regular fuzzy set is not good enough to offset its limitations. Although his approach has several weaknesses, we should give him a credit for a pioneering work in a linguistic approximation.

#### Eshragh and Mamdani's Approach

In their approach, the labels are made up of a combination of predetermined linguistic terms and appropriate logical connectives "AND" and/or "OR." Thus, the linguistic terms are formed by combining the hedges (or modifiers) and linguistic terms (or primary subsets) provided by the users. The hedges used in their system are:

"not," "very," "indeed," "more or less," "above," and "below."

The assignment of a linguistic label to a fuzzy set is achieved by labelling its segments comprising a fuzzy set. Then the labels obtained for the segments are appropriately concatenated to form a linguistic statement using connectives "AND" and "OR." Unlike Wenstop's approach, this approach does neither require normality nor restrict the number of primary subsets and their labels allowed.

They developed a heuristic search program called "LAM5" finding the best label for an unlabelled set. The procedure to find the best label consists of three steps: (1) a generator; (2) a search procedure; and (3) an evaluation method. At step 1, the primary subsets provided by the user are arranged so that the first and last primary subsets are S-type and the rest are II-Type. S-type includes  $S^{-}$ -set and  $S^{+}$ -set, where S-type and II-types are shown in Figure 11.

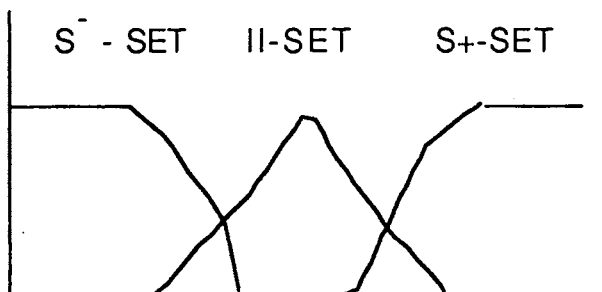


Figure 11. S-type and II-type

The hedges are combined with the primary subsets to construct a set of labels. The subsets associated with these labels are represented using the three parameters:  $\alpha$ ,  $\beta$ ,  $\tau$ , where  $\alpha$ ,  $\beta$ , and  $\tau$  are points in the universe of discourse at which the membership function attains 0, 0.5, and 1, respectively. This three parameter representation for each label is then stored permanently in a back up file.

The second step is a search procedure consisting of two phases. The first phase is an exhaustive search. If a given subset has the similar parametric representation to those of primary or negated primary subsets, then it will be tested for a perfect match. If a perfect match occurs, then the search is terminated and the second phase is avoided. Otherwise, the second phase is activated. In the second phase, the input is decomposed into a number of segments, and parameters of a given segment are found.

In the third step, i.e., an evaluation phase, first, we locate all the labels, among those labels stored in file, which have the same  $\alpha$ ,  $\beta$ , and  $\tau$  values as the unlabelled set. When the first of such labels is found, before any attempt is made in finding the next one, the distance between the relevant section of subset represented by that label and the segment under test, is computed using the least squares method. If the distance is zero, then that label is accepted as the label for that segment, and the evaluation phase is terminated. Otherwise, the distance is

noted. If this distance is smaller than the previous one, the new label is considered as a more suitable one and replaces any one previously found.

Their approach is more sophisticated than Wenstop's approach for the reasons listed below:

- (1) It allows the users to choose their own primary subsets and their labels. This is a significant improvement due to its flexibility with regard to the selection of primary subsets and their labels.
- (2) This approach employs the hedges which increase the level of granularity.
- (3) It employs the three-parameter representation for each primary subset, i.e,  $\alpha$ ,  $\beta$ , and  $\tau$ .

On the other hand, the drawbacks of their approach are shown below. First, the fact that it allows users to choose the primary subsets and their labels can be a weakness as well as an advantage. The selection of primary subsets affects the output generated by the system significantly. For instance, a term (or label) "likely" could be defined as  $\{0.1/0.1, 0.5/0.5, 0.9/0.9\}$  or  $\{0.2/0.1, 0.6/0.5, 1.0/0.9\}$ .

The output generated by a system can be affected by which fuzzy set represents a term "likely." Thus, the flexibility regarding the selection of the primary subsets, in fact, can pose the problem to some extent. This leads to an argument that the selection of the labels and primary

subsets should have a theoretical basis like a parametric representation introduced by Bonissone and Decker [1986].

Second, Eshragh and Mamdani used the three-parameter representation for each primary subset and its label. However, this representation may not be good enough to represent the entire distribution, because this representation does not contain the information about the segment between  $\alpha$  and  $\beta$ , or between  $\beta$  and  $\tau$ . The representation which employs the parameters including the mean, a variance, and the skewness could be better than  $[\alpha, \beta, \tau]$  representation. In fact, Bonissone [1979] adopted this approach by using four parameters including the mean and skewness.

#### Bonissone's Approach

Bonissone [1979] introduced an approach based on the feature selection and pattern recognition. He identified four parameters to represent the linguistic values: the power, entropy, first moment, and skewness. These four features were selected after several experiments looking for an efficient representation. Since an infinite number of features including the mean, a variance, and the skewness are available, we need to find a limited number of features to represent the linguistic values efficiently.

The power of a fuzzy set is defined as the sum of the membership values. For example, the power of a fuzzy set  $A=$

$\{0.1/0.2, 0.5/0.6, 0.9/1\}$  is 1.5 ( $=0.1+ 0.5+ 0.9$ ). The fuzzy entropy of a fuzzy set A is defined as:  
 $\text{Entropy}(A) = \sum_i S(\mu_A(u_i))$ , where  $S(X) = -(X \cdot \ln X) - (1-X) \cdot \ln(1-X)$ . The first moment indicates the center of the probability distribution and the skewness is a measure of asymmetry of the distribution with respect to its mean.

The moments are defined as follows [Mood et al., 1963]:  
 If X is a random variable, the rth moment of X is defined as  $E(X^r)$ , where E denotes the expectation, if the expectation exists. Similarly, if X is a random variable, the rth (central) moment of X about a is defined as  $E[(X-a)^r]$ . The first moment of X, i.e.,  $E(X)$  is the mean of X and the third moment about  $\mu$  denoted  $E[(X-\mu)^3]$  is the skewness of X.

Figure 12 shows the positive and negative skewness, where a curve shaped like  $f_1(X)$  is said to be skewed to the left (or a negative skewness), whereas one shaped like  $f_2(X)$  is said to be skewed to the right (or a positive skewness).

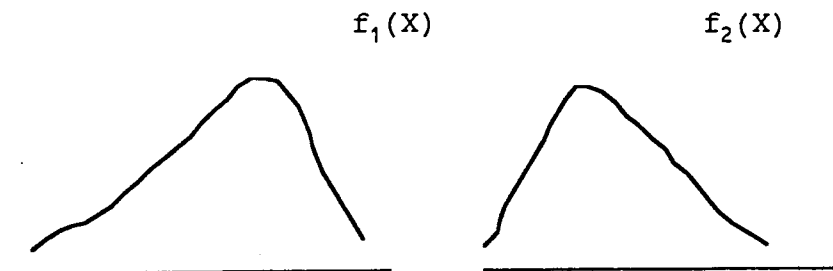


Figure 12. Positive and negative skewness

Bonissone's search procedure consists of two steps: (1) The first step consists in prescreening the term set; and (2) The second step determines a modified Bhattacharyya distance between the distribution of the unlabelled set and the distribution of each of these preselected labels. Then the label of a fuzzy set with a minimum modified Bhattacharyya distance is selected as the best label.

The goal of the first step, namely, the prescreening process is to reduce a search space. This is achieved by evaluating the four parameters (or features) of the unlabelled fuzzy set and by using a weighted Euclidean distance in the feature space followed by selecting the terms in a term set whose weighted Euclidean distance is within a desired tolerance level. Thus, the first step selects the terms such that the weighted Euclidean distance  $d(A, A^*) \leq E$ , where  $A$  and  $A^*$  denote the term set and unlabelled set, respectively and  $E$  denotes an acceptable tolerance level.

The weighted Euclidean distance  $d$  is defined as:  
 $d(A, A') = [\sum_i w_i (P_i - P'_i)^2]^{0.5}$ , where  $A$  and  $A'$  are fuzzy sets,  $w_i$ 's are the weights assigned to four parameters. The  $w_i$ 's can be obtained from the user indirectly by means of pairwise comparison tests. The Bhattacharyya distance is defined as:  $d(p_A(u), p_B(u)) = -\ln R$ , where  $R$  is called the Bhattacharyya coefficient.  $R$  is defined as:  $R(p_A, p_B) = \sum_i [p_A(u_i) * p_B(u_i)]^{1/2}$ , where  $p_A(u_i)$  and  $p_B(u_i)$  denote the

membership function of the fuzzy set A and B, respectively.

For example, if  $A = \{0.1/0.2, 0.5/0.6, 0.8/0.9\}$  and  $B = \{0.3/0.2, 0.6/0.6, 0.9/0.9\}$ , then  $d(A,B) = -\ln [(0.1*0.3 + 0.5*0.6 + 0.8*0.9)^{0.5}]$ . However, Bonissone argued that since this measure does not satisfy the triangle inequality, therefore it is not a metric. Instead, Bonissone defined a modified Bhattacharyya distance satisfying all the axioms of a metric:  $d(p_1, p_2) = [1 - R(p_1, p_2)]^{0.5}$ .

The procedure to determine  $w_i$ 's, namely, the weights assigned to four parameters is presented next. Yager [1977] showed that a method developed by Saaty can be applied to the determination of the  $w_i$ 's. Thus, Saaty's procedure is described below [Yager, 1977]. Assume we have P objects and we want to construct a scale rating of these objects as to their importance with respect to a certain criterion. We ask the decision-maker to compare the objects in a paired comparison.

When an object i is compared with another object j, the following values  $a_{ij}$  and  $a_{ji}$  are assigned: (1)  $a_{ij} = 1/a_{ji}$ ; and (2) if i is more important than j, we assign a number to  $a_{ij}$  from Table 4 [Yager, 1977]. Having obtained  $a_{ij}$ 's, a  $P \times P$  matrix B is constructed so that (1)  $b_{ii} = 1$ ; (2)  $b_{ij} = a_{ij}$ ,  $i \neq j$ ; and (3)  $b_{ji} = 1/b_{ij}$ . Saaty showed that the eigenvector corresponding to the maximum eigenvalue of a matrix B indicates the importance of P objects.

In the problem of finding values of a scalar parameter



$\Gamma$  for which there exist vectors  $X \neq 0$  satisfying  $AX = \Gamma X$ , where  $A$  is a given  $N \times N$  matrix, the values of  $\Gamma$  are called the "eigenvalues" and the vectors  $X \neq 0$  which satisfy  $AX = \Gamma X$  are called "eigenvectors" of the matrix  $A$  [Hadley, 1961]. Because a computer program to compute the eigenvector corresponding to the maximum eigenvalue of a given matrix is not straightforward, it is listed in Appendix A.

TABLE 4  
LISTING OF INDEXES REPRESENTING  
THE IMPORTANCE

Intensity of importance	Definition
1	Equal importance
3	Weak importance of one over the other
5	strong importance of one over the other
7	Demonstrated importance of one over the other
9	absolute importance of one over the other
2,4,6,8	Intermediate values: believe two adjacent judgement

Suppose three people  $X$ ,  $Y$ , and  $Z$  are being rated on a scale as to their importance to an organization [Yager, 1979]:

$Y$  is weakly more important than  $X$      $A_{12} = 1/3, A_{21} = 3$

$Z$  is somewhere between equal and     $A_{13} = 1/2, A_{31} = 2$

and weakly more important than X

Y is weakly more important than Z.  $A_{23} = 3$ ,  $A_{32} = 1/3$ .

Then the matrix B becomes:

$$B = \begin{matrix} & \begin{matrix} X & Y & Z \end{matrix} \\ \begin{matrix} X \\ Y \\ Z \end{matrix} & \begin{bmatrix} 1 & 1/3 & 1/2 \\ 3 & 1 & 3 \\ 2 & 1/3 & 1 \end{bmatrix} \end{matrix}$$

We then need to solve the eigenvalue problem  $BX = \Gamma_{\max} X$ , where  $\Gamma_{\max}$  is the maximum eigenvalue of the matrix B and compute the eigenvector corresponding to  $\delta_{\max}$ . In this example,  $W = (0.16 \ 0.59 \ 0.25)^T$ . This result indicates that the importance of three people X, Y, and Z to an organization are 0.16, 0.59, and 0.25, respectively.

Bonissone's scheme has several strengths and weaknesses, especially comparing to Eshragh and Mamdani's scheme. The advantages are discussed first, followed by the discussion of the disadvantages. First, it adopts four features, namely, power, entropy, first moment, and skewness to represent each subset. It seems that this representation is better than Eshragh and Mamdani's scheme.

Second, it adopts the prescreening process to reduce a search space. As the terms employed increase, the prescreening process will play an important role in reducing the search time. Third, it developed a modified Bhattacharyya distance which is a good metric. On the other hand, it has a drawback in that it does not employ the

hedges, whereas the hedges are adopted in Eshragh and Mamdani's approach.

### Parametric Representation of Linguistic Probability

Bonissone and Decker [1986] presented a parametric representation of the linguistic probability. This parametric representation is achieved by the 4-tuple  $(a, b, \alpha, \beta)$ , where the first two parameters indicate the interval in which the membership value is 1.0, whereas the third and fourth parameters indicate the left and right width of the distribution, respectively. Its membership distribution is shown in Figure 13.

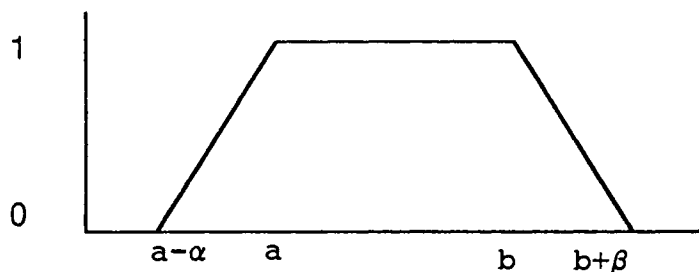


Figure 13. A parametric representation  $[a, b, \alpha, \beta]$

As discussed earlier, Eshragh and Mamdani introduced the parametric representation achieved by 3-tuple  $[\alpha, \beta, \tau]$ .

Both parametric representations introduced by Bonissone and Decker, and Eshragh and Mamdani have the advantage in that they are computationally efficient ways to characterize the linguistic probability. Now the question as to which parametric representation needs to be addressed. The parametric representation achieved by 4-tuple  $[a, b, \alpha, \beta]$  is better than the parametric representation achieved by 3-tuple  $[\alpha, \beta, \tau]$  for the following reasons.

First of all, 4-tuple representation  $[a, b, \alpha, \beta]$  does represent the entire membership function better than 3-tuple representation  $[\alpha, \beta, \tau]$ , simply because it has one more parameter. On the other hand, it can be argued that 3-tuple representation is better than 4-tuple representation in terms of the memory requirement. However, as long as the number of the linguistic values employed by the system are of reasonable size, the issue of memory requirement is of little concern.

Furthermore, the 4-tuple representation is the only representation which is based on the results of psychological experiments on the use of linguistic probabilities [Beyth-Marom, 1966]. Note that the membership function of a fuzzy set affects the output significantly.

#### Proposed Approaches

Because Pearl's algorithms use the probability as the measure of the belief, the determination of the conditional

probability distributions of the links and the prior probability distribution of the root node need to be determined in advance before Pearl's algorithms are applied. The precision inherent in the probabilities affects the precision of the belief of a node of interest to a great extent.

The main problem with Pearl's algorithms is how accurately users/experts can determine the prior probability distribution and conditional probability distributions. Several researchers including Szolovits and Pauker argued that when users/experts must provide these measures, an assumption of "fake precision" must usually be made. Szolovits and Pauker [1978] noted that "...while people seem quite prepared to give qualitative estimates likelihood, they are often notoriously unwilling to give precise numerical estimates to outcomes."

Therefore, they argued that any scheme that relies on the user providing consistent and precise numerical quantifications of the confidence level of his/her conditional or unconditional statements is bound to fail. For example, in response to the question "how much does the report on U.S. trade deficit in January 1989 affect the stock market in March 1989?", the expert may feel more comfortable with providing an answer of "very likely" rather than 80 percent.

For the sake of argument, suppose that accurate values

of the probabilities are obtainable from the users/experts in the actual applications, although it will be rarely true primarily due to an assumption of "fake precision." Pearl's algorithms can rectify the error inherent in the prior probability distribution by updating it, in a time proportional to the longest path in the network. Because the time it takes to update the prior probability distribution is proportional to constant, i.e., the longest path in the network, the problem due to an inaccurate estimation of the prior probability distribution is not significant enough to justify the use of linguistic probabilities instead of the probability.

However, the problem with the estimation of conditional probability distributions may be significant enough to justify the use of linguistic probabilities. For example, suppose that thirty rules need to be applied to compute the belief of one proposition of interest. Also suppose that the longest path in the network is 50. When there is an error in one conditional probability distribution associated with one link, its impact can be updated in a time proportional to 50. Note that there is the conditional probability distribution associated with each link.

Suppose that there are errors in all thirty conditional probability distributions. This is likely to occur especially if we adopt the definition of probability as a measure of belief. The definition of probability as a

measure of belief was favored by a number of researchers including Cheeseman who have strongly defended the probability theory. In this definition of probability, the probability is a (subjective) measure of an entity's belief in that proposition given the evidence.

The time required to impart the impacts to all nodes becomes proportional to 1,500 ( $=50*30$ ). We can imagine how much time is required in the propagation scheme, if one hundred rules need to be applied. Although the time required is proportional to constant, i.e., the longest path of the network, the updating time gets huge as the number of rules employed by the system increases. Furthermore, note that this argument is based on the unrealistic assumption that the accurate values of probabilities are obtainable.

The major advantages of using the linguistic probabilities is presented below. Phillips and Edwards [1966] observed the conservatism which is consistently present among the suppliers of subjective assessments, when dealing with subjective assessment of the probability. The behavior in which people tend to stick to the original (a priori) assessment regardless of new amount of evidence that should cause a revision of their belief, is called observed conservatism.

Zimmer [1985] performed an experiment to compare the linguistic probabilities with numerical probabilities to determine if the observed conservatism in the belief

revision was a phenomenon intrinsic in the perception of the events or due to the type of representation (i.e., numerical rather than verbal expressions). The results indicated that people are much closer to the optimal Bayesian revision when they are allowed to use linguistic probabilities.

The proposed algorithms adopt the linguistic probability instead of the (ordinary) probability, where "likely," "unlikely," "probable," and "very probable" are the examples of linguistic probability. The linguistic probability is defined as a fuzzy subset of  $[0,1]$ . For instance, linguistic probability "likely" could be defined as  $\{0.3/0.6, 0.5/0.7, 0.7/0.8, 0.9/0.9, 1/1\}$ .

For illustrative purposes, the following example is used throughout this section to contrast Pearl's algorithm for a tree structure to the proposed algorithm for a tree structure:

(example) Assume that in a certain trial there are three suspects, one of whom has definitely committed a murder, and that the murder weapon showing some fingerprints, was later found by the police. Let A stand for the identity of the last user of the weapon, namely, the killer. Let B stand for the identity of the last holder of the weapon, i.e., the person whose finger prints were left on the weapon. Let C represent the possible readings that may be obtained in a fingerprint-testing laboratory.



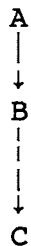


Figure 14. Example of the Network Diagram

A tree diagram of this example is shown in Figure 14. The arc from node A to node B contains the conditional probability distribution  $P(B|A)$ . The node C is a dummy node and is indicated by a dotted link. Suppose that  $P(B|A)$  is:

$$P(B_j|A_i) = \begin{cases} 0.8 & \text{if } A_i = B_j, \quad i, j = 1, 2, 3 \\ 0.1 & \text{if } A_i \neq B_j, \quad i, j = 1, 2, 3 \end{cases}$$

or stated another way

$$P(B|A) = \begin{matrix} & & B_1 & B_2 & B_3 \\ \begin{matrix} A_1 \\ A_2 \\ A_3 \end{matrix} & \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \end{matrix}$$

In the matrix  $P(B|A)$ , three rows represent  $A_1$ ,  $A_2$ , and  $A_3$ , respectively, where  $A_i$  denotes an event that the identify of the murderer is the  $i$ th suspect. Similarly, three columns represent  $B_1$ ,  $B_2$ , and  $B_3$ , respectively, where  $B_i$  denotes an event that the last holder of weapon is the  $i$ th suspect.

Also suppose that  $P(B|C)$  is  $(1 \ 1 \ 1)$ , because the laboratory report is not available at this moment.  $P(B|C) = (1 \ 1 \ 1)$  means that if the laboratory report indicates that the fingerprint of the  $i$ th suspect is found on the weapon, then the  $i$ th suspect is the last holder of the weapon with certainty. Let us assume that the prior probability distribution of node A is  $(0.8 \ 0.1 \ 0.1)$ . This implies that the probability the first suspect is a murderer is 0.8, while the probability the second or third suspect is a murderer is 0.1.

Pearl's algorithm and a proposed algorithm are presented below, where Pearl's algorithm is shown in the left-hand side and a proposed algorithm is shown in the right-hand side. Note that the linguistic probabilities are chosen arbitrarily for illustrative purposes. Also note that the vectors  $(A_1 \ A_2 \ A_3), (B_1 \ B_2 \ B_3), \dots, (M_1 \ M_2 \ M_3)$  denote the resulting vectors obtained from the relevant operations.

$$P(B|A) = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \quad P(B|A) = \begin{bmatrix} \text{likely} & \text{unlikely} & \text{unlikely} \\ \text{unlikely} & \text{likely} & \text{unlikely} \\ \text{unlikely} & \text{unlikely} & \text{likely} \end{bmatrix}$$

$$\pi(A) = (0.8 \ 0.1 \ 0.1) \quad \pi(A) = (\text{likely} \ \text{unlikely} \ \text{unlikely})$$

(B: an anticipatory node before obtaining any fingerprint information)

$$\delta(B) = (1 \ 1 \ 1) \quad \delta(B) = (\text{certain} \ \text{certain} \ \text{certain})$$

( $\delta_B(A)$ : message sent from a node B to a node A)

$$\delta_B(A) = \delta(B) * P(B|A) = \quad \delta_B(A) = (\text{certain} \ \text{certain} \ \text{certain})$$

$$(1 \ 1 \ 1) \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \quad \begin{bmatrix} \text{likely} & \text{unlikely} & \text{unlikely} \\ \text{unlikely} & \text{likely} & \text{unlikely} \\ \text{unlikely} & \text{unlikely} & \text{likely} \end{bmatrix}$$

$$= (1 \ 1 \ 1) \quad = (A_1 \ A_2 \ A_3)$$

(B: only child node of node A)

$$\delta(A) = \delta_B(A) = (1 \ 1 \ 1) \quad \delta(A) = \delta_B(A) = (A_1 \ A_2 \ A_3)$$

(Computation of the belief of node A)

$$BEL(A) = \alpha * \delta(A) * \pi(A) = \quad BEL(A) = \alpha * \delta(A) * \pi(A) = (A_1 \ A_2 \ A_3) *$$

$$(1 \ 1 \ 1) * (0.8 \ 0.1 \ 0.1) \quad (\text{likely} \ \text{unlikely} \ \text{unlikely})$$

$$= (0.8 \ 0.1 \ 0.1) \quad = (B_1 \ B_2 \ B_3)$$

( $\pi_B(A)$ : message sent from node A to node B): top-down propagation

$$\pi_B(A) = \alpha' * BEL(A) / \delta_B(A) \quad \pi_B(A) = \alpha' * BEL(A) / \delta_B(A)$$

$$= \alpha' * (0.8 \ 0.1 \ 0.1) / \quad = \alpha' * BEL(A) / (A_1 \ A_2 \ A_3)$$

$$(1 \ 1 \ 1) = (0.8 \ 0.1 \ 0.1) \quad = (C_1 \ C_2 \ C_3)$$

( $\pi(B)$ : support attributed to B by its parent node)

$$\pi(B) = \beta * \pi_B(A) * P(B|A) \quad \pi(B) = \beta * \pi_B(A) * P(B|A) = \beta * (C_1 \ C_2 \ C_3) *$$

$$= \beta * (0.8 \ 0.1 \ 0.1)$$

$$\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} \text{likely} & \text{unlikely} & \text{unlikely} \\ \text{unlikely} & \text{likely} & \text{unlikely} \\ \text{unlikely} & \text{unlikely} & \text{likely} \end{bmatrix}$$

$$= (0.66 \ 0.17 \ 0.17) \quad = (D_1 \ D_2 \ D_3)$$

(Computation of the belief of node B)

$$BEL(B) = \alpha * \delta(B) * \pi(B) \quad BEL(B) = \alpha * (\text{certain} \ \text{certain} \ \text{certain})$$

$$= \alpha * (1 \ 1 \ 1) * \quad (D_1 \ D_2 \ D_3) = (E_1 \ E_2 \ E_3)$$

$$(0.66 \ 0.17 \ 0.17) = (0.66 \ 0.17 \ 0.17)$$

(Now a laboratory report arrives and summarizes the test

results):

( $\delta_c(B)$ : message from the dummy node C to node B)

$$\delta_c(B) = \delta(B) \quad \delta_c(B) = \delta(B) = (\text{very likely likely likely}) \\ = (0.8 \ 0.6 \ 0.5)$$

(Updated belief of node B)

$$\text{BEL}(B) = \alpha * \delta(B) * \pi(B) \quad \text{BEL}(B) = \alpha * \delta(B) * \pi(B) = \alpha * \\ = \alpha * (0.8 \ 0.6 \ 0.5) * (\text{very likely likely likely}) * (D_1 \ D_2 \ D_3) \\ (0.66 \ 0.17 \ 0.17) \quad = (F_1 \ F_2 \ F_3) \\ = (0.738 \ 0.142 \ 0.119)$$

( $\delta_B(A)$ : message from node B to node A): bottom-up propagation

$$\delta_B(A) = \delta(B) * P(B|A) \quad \delta_B(A) = \delta(B) * P(B|A) = (\text{very likely} \\ = (0.8 \ 0.6 \ 0.5) \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} \text{likely likely} \\ \text{likely unlikely unlikely} \\ \text{unlikely likely unlikely} \\ \text{unlikely unlikely likely} \end{bmatrix} \\ = (0.75 \ 0.61 \ 0.54) \quad = (G_1 \ G_2 \ G_3)$$

(B: only child node of node A)

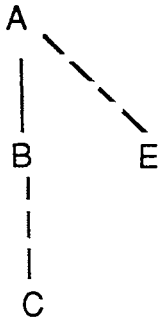
$$\delta(A) = \delta_B(A) \quad \delta(A) = \delta_B(A) = (G_1 \ G_2 \ G_3)$$

(Updated belief of node A)

$$\text{BEL}(A) = \alpha * \delta(A) * \pi(A) \quad \text{BEL}(A) = \alpha * (G_1 \ G_2 \ G_3) (\text{likely unlikely} \\ \text{unlikely}) \\ = \alpha * (0.75 \ 0.61 \ 0.54) * = (H_1 \ H_2 \ H_3) \\ (0.8 \ 0.1 \ 0.1) \\ = (0.839 \ 0.085 \ 0.076)$$

(suspect A<sub>1</sub> produces a very strong alibi in his/her favor, suggesting that there are only 1:10 odds that he could have

committed the crime. We link a dummy node E directly to A.  
E is an event producing a alibi)



$(\delta_E(A)$ : message from node E to node A)

$$\delta_E(A) = (0.1 \quad 1 \quad 1) \quad \delta_E(A) = (\text{very unlikely} \quad \text{certain} \quad \text{certain})$$

$(\delta(A)$ : support attributed to A by its child nodes)

$$\delta(A) = \delta_E(A) * \delta_B(A) \quad \delta(A) = (\text{very unlikely} \quad \text{certain} \quad \text{certain}) * (G_1 \quad G_2 \quad G_3) =$$

$$= (0.1 \quad 1 \quad 1) * (0.75 \quad 0.61 \quad 0.54) \quad (I_1 \quad I_2 \quad I_3)$$

$$= (0.075 \quad 0.61 \quad 0.54)$$

(Updated belief of node A)

$$\text{BEL}(A) = \alpha * \delta(A) * \pi(A) \quad \text{BEL}(A) = \alpha * (I_1 \quad I_2 \quad I_3)$$

$$= \alpha * (0.075 \quad 0.61 \quad 0.54) * (\text{likely} \quad \text{unlikely} \quad \text{unlikely})$$

$$= (J_1 \quad J_2 \quad J_3)$$

$$(0.8 \quad 0.1 \quad 0.1)$$

$$= (0.343 \quad 0.349 \quad 0.309)$$

( $\pi_B(A)$ ): message from node A to node B): top-down propagation

$$\pi_B(A) = \alpha * \delta_E(A) * \pi(A) \quad \pi_B(A) = \alpha * (\text{very unlikely certain certain}) *$$

$$= \alpha * (0.1 \quad 1 \quad 1) * \quad (\text{likely unlikely unlikely})$$

$$(0.8 \quad 0.1 \quad 0.1) \quad = \alpha * (K_1 \quad K_2 \quad K_3)$$

$$= \alpha * (0.08 \quad 0.1 \quad 0.1)$$

( $\pi(B)$ ): support attributed to B by its parent node)

$$\pi(B) = \beta * \pi_B(A) * P(B|A) \quad \pi(B) = \beta * (K_1 \quad K_2 \quad K_3)$$

$$= \beta * (0.08 \quad 0.1 \quad 0.1)$$

$$\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

$$\begin{bmatrix} \text{likely} & \text{unlikely} & \text{unlikely} \\ \text{unlikely} & \text{likely} & \text{unlikely} \\ \text{unlikely} & \text{unlikely} & \text{likely} \end{bmatrix}$$

$$= (0.3 \quad 0.35 \quad 0.35) \quad = (L_1 \quad L_2 \quad L_3)$$

(Updated belief of node B)

$$BEL(B) = \alpha * \delta(B) * \pi(B) = BEL(B) = \alpha * (\text{very likely likely})$$

$$\alpha * (0.8 \quad 0.6 \quad 0.5) * \quad \text{likely})$$

$$(0.3 \quad 0.35 \quad 0.35) \quad = (L_1 \quad L_2 \quad L_3) = (M_1 \quad M_2 \quad M_3)$$

$$= (0.384 \quad 0.336 \quad 0.28)$$

After updating the belief of node B, the propagation scheme is terminated.

There are two fundamental problems with the proposed approach. The first problem is to determine the types of operations performed on the linguistic probabilities. The second problem is to find a label of an unlabelled set generated by the computational model.

Regarding the operations performed on the linguistic probabilities, we need to resolve several problems presented below. First, we need to determine whether or not the sum

of the linguistic probabilities equals 1. In the example given above, for instance, the question is whether the sum of elements in the first row of matrix  $P(B|A)$ , i.e., likely+ unlikely+ unlikely equals 1, whereas the sum of the probabilities of all possible events should be 1 in the probability theory.

Second, we need to determine how to handle the multiplication of matrices whose elements are linguistic probabilities. This is reflected in the computation of the row vector  $(A_1 A_2 A_3)$  in the example given above. Third, we need to determine how to handle the product of two linguistic probabilities. This is reflected in the computation of the row vector  $(B_1 B_2 B_3)$  in the example given above. Finally, we need to determine how to handle the division of one linguistic probability by another linguistic probability. This is shown in the computation of the vector  $(C_1 C_2 C_3)$  in the example given above.

The second fundamental question is to find a label of an unlabelled set. For example, we could label  $B_1$  as "likely." The process of finding a label of an unlabelled set is called a linguistic approximation. In the proposed algorithms employing the linguistic probabilities, we do not attempt to modify Pearl's algorithms, but attempt to resolve the problems associated with using the linguistic probabilities.

These two fundamental problems also occur in a propose

algorithm for the network structure. The next two sections deal with these two problems, where the first problem is discussed first followed by the discussion of the second problem, namely, a linguistic approximation.

### Interpretations of Linguistic Probabilities

The issue of determining the types of operations performed on the linguistic probabilities is the kernel of the proposed algorithms, because research is scarce in the literature regarding this problem. On the other hand, several approaches [Bonissone, 1979; Bonissone and Decker, 1986; Eshragh and Mamdani, 1979; Wenstop, 1980] have been developed in regard to a linguistic approximation.

When these algorithms were initiated, the development of a collection of theorems was expected to resolve this issue. However, after several attempts, it was concluded that such a theorem base was not forthcoming. It was decided to employ a heuristic approach. Several researchers including Bonissone and Decker [1986] indicated that the linguistic probability can be regarded as a fuzzy number. A fuzzy number is a fuzzy subset of real numbers, e.g., "approximately 0.9." Dubois and Prade [1980a] defined a number of operations performed on fuzzy numbers. It seems that an alternative approach is to treat the linguistic probability as the ordinary fuzzy set. A number of operations on fuzzy sets including union and intersection



have been defined by a number of researchers including Zadeh [1965].

Depending on which interpretation of the linguistic probability we follow, the result becomes totally different due to the fact that the operations defined on fuzzy numbers are completely different from those defined on fuzzy sets. In the absence of a set of theorems to determine the best interpretation, it seems feasible to use a simulation approach. The complete description of a simulation model will be done later in this section.

#### Fuzzy Set Interpretation

Under this interpretation, we treat the linguistic probability as the ordinary fuzzy set. The basic operations performed on fuzzy sets are as follows: equal, contained, union, intersection, complement, product, bounded sum, bounded difference, left-square, convex combination, and Cartesian product. In this interpretation, the sum of the fuzzy sets obviously needs not to be 1. For example, the sum of two fuzzy sets "young" and "middle-aged" needs not to be 1.

As far as the multiplication of matrices consisting of the linguistic probabilities is concerned, It seems that a fuzzy matrix theory [Kim, 1982; Kim and Roush, 1980] is applicable. A fuzzy matrix is a matrix whose entries lie in  $[0,1]$ . The example of a fuzzy matrix is:

$$A = \begin{bmatrix} 0.6 & 0.7 \\ 0.5 & 0.8 \end{bmatrix}$$

The multiplication of fuzzy matrices is defined as follows:

$$A * B = [\sup_k \{\inf(a_{ik}, b_{kj})\}], \quad \text{where } a_{ik} = \text{element of matrix A}$$

$$b_{kj} = \text{element of matrix B}$$

The other operations defined on fuzzy matrices are as follows:

$$\text{addition of matrices: } A + B = [\sup\{a_{ij}, b_{ij}\}]$$

product of scalar and matrix:  $cA = [\inf\{c, a_{ij}\}]$ , where  $c =$  a scalar.

The "sup" and "inf" denote the supremum and infimum defined below, respectively [Pinter, 1971]. The supremum of a set B in a set A is defined as the least upper bound of a set B in a set A. The infimum of a set B in a set A is defined as the greatest lower bound of a set B in a set A. A greatest element in the class of lower bounds of a set B in a set A is called the greatest lower bound of a set B in a set A. A least element in the class of upper bounds of a set B in a set A is called the least upper bound of a set B in a set A.

A scheme to handle the multiplication of matrices consisting of the linguistic probabilities is presented below. Because the basic operations performed on fuzzy sets include intersection and union, our intention is to propose a scheme built around these basic operations.

Kim and Roush [1980] showed that many results of a fuzzy matrix theory are valid for matrices over any commutative semiring. A semiring is a set  $R$  provided with two binary operations '+' and '\*' from  $R \times R$  to  $R$ , which satisfy the following five properties: (1)  $a+b = b+a$ ; (2)  $a+(b+c) = (a+b)+c$ ; (3)  $a*(b*c) = (a*b)*c$ ; (4)  $a*(b+c) = a*b+a*c$ ; and (5)  $(b+c)*a = b*a+c*a$ . A semiring is said to be commutative if the law  $a*b = b*a$  holds.

If we define '+' as union operation, i.e.,  $F(A+B) = \max[F(A), F(B)]$  and '\*' as intersection operation, i.e.,  $F(A*B) = \min[F(A), F(B)]$ , it can be easily shown that a set of linguistic probabilities is a commutative semiring.  $F$  denotes the membership function. Thus, the membership function of the union of  $A$  and  $B$  consists of the maximum values of the membership functions of  $A$  and  $B$ . On the other hand, the membership function of the intersection of  $A$  and  $B$  consists of the minimum values of the membership functions of  $A$  and  $B$ .

Kim and Roush [1980] also showed the fuzzy algebra  $[0,1]$  under the operations  $a+b = \sup\{a,b\}$ ,  $a*b = \inf\{a,b\}$  is a commutative semiring. If we perform three operation on fuzzy matrices  $A_1$  and  $A_2$  using the operations  $a+b = \sup\{a,b\}$  and  $a*b = \inf\{a,b\}$  defined in the fuzzy algebra  $[0,1]$ , we can obtain the following results.

$$A_1 = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 0.5 & 0.6 \\ 0.7 & 0.8 \end{bmatrix}$$

(1) multiplication of  $A_1$  and  $A_2$ :

$$\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} \begin{bmatrix} 0.5 & 0.6 \\ 0.7 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.1*0.5+0.2*0.7 & 0.1*0.6+0.2*0.8 \\ 0.3*0.5+0.4*0.7 & 0.3*0.6+0.4*0.8 \end{bmatrix}$$

(2) addition of  $A_1$  and  $A_2$ :

$$\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.6 \\ 0.7 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.1+0.5 & 0.2+0.6 \\ 0.3+0.7 & 0.4+0.8 \end{bmatrix}$$

(3) product of a scalar and  $A_1$ :

$$(0.5) \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.5*0.1 & 0.5*0.2 \\ 0.5*0.3 & 0.5*0.4 \end{bmatrix}$$

If we define '\*' as the infimum and '+' as the supremum, it can be easily shown that the results of these three operations are consistent with those obtained from the operations performed on fuzzy matrices.

As shown earlier, a set of the linguistic probabilities is a commutative semiring provided that '+' is an union operation and '\*' is an intersection operation. The fuzzy algebra  $[0,1]$  under the operations  $a+b = \sup\{a,b\}$  and  $a*b = \inf\{a,b\}$  is also a commutative semiring. Since the latter commutative semiring produces the results which are identical to those obtained from the operations on fuzzy

matrices, it seems reasonable to define '+' as an union operation and '\*' as an intersection operation for fuzzy matrices consisting of the linguistic probabilities.

Thus, the multiplication of matrices  $A_1$  and  $B_1$  consisting of the linguistic probabilities can be defined as follows:

$$A_1 = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \quad B_1 = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} = \begin{bmatrix} (a_1 \cap b_1) \cup (a_2 \cap b_3) & (a_1 \cap b_2) \cup (a_2 \cap b_4) \\ (a_3 \cap b_1) \cup (a_4 \cap b_3) & (a_3 \cap b_2) \cup (a_4 \cap b_4) \end{bmatrix}$$

where  $a_i$  and  $b_i$  denote the linguistic probabilities, and ' $\cap$ ' and ' $\cup$ ' denote an intersection and union operation, respectively. Although this argument does not fully justify a proposed scheme, it seems acceptable provided that no scheme has been developed.

In the process of developing a proposed scheme, it is shown that if we define '+' as an union and '\*' as an intersection, a set of linguistic probabilities is a commutative semiring. Similarly, if we define '+' as an intersection and '\*' as an union, we can also easily show that a set of linguistic probabilities is also a commutative semiring. If we follow these latter definitions of '+' and '\*', the multiplication of matrices consisting of linguistic probabilities can be defined as follows:

$$\begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} = \begin{bmatrix} (a_1 \cup b_1) \cap (a_2 \cup b_3) & (a_1 \cup b_2) \cap (a_2 \cup b_4) \\ (a_3 \cup b_1) \cap (a_4 \cup b_3) & (a_3 \cup b_2) \cap (a_4 \cup b_4) \end{bmatrix}$$

Given the fuzzy sets  $A = \{0.2/0.1, 0.4/0.5, 0.8/0.9\}$  and  $B = \{0.3/0.1, 0.6/0.5, 1.0/0.9\}$ ,  $A \cup B = \{\max(0.2, 0.3)/0.1, \max(0.4, 0.6)/0.5, \max(0.8, 1.0)/0.9\}$  and  $A \cap B = \{\min(0.2, 0.3)/0.1, \min(0.4, 0.6)/0.5, \min(0.8, 1.0)/0.9\}$ . This shows that the union operation tends to increase the resulting membership values, while the intersection operation tends to decrease the resulting membership values. The fact that the membership values increase as a result of the union operation indicates that the union is a nondecreasing operator, whereas the intersection is a nonincreasing operator.

Because Pearl's algorithms use the probability as a measure of belief, the numbers dealt with by his algorithms lie between 0 and 1. Thus, the addition of the probabilities is a nondecreasing operator, whereas the multiplication of the probabilities is a nonincreasing operator. If we simply consider how the multiplication of the (ordinary) matrices is performed, treating '+' and '\*' as the union and the intersection seems to make more sense than the other way. The belief of node A is computed via the product of  $\delta(A)$  by  $\pi(A)$ . In this computation, two linguistic probabilities need to be multiplied. Because the multiplication of the probabilities and the intersection

operation are both nonincreasing operators, it seems reasonable to use the intersection operation for the product of two linguistic probabilities.

The division operation on fuzzy sets needs to be defined in order to divide one row vector consisting of the linguistic probabilities by another row vector. The proposed scheme for this division operation is provided below. The division operation on matrices A and B consisting of the linguistic probabilities is denoted  $A/B$  and is defined by  $F(A/B) = F(A)/F(B)$ , where  $F(A)$  = membership function of A and  $F(B)$  = membership function of B.

For example, if  $A = \{0.1/0.2, 0.3/0.4, 0.5/0.6\}$  and  $B = \{0.7/0.2, 0.8/0.4, 0.9/0.6\}$ ,  $A/B = \{(0.1/0.7)/0.2, (0.3/0.8)/0.4, (0.5/0.9)/0.6\}$ . If any of the resulting membership values exceeds one, a normalization is necessary. For instance, given  $A = \{0.1/0.2, 0.3/0.4, 0.5/0.6\}$  and  $B = \{0.2/0.2, 0.2/0.4, 0.5/0.6\}$ ,  $A/B = \{0.5/0.2, 1.5/0.4, 1.0/0.6\} = \{(0.5/1.5)/0.2, 1.0/0.4, (1.0/1.5)/0.6\}$ .

#### Fuzzy Number Interpretation

The second approach advocated by researchers including Dubois and Prade [1980a], and Bonissone and Decker [1986] is to treat the linguistic probability as a fuzzy number. A fuzzy number is a fuzzy subset of real numbers. As discussed earlier, the operations performed on the fuzzy numbers are: inverse, scalar, multiplication, exponential,

absolute value, extended addition, extended multiplication, extended subtraction, extended division, extended max, extended min, and extended power function [Dubois and Prade, 1980a].

The extended addition, extended multiplication, and extended division can be performed on the linguistic probabilities in the proposed algorithms in the same manner that the probabilities are manipulated in Pearl's algorithms via the addition, multiplication, and division. One issue that has not been resolved yet is how to deal with a normalizing constant.

Suppose that the linguistic probabilities "unlikely," "maybe," and "likely" are defined as follows:

"unlikely"= {1/0.2, 0.8/0.4, 0.6/0.6, 0.4/0.8, 0.2/1}

"maybe"= {0.6/0.1, 0.8/0.3, 1/0.5, 0.8/0.7, 0.6/0.9}

"likely"= {0.6/0.5, 0.8/0.6, 1/0.7, 0.8/0.8, 0.6/0.9}

If we treat the linguistic probability as a fuzzy number, "unlikely," "maybe," and "likely" could be regarded as "approximately 0.2," "approximately 0.5," and "approximately 0.7," respectively. Thus, the sum of these linguistic probabilities is "approximately 0.2"+ "approximately 0.5"+ "approximately 0.7" which results in "approximately 1.4." This indicates that a normalizing constant is necessary if the linguistic probability is treated as a fuzzy number.

Once we have the conclusion that a normalizing constant



is necessary, the next question occurs as to how we determine a normalizing constant and how we normalize the linguistic probabilities. An approach to resolve this problem is proposed below.

Zadeh [1978a] argued that we can derive the probability distribution from a possibility distribution and Moral [1986] proposed the use of principle of ME in deriving the probability distribution from a possibility distribution. Zadeh's approach and Moral's approach are discussed in detail in Chapter 3 later. A proposed approach is primarily built around these two approaches.

Suppose we add two linguistic probabilities  $L_1$  and  $L_2$  that are represented by  $\{a_1/p_1, a_2/p_2, a_3/p_3, a_4/p_4\}$  and  $\{b_1/p_1, b_2/p_2, b_3/p_3, b_4/p_4\}$ , respectively. And suppose the result obtained from adding  $L_1$  to  $L_2$  is  $\{w_1/p_1, w_2/p_2, w_3/p_3, w_4/p_4\}$ . The first step is to determine the (second-order) probability distribution from this possibility distribution  $\{w_1/p_1, w_2/p_2, w_3/p_3, w_4/p_4\}$ . The resulting probability distribution is denoted  $\{z_1/p_1, z_2/p_2, z_3/p_3, z_4/p_4\}$ , where  $\sum z_i = 1$ . The second step is to compute the mean of this (second-order) probability distribution, i.e.,  $z_1 * p_1 + z_2 * p_2 + z_3 * p_3 + z_4 * p_4$ , where the mean is denoted  $M$ .

If  $M$  is 1, the procedure is terminated. Otherwise, we go to the third step performing a normalization. The third step performs a normalization and produces a normalized  $L_1$  and  $L_2$  which are expressed as  $\{a_1/(p_1/M), a_2/(p_2/M),$

$a_3/(p_3/M)$ ,  $a_4/(p_4/M)$  and  $\{b_1/(p_1/M), b_2/(p_2/M), b_3/(p_3/M), b_4/(p_4/M)\}$ . In other words, in the third step, we multiply each second element of a fuzzy pair by  $(1/M)$ . Its line of reasoning is provided below.

Suppose that a fuzzy number "approximately 3" is defined as  $\{m_1/e_1, m_2/e_2, m_3/e_3, m_4/e_4\}$ . Another fuzzy number defined as  $\{m_1/(2e_1), m_2/(2e_2), m_3/(2e_3), m_4/(2e_4)\}$  denotes a fuzzy number "approximately 6." This can be easily shown by simply applying Moral's approach.

#### Proposed Simulation model

Because it has been concluded that it will not be possible to prove mathematically which interpretation performs better, the use of a simulation model seems acceptable to this kind of situation, even though this technique does not offer conclusive proof. As discussed earlier, Wise's framework compares the results obtained from the non-Bayesian approaches to that obtained from the probability theory supplemented by ME/MXE.

Wise also developed the following conversion scheme between FST and the probability theory: probability = fuzzy membership value. Suppose that a fuzzy set "young" is defined as  $\{1/10, 0.9/20, 0.8/30, 0.5/40\}$ . In Wise's scheme, for example, the probability of "age of 20" is 0.9. Assuming that his scheme is correct, the sum of the membership values should equal 1, because the sum of the

probabilities equals 1. However, in FST, the sum of the membership values is not necessarily equal to one. In fact, in this example, the sum of the membership values is 3.2 ( $=1+0.9+0.8+0.5$ ).

It seems that Wise expected this kind of criticism on his conversion scheme between FST and the probability theory. He [p.44, Wise, 1986] noted "...the objection is often made that fuzzy memberships need not add to one, as must probabilities over an exclusive and exhaustive set of events. The natural reply is that they do not represent probabilities over an exclusive and exhaustive set-it may be incomplete if they add to less than 1, it is not exclusive if they add to more than 1.0."

FST is a generalization of the (ordinary) set theory in that FST allows a partial membership. In the ordinary set theory, for example, given a set  $S = \{1,2,4,6\}$ , the membership values of the elements  $\{1,4,6,7\}$  are 1 (true), 1 (true), 1 (true), and 0 (false), respectively. Thus, the sum of the membership values is 3. The sum of the membership values in the ordinary set theory needs not to be one, rather can be any integer. Since FST is a generalization of the (ordinary) set theory, we can argue that the sum of the membership values needs not to be one in FST. In essence, it has nothing to do with "exclusive and exhaustive set of events."

The probability/possibility consistency principle

introduced by Zadeh [1978a] is discussed first, followed by the presentation of a proposed mapping scheme. Zadeh [1978a] concluded that "Thus, a high degree of possibility does not imply a high degree of probability, nor does a low degree of probability imply a low degree of possibility. However, if an event is impossible, it is bound to be improbable." From "if an event is impossible, it is bound to be improbable," a possibility was interpreted as an upper bound of the probability, leading to the conclusion that the probability distribution can be derived from a possibility distribution, but not vice versa.

This connection between the possibilities and the probabilities is called the probability/possibility consistency principle. After Zadeh introduced the probability/possibility consistency principle, researchers in FST have taken his conclusion regarding the probability/possibility consistency principle for granted. The possibility/probability consistency measure represents the degree of consistency of the probability distribution with a possibility distribution.

Three measures of the probability/possibility consistency developed by Zadeh, Dubois and Prade, and Sugeno are as follows [Moral, 1986]:

$$(1) \text{ Zadeh's concept: } C_2(\text{Pos}, P) = \sum_{a \in U} P(a) * \text{pos}(a), \text{ where pos} =$$

possibility.

(2) Dubois and Prade's concept:

$$C_{DP}(\text{pos}, p) = \begin{cases} 1 & \text{if } \text{POS}(A) \geq P(A), A \in U \\ 0 & \text{otherwise} \end{cases}$$

where  $\text{POS}(A) = \sup_{a \in A} \text{pos}(a)$ ,  $P(A) = \sum_{a \in A} p(a)$ , and  $p =$  probability

(3) Sugeno's measure:

$$C_s(\text{pos}, p) = \int_U f(\text{pos}) \cdot g(p),$$

where  $f$  and  $g$  are non-decreasing mappings  $f, g: [0, 1] \rightarrow [0, 1]$  with  $f(0) = g(0) = 0$ ,  $f(1) = g(1) = 1$  and  $\int_U$  stands for Sugeno's fuzzy integral.

In Zadeh's measure and Sugeno's measure, a higher value of consistency measure implies a high degree of consistency of the probability distribution with a possibility distribution, where the consistency measure does exceed 1.

It seems that the statement "if an event is impossible, it is bound to be improbable" is an overstatement, in the sense that the rather more accurate statement could be "a low degree of possibility implies a low degree of probability." The former is a special case of the latter. From the statements "a high degree of possibility does not imply a high degree of probability." and "a low degree of possibility implies a low degree of probability.", we can conclude that there exists a weak relationship between a possibility distribution and the probability distribution.

Thus, the probability distribution can be derived from a possibility distribution, which is in accordance with

Zadeh's probability/possibility consistency principle. Although it was not discussed by Zadeh [1978a], it makes sense intuitively to say that a high probability implies a high possibility. Two statements, namely, "a high degree of probability implies a high of possibility" and "a low probability does not imply a low possibility" indicate that there exists a weak relationship between the probability distribution and a possibility distribution. Thus, we can conclude that a possibility distribution can be derived from the probability distribution, although it is not in accordance with the commonly accepted statement "a possibility distribution cannot be derived from the probability distribution."

The analytical solution in deriving the probability distribution from a possibility distribution from a possibility distribution are presented first, followed by the discussion of the analytical solutions in deriving a possibility distribution from the probability distribution. Moral [1986] discussed how to construct the probability distribution from a possibility distribution. His scheme is based on ME.

The resulting probability distribution is the solution of the following non-LP problem:

$$\begin{aligned} \text{Maximize } H &= -K \sum_{a \in U} P(a) * \ln(P(a)) \\ \text{s.t.} & \\ & \sum_{a \in U} P(a) = 1 \end{aligned}$$

"P is consistent with possibility," where  $K = \text{constant}$  and  $P = \text{probability}$ . Any of three probability/possibility consistency measures can be substituted into the constraint "P is consistent with possibility."

The above non-LP problem is a fuzzy non-LP problem, because the constraint "P is consistent with possibility" is fuzzy. Especially if Zadeh's measure or Sugeno's measure is used, the constraint "P is consistent with possibility" is fuzzy, because the higher  $\alpha$ , the more "P is more consistent with possibility."  $\alpha$  represents the measure of probability/possibility consistency. On the other hand, the use of Dubois and Prade's measure results in a crisp (or non-fuzzy) non-LP problem. The fuzzy non-LP problem has the limitation in that its computational time is prohibitive.

Thus, Verdegay proposed an approach to solve the  $\alpha$ -cut of the original problem shown below:

$$\begin{aligned} \text{Max:} \quad H &= -K \sum_{a \in U} P(a) * \ln(P(a)) \\ \text{s.t.} \quad \sum_{a \in U} P(a) &= 1 & (1) \\ C(\pi, P) &\geq \alpha \end{aligned}$$

where  $\alpha \in [0, 1]$ ,  $C(\pi, P) = \text{consistency measure}$ ,  $\pi = \text{possibility}$ , and  $P = \text{probability}$ . Thus, in order to obtain the only probability distribution, we can use a single value of  $\alpha$ , e.g., 0.9. The analytical solution of  $\alpha$ -cut of the original problem given in (1) is presented below for three measures of consistency.

First, if Zadeh's measure is used, (1) becomes as follows:

$$\begin{aligned} \text{Max:} \quad & H = -K \sum_{a \in U} P(a) * \ln(P(a)) \\ \text{s.t.} \quad & \sum_{a \in U} P(a) = 1 \\ & \sum_{a \in U} P(a) \pi(a) \geq \alpha \end{aligned} \quad (2)$$

The analytical solution of (2) is provided below [Moral, 1986].

If  $\alpha \in [0,1]$  and  $u$  is the root of  $f(u) = \alpha$  provided that  $f(u) = [\sum \pi(a) * u^{\pi(a)}] / [\sum u^{\pi(a)}]$ , then the solution can be expressed as  $P(a) = u^{\pi(a)} / [\sum u^{\pi(b)}]$ , only if  $\alpha \geq [\sum \pi(a)] / |U|$ .

(3)

$|U|$  denotes the cardinality of universe of discourse  $U$ , namely, the number of elements belonging to  $U$ . However, if  $\alpha < [\sum \pi(a)] / |U|$ ,

then  $P(a) = 1/|U|$ . (4)

Second, if Dubois and Prade's measure is used, (1) becomes the following:

$$\begin{aligned} \text{Max:} \quad & H = -K \sum_{a \in U} P(a) * \ln(P(a)) \\ \text{s.t.} \quad & \sum_{a \in U} P(a) = 1 \\ & II(A) \geq P(A) \end{aligned} \quad (5)$$

Moral [1986] showed that  $II(A) \geq P(A)$  is equivalent to:

$$\pi(b_i) \geq p(b_1) + \dots + p(b_i), \quad i = 1, \dots, n$$

where  $U = \{b_1, \dots, b_n\}$  with  $\pi(b_1) \leq \dots \leq \pi(b_n)$  (6)



He also showed that the following distribution satisfies the Kuhn-Tucker conditions for the problem (5):

$$p(b_1) = \text{Min} \{ \pi(b_i)/i \}$$

$$i=1, \dots, n$$

$$p(b_k) = \text{Min} \{ [\pi(b_i) - p(b_1) - \dots - p(b_{k-1})] / (i-k+1) \}, k=2, \dots, n \quad (7)$$

$$i=k, \dots, n$$

In other words, (7) is an optimal solution of the problem (5). Kuhn-Tucker conditions produce an optimal solution of non-LP problems. For readers interested in Kuhn-Tucker conditions, refer to any advanced book in Operations Research.

Finally, if Sugeno's measure is used, (1) becomes the following:

$$\text{Max}; \quad H = -K \sum_{a \in U} P(a) * \ln(P(a))$$

$$\text{s.t.} \quad \sum_{a \in U} P(a) = 1 \quad (6)$$

$$\int_U f(\pi) \circ g(P) \geq \alpha$$

Moral showed that denoting  $U = \{a_1, \dots, a_n\}$  with  $\pi(a_1) \geq \pi(a_2) \geq \dots \geq \pi(a_n)$ ,

the condition  $\int_U f(\pi) \circ g(p) \geq \alpha$  is equivalent

to  $f(p(a_1) + \dots + p(a_{i_\alpha})) \geq \alpha$ , where  $i(\alpha) = \text{Min} \{ i \mid g(\pi(a_i)) \geq \alpha \}$  (9)

And if we define  $h(\alpha) = \text{Inf} \{ t \in [0, 1] \mid f(t) \geq \alpha \}$ , an optimal solution of the problem (8) is as follows:

if  $i(\alpha)/|U| \geq h(\alpha)$ , then  $p_\alpha(a) = 1/|U|$

$$\text{otherwise, } p_{\alpha}(a_k) = \begin{cases} h(\alpha)/i(\alpha) & \text{if } k \leq i(\alpha) \\ [1-h(\alpha)]/[|U|-i(\alpha)] & \text{if } k > i(\alpha) \end{cases}$$

Some numerical results of the above procedures are shown in Table 5, Table 6, and Table 7 [Moral, 1986].

TABLE 5  
PROBABILITY DISTRIBUTION USING  $C_2$   
(=ZADEH'S MEASURE)

	X1	X2	X3	X4
$\pi$	1	0.9	0.7	0.2
$p$	0.27	0.27	0.27	0.2

TABLE 6  
PROBABILITY DISTRIBUTION USING  $C_{DP}$   
(=DUBOIS AND PRADE'S MEASURE)

		X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>
$\pi$		1	0.9	0.7	0.2
$p_{\alpha}$	$\alpha$				
	0.70	0.25	0.25	0.25	0.25
	0.76	0.30	0.28	0.24	0.17
	0.85	0.41	0.32	0.20	0.07
	1.00	1	0	0	0

TABLE 7  
 PROBABILITY DISTRIBUTION USING  $C_s$   
 (=SUGENO'S MEASURE)

		$X_1$	$X_2$	$X_3$	$X_4$
$\pi$		1	0.9	0.7	0.2
$P_\alpha$	$\alpha$				
	0.76	0.38	0.38	0.12	0.12
	0.85	0.43	0.43	0.07	0.07
	0.95	0.95	0.02	0.02	0.02
	1.00	1	0	0	0

These three tables showed that the derived probability distribution varies depending on which measure is used and/or which value of  $\alpha$  is used.

Similarly, we can derive a possibility distribution from the probability distribution using ME. Thus, a possibility distribution is the solution of the following non-LP problem:

$$\text{Maximize } H = -K \sum_{a \in U} \text{Pos}(a) * \ln(\text{pos}(a)) \quad (10)$$

s.t. "Pos is consistent with P," where pos= possibility and P= probability.

Any of three probability/possibility consistency measures can be substituted into the constraint "Possibility is consistent with probability." Since the use of different consistency measure leads to the different probability distribution. Furthermore, if Zadeh's measure or Sugeno's

measure is used, the same value of  $\alpha$  should be used to maintain the consistency.

The analytical solutions of problem (10) are to be discussed below for three consistency measures. First, suppose that Dubois and Prade's measure is used. Equation (7) indicates that the use of their consistency measure makes it extremely difficult to derive an analytical solution of problem (10). This is because the determination of  $\pi(b_i)$ 's given  $p(b_i)$ 's is extremely difficult due to the fact that we cannot guarantee that  $\pi(b_1) \leq \dots \leq \pi(b_n)$  is satisfied before  $\pi(b_i)$ 's are obtained. This eliminates Dubois and Prade's measure from further consideration for problem (10).

Second, suppose Sugeno's measure is used for problem (10). Equation (9) holds if  $\pi(a_1) \geq \pi(a_2) \geq \dots \geq \pi(a_n)$ . This condition of equation (9) cannot be guaranteed to be true before  $\pi(a_i)$ 's are determined. Thus, we can also conclude that Sugeno's measure is inappropriate for problem (10) in the same manner that Dubois and Prade's measure is not considered as a potential candidate of consistency measure.

Finally, consider Zadeh's measure. The problem (10) becomes the following: Max:  $H = -K \sum_{a \in U} \text{Pos}(a) * \ln(\text{Pos}(a))$

$$\text{s.t } \sum_{a \in U} \text{Pos}(a) * p(a) \geq \alpha, \alpha \in [0,1] \quad (11)$$

The application of extension of the Lagrangian method

to problem (11), provides the following analytical solution. For the readers interested in the extension of Lagrangian method, refer to Taha [1982]. If  $x_i = 0.368$ ,  $i=1, \dots, n$  satisfies the constraint of problem (9), then it is an optimum solution. However, if  $x_i = 0.368$ ,  $i=1, \dots, n$  does not satisfy the constraint of problem (11), then optimal solution is:  $x_i = e^{*\theta p_i - 1}$ ,  $i=1, \dots, n$  such that  $p_1^* [e^{*(\theta p_1 - 1)}] + \dots + p_n^* [e^{*(\theta p_n - 1)}] = \alpha$ . (12)

A proposed scheme to convert the probability into a possibility distribution is presented below, followed by the description of a proposed scheme to convert a possibility distribution into the probability. Suppose that the given probability is  $P_1$ . The first step is to express the (single-valued) probability as the second-order probability distribution.  $P_1$  can be expressed as the following second-order probability distribution:  $\{1/P_1, 0/P_2, \dots, 0/p_n\}$ , where first and second element of a pair denote the second-order probability and first-order probability, respectively.

The second and final step is to derive a possibility distribution from the (second-order) probability distribution using a proposed scheme. For instance, we can derive the equivalent possibility distribution,  $\{u_1/P_1, u_2/P_2, \dots, u_n/P_n\}$  from the probability distribution  $\{1/P_1, 0/P_2, \dots, 0/P_n\}$ , where  $u_i$  denotes the membership value.

The scheme to convert a possibility distribution into the probability is discussed below. This conversion scheme

is used when the final result obtained from the probability theory is compared to that obtained from FST. Suppose the final result obtained from FST is  $\{u_1/p_1+u_2/p_2+\dots+u_n/p_n\}$ .

Using a proposed conversion scheme based on ME, the (second-order) probability distribution  $\{q_1/p_1+q_2/p_2+\dots+q_n/p_n\}$  can be derived from  $\{u_1/p_1, u_2/p_2, \dots, u_n/p_n\}$ . This second-order probability distribution can be converted into a single-valued probability, by computing the mean of the distribution  $\{q_1/p_1+ \dots+q_n/p_n\}$ . Then the mean can be used as the probability-equivalent measure.

The proposed simulation model uses the probability as a measure of belief. The scheme to convert the probability into a possibility distribution derives a (probability-equivalent) possibility distribution (or linguistic probability). Then we can perform the relevant operations on the probabilities and the linguistic probabilities (or possibility distributions).

Different types of operations are performed on the linguistic probabilities depending on whether the linguistic probability is treated as a fuzzy set or a fuzzy number. As the final results, we have the final output expressed as the probability, two possibility distributions for two interpretations. Now we can convert two possibility distributions into a single-valued probability using the proposed scheme. Finally, we can compute mean squared error, mean absolute error, normalized mean squared error,

and normalized mean absolute error of two interpretations. The interpretation with the smaller error is concluded as the best interpretation. The flow chart of proposed simulation model is shown in Appendix.

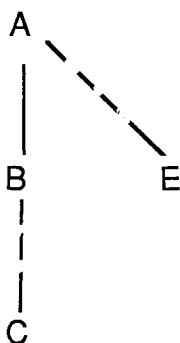
It can be argued that because the proposed schemes between a possibility distribution and the probability distribution are based on the weak relationships between two distributions, the conclusion obtained from this simulation model may not be convincing. However, since the same conversion schemes are applied to both interpretations, it does not seem to affect the results unfairly.

It is also possible to criticize the use of a proposed simulation model, in that the proposition of fuzzy-set-based algorithms is not consistent with the use of Wise's framework as a basis of a simulation model. The argument may be that the fuzzy-set-based algorithms fundamentally favors FST over the probability theory, whereas Wise's line of reasoning is that the probability theory is all we need in uncertainty management in ESSs, nothing else.

Our position in regard to the Bayesian approach and non-Bayesian approaches is to accept the probability theory as a normative approach, but at the same time, admit that the problems with the probability theory, especially the second-order probability theory. This can lead to the adoption of the non-Bayesian approaches in some applications.

### Analysis of Results

An example of Bayesian network discussed earlier in this chapter is adopted in a simulation model. Its Bayesian network is shown below. A node A stands for the identity of the last user of the weapon and a node B represents the identity of the last holder of the weapon.



A node C denotes the possible readings that may be obtained from a fingerprint laboratory and a node E stands for an event producing an alibi. We assume that the alibis are obtained after the information about a fingerprint is obtained. In this experiment, we assume that there are three suspects.

In this experiment, the prior probability distribution of a node A can take either  $(0.8 \ 0.1 \ 0.1)$  or  $(0.6 \ 0.2 \ 0.2)$ .  $P(B|A)$  takes the following three distributions:



$$\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \text{ or } \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} \text{ or } \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}$$

The initial "descendant belief" of node B is (1 1 1), but later changed to either (0.8 0.6 0.5) or (0.6 0.8 0.5). This "descendant belief" is changed due to the result obtained from a fingerprint laboratory. After the alibis for three suspects are obtained, a node E is connected to a node A. The message sent from a node E to a node A takes either (0.1 1.0 1.0) or (0.1 0.1 1.0).

The values employed in this experiment are determined arbitrarily, because our goal is simply to test which interpretation of the linguistic probability generates the better results between a fuzzy number interpretation and the fuzzy set interpretation. Thus, there are thirty-six different cases in this experiment.

As discussed earlier, four measures of error are adopted, namely, the average absolute error, average squared error, average normalized absolute error, and average normalized squared error. In each of thirty-six cases, these four measures are computed. Then the averages of these four measures are used in determining which interpretation performs better.

The averages of these four measures are computed at a node A and node B. That is, we compute the errors between the probability obtained from Pearl's algorithm and the

linguistic probability obtained from the proposed algorithm.  
The results are given below.

Node A:	fuzzy number interpretation	fuzzy set interpretation
average absolute error	0.1382	0.2366
average squared error	0.0473	0.1162
normalized average absolute error	0.5298	0.2959
normalized average squared error	0.8098	0.5948
Node B:	fuzzy number interpretation	fuzzy set interpretation
average absolute error	0.0531	0.2176
average squared error	0.0066	0.1031
normalized average absolute error	0.8117	0.2585
normalized average squared error	0.9590	0.5858

Note that the normalized error takes 1 for zero error,

0 for the random guessing, and -1 for the worst outcome. These results indicate that a fuzzy number interpretation performs better than the fuzzy set interpretation. That is, we can conclude that the operations defined on fuzzy numbers perform better for the linguistic probability than the operations defined on the fuzzy sets.

In fact, these results support Bonissone and Decker's position [1986] that the linguistic probability is treated as a fuzzy number. However, it seems that Bonissone and Decker's assertion that the linguistic probability is treated as a fuzzy number, is not based on any kind of theory or experiment. Thus, this experiment provides the experimental support for Bonissone and Decker's argument, although this experiment does not prove it. Several questions can be raised with regard to this experiment.

The first question is as to whether the result obtained from this simple binary tree is strong (or valid) enough to conclude that a fuzzy number interpretation performs better than the fuzzy set interpretation. Even if the adopted Bayesian network looks like a simple binary tree, lots of computations are performed until the beliefs at node A and node B are determined. Therefore, it can be argued that this simple binary tree can generate the valid result.

The second question is as to whether the result obtained from a simple binary tree can be applied to Pearl's algorithm for the network structure. As shown earlier, the

algorithm performed in the network structure is basically of the same nature to that of an algorithm for a tree structure. Note that we do not attempt to develop the new algorithms for the Bayesian networks, but attempt to resolve the problems associated with the adoption of the linguistic probability.

The third question is as to whether just one experiment can generate the valid results. That is, it can be argued that the additional experiments need to be done by changing the size of the vector. For instance, it could be argued that if the size of the vector is changed to the even number, e.g., 4, we might have the different conclusion.

Although it is not shown here, some experiment is performed for the vector of size 2. Because its results are basically identical to that of the vector of size 3, we do not test thirty-six cases. Based on the experiments performed, we could conclude that the result will be identical regardless of the size of the vector.

The final question can occur as to why we need to use the Bayesian network. That is, we can employ an experiment in which a combination of the addition, multiplication, and division is performed on the linguistic probabilities. It seems that this simple experiment may also produce the meaningful results.

However, the number of operations performed in the binary tree until the final beliefs at node A and node B are

determined, is relatively large. Furthermore, we propose the procedure to perform a normalization. Thus, the use of this simple binary tree enables us to test how well this proposed normalization procedure performs. In this regard, we can argue that the experiment based on the binary tree is more appropriate than the simple experiment in which a collection of addition, multiplication, and division are performed.

### Linguistic Approximation

A linguistic approximation is the process of finding a label whose meaning is the same or the closest to the meaning of an unlabelled fuzzy set generated by the computational model. For example, a fuzzy set  $\{0.3/0.2, 0.4/0.4, 0.5/0.6, 0.8/0.9\}$  could be labelled as "more or less true" or "very true."

Wenstop[1980], Eshragh and Mamdani [1979], Bonissone and Decker [1986], and Bonissone [1979] have developed the procedures for a linguistic approximation. A proposed approach is not meant to be an alternative scheme to all existing approaches, but is a practical scheme which employs the strengths of the existing approaches to a large extent. In fact, a proposed approach is primarily built around Eshragh and Mamdani's approach and Bonissone's approach.

The primary characteristics of a proposed approach are as follows:

(1) It adopts the hedges like Eshragh and Mamdani's approach to increase its granularity.

(2) Like Eshragh and Mamdani's approach, it employs the search step performing a perfect match between an unlabelled set and each primary subset. If a perfect match exists, this step will reduce a search time to a great extent.

(3) After the primary subsets are combined with the hedges, the proposed search procedure consists of two phases, that is, prescreening procedure and a procedure to find the best label using a modified Bhattacharyya distance. The prescreening procedure can cope with the increasing complexity of an exhaustive search. This search procedure will reduce a search time as the number of primary subsets and the hedges increase.

(4) The parametric representation of primary subset [Beyth-Marom, 1966] and three term sets [Bonissone and Decker, 1986], are employed in a proposed approach, where each term set consists of a number of the linguistic probabilities.

(5) As Bonissone [1979] suggested, four features, that is, the power, fuzzy entropy, first moment, and skewness are employed to represent each term.

Bonissone and Decker developed three term sets, i.e., five element term set, nine element term set and thirteen element term set, where each term in a term set is represented by the 4-tuple  $[a, b, \alpha, \beta]$ . For the elements in a term set, the two measures of dispersion, namely, the

interquartile range ( $C_{25} - C_{75}$ ) and the 80 percent range ( $C_{10} - C_{90}$ ), were used to define respectively the interval  $[a, b]$  and  $[(a-\alpha), (b-\beta)]$  of each linguistic probability. The five element term set  $L_1$ , the nine element term set  $L_2$ , and the thirteen element term set  $L_3$  are shown in Table 8, Table 9, and Table 10, respectively.

TABLE 8  
FIVE ELEMENT TERM SET:  $L_1$

impossible	(0 0 0 0)
unlikely	(.01 .25 .01 .1)
maybe	(.4 .6 .1 .1)
likely	(.75 .99 .1 .01)
certain	(1 1 0 0)

The characteristics of a proposed approach are discussed in detail below. The 4-tuple representation is adopted primarily because it is the only representation based on the experiment. The three term sets are employed, mainly because it is likely to increase the chances that we can get the consistent answers from several experts. That

is , from the users' perspective, it facilitates them to choose the appropriate answer from the terms provided.

A proposed approach allows the users to select the term set among three term sets in the actual application. The main advantage obtained from allowing the users to select a term set is given below. These term sets provide the different levels of granularity, i.e., the finest level of distinction among different quantifications of uncertainty.

TABLE 9  
NINE ELEMENT TERM SET:  $L_2$

impossible	(0 0 0 0)
extremely_unlikely	(.01 .02 .01 .05)
very_low_chance	(.1 .18 .06 .05)
small_chance	(.22 .36 .05 .06)
it_may	(.41 .58 .09 .07)
meaningful_chance	(.63 .80 .05 .06)
most_likely	(.78 .92 .06 .05)
extremely_likely	(.98 .99 .05 .01)
certain	(1 1 0 0)

This enables the users to choose a collection of the



linguistic probabilities which has an appropriate level of granularity for the specific application.

TABLE 10  
THIRTEEN TERM SET:  $L_3$

impossible	(0 0 0 0)
extremely_unlikely	(.01 .02 .01 .05)
not_likely	(.05 .15 .03 .03)
very_low_chance	(.1 .18 .06 .05)
small_chance	(.22 .36 .05 .06)
it_may	(.41 .58 .09 .07)
likely	(.53 .69 .09 .12)
meaningful_chance	(.63 .80 .05 .06)
high_chance	(.75 .87 .04 .04)
most_likely	(.78 .92 .06 .05)
very_high_chance	(.87 .96 .04 .03)
extremely_likely	(.98 .99 .05 .01)
certain	(1 1 0 0)

As Eshragh and Mamdani's approach employs the hedges, a proposed approach employs the following hedges: "highly," "very," "more or less," and "slightly." The hedges which are combined with some of the terms in each term set are

shown in Table 11, Table 12, and Table 13.

TABLE 11  
COMBINATION OF THE HEDGES WITH A TERM SET  $L_1$

	highly	very	more or less	slightly
impossible				
unlikely	*	*	*	*
maybe				
likely	*	*	*	*
certain				

The selection of hedges is restricted by a number of hedges which are well-defined and/or whose definitions are consistent among FST researchers. Note that '\*' in these tables indicates that the corresponding hedge is combined with the corresponding term.

First of all, there are not lots of hedges which are well defined. Furthermore, in some cases, the hedges have the different definitions among researchers. As the number of hedges which are well defined increases, we can employ more hedges in this proposed scheme.

TABLE 12  
COMBINATION OF THE HEDGES WITH A TERM SET  $L_2$

	highly	very	more or less	slightly
impossible				
extremely_unlikely			*	*
very_low_chance			*	*
small_chance	*	*	*	*
it_may				
meaningful_chance	*	*	*	*
most_likely			*	*
extremely_likely			*	*
certain				

The sole purpose of employing the hedges in a proposed approach is to increase the level of granularity. When the users use the linguistic probability as the input to the system, the users will be asked whether they want the hedges. Users have the option of choosing some of the available hedges or all the available hedges or no hedges at all. The default option is not to adopt any hedges. Similarly, when the system generates the output, the users have the option regarding the use of hedges in the final output. The default option is not to use any hedges.

TABLE 13  
COMBINATION OF THE HEDGES WITH A TERM SET  $L_3$

	highly	very	more or less	slightly
impossible				
extremely_unlikely			*	*
not_likely	*	*	*	*
very_low_chance			*	*
small_chance	*	*	*	*
it_may				
likely	*	*	*	*
meaningful_chance	*	*	*	*
high_chance			*	*
most_likely			*	*
very_high_chance			*	*
extremely_likely			*	*
certain				

Bonissone and Decker [1986] presented a simplified approach which adopts only two parameters, the power and first moment when these three term sets are used. They did not provide clear reasons why only two features are adopted in a simplified approach. However, since a term set has the small cardinality, i.e., number of terms in a term set is small, the use of four parameters does not seem to pose any

problem in terms of computational time and storage requirements. Thus, the use of four features is adopted in a proposed approach.

Bonissone [1979] who proposed the use of four parameters, did not discuss specifically how to compute these four parameters. The power and entropy can be computed easily from the distribution of the linguistic value. However, the computation of the first moment and skewness is not straightforward due to the fact that these are defined for a random variable which has an associated probability distribution. The schemes to compute the first moment and skewness are presented below.

As discussed earlier, we can derive the probability distribution from a possibility distribution using ME. The proposed approach first derives the probability distribution from a possibility distribution and the first moment and skewness of the derived probability distribution can be the first moment and skewness of the linguistic probability.

As Eshragh and Mamdani's approach adopts an exhaustive search in the first phase of the search procedure by testing a perfect match between the unlabelled set and primary subsets and adopts an heuristic search in the second phase of search procedure to find the best label, a proposed approach adopts the similar strategy.

In a search procedure, the first step checks a perfect match between the unlabelled set and (primary) terms in the

term set. If a perfect match is found, a search procedure is terminated. Otherwise, the second step is activated. This first step can reduce the search time if a perfect match exists. The second step consists of two phases, namely, prescreening and a search of the best label.

The prescreening procedure is performed based on the weighted Euclidean distance. This phase produces a number of terms whose weighted Euclidean distance is within a predetermined tolerance level. The main goal of this prescreening procedure is to reduce the search space in an exhaustive search. The second phase, namely, a search of the best label selects the best label by computing a modified Bhattacharyya distance of each term selected at the first phase. Thus, the main search routine is built around a combination of Eshragh and Mamdani's approach, and Bonissone's approach.

Unless specified otherwise by the users, a single-valued probability corresponding to a term in the term set is provided to the users for reference. This can be achieved by converting the membership function of a term to the probability distribution, followed by the computation of the mean of the second-order probability distribution.

A proposed approach to a linguistic approximation can be summarized as follows:

Step 1 (input):

Users are asked to choose one term set among three term

sets. Also users are asked to indicate whether they want the hedges in the input and output. If the users do not respond with regard to the input, a default option, namely, no use of the hedges is adopted. Similarly, if the users do not respond in regard to the output, also a default option, that is, no use of the hedges is adopted. Furthermore, the users are allowed to choose the hedges they want.

Unless otherwise specified, a single-valued probability corresponding to each term is provided to the users. The users also can assign the weights to four features. If not specified, the default option, namely, equal weights to four features are used.

Step 2 (search and evaluation procedure):

Phase 1: If the unlabelled set shows the characteristics similar to those of terms in the term set, then it will be tested against the appropriate terms in a term set for a perfect match. If a perfect match occurs, then the search is terminated. Otherwise, proceed to phase 2.

Phase 2: First, prescreen all the (linguistic) terms using a weighted Euclidean distance. Users can input the desired tolerance level, if they choose to. Otherwise, the default value is used. Finally, select the best label for the unlabelled set using a modified Bhattacharyya distance.

Step 3 (output):

Generate the output including the best label and its corresponding single-valued probability if needed.

The flow chart of a proposed approach to a linguistic approximation is shown in Appendix.

### Implementation

Two main computer programs are written for this dissertation. The first program is the program performing a simulation used for selecting the best one from a fuzzy set interpretation of the linguistic probability and a fuzzy number interpretation of the linguistic probability.

The second program is the implementation of a proposed linguistic approximation scheme. This program can be added to any system based on FST as a module performing a linguistic approximation. These programs are coded in the programming language C. Specifically, the programming is done in the Turbo C system that is one of the popular C system for personal computer users.

Turbo C system can be used to develop both MS-DOS and UNIX compatible software. The fact it can be used to develop both MS-DOS and UNIX compatible software implies that when we upload C programs developed in Turbo C system into VAX system of the mainframe, the programs can be run correctly under VAX system with little modifications, i.e., C is highly portable. This is in contrast with other programming languages, e.g., PL/I which is different depending on whether IBM system or VAX system is used. Thus, PL/I is said to have less compatibility than C. The



flow charts of these two programs are shown in Appendix. A computer program determining a eigenvector corresponding to the maximum eigenvalue of a given matrix is listed in Appendix.

## CHAPTER IV

### SUMMARY AND CONCLUSIONS

There are two major objectives of this study. The first objective is to determine the operations appropriate for the linguistic probabilities. It is achieved using a proposed simulation model. The second objective is to develop an approach to a linguistic approximation. This proposed approach is implemented in C language.

The conclusion obtained from the simulation model is that the operations defined on the fuzzy numbers perform better for the linguistic probabilities than the operations defined on the fuzzy sets. This conclusion is consistent with the position taken by Bonissone and Decker [1986]. The contribution made by this simulation model is to provide the experimental results to support the argument that the linguistic probability is treated as a fuzzy number.

A proposed scheme to a linguistic approximation employs the four-feature representation, namely, first-moment, entropy, skewness, and power. It also employs the four parameter representation for each term, namely,  $[a, b, \alpha, \beta]$ . It seems that under this four-parameter representation, the skewness is not a good feature, because its variance is very small. On the other hand, first-moment, entropy, and power seem to be good features representing a term.

The contributions made by this dissertation can be summarized as follows:

(1) It provides a basis to consolidate the argument that the linguistic probability is treated as a fuzzy number through the experiment.

(2) The proposed simulation model can be employed in other problems of FST, especially if it is not possible to develop a set of theorems to prove that one approach performs better than the other. Although the results obtained from this simulation model do not prove anything, it at least provides some justifications to favor one approach over the other.

(3) Although a proposed approach to a linguistic approximation is not a totally new approach, it could be a very comprehensive approach in the sense that it attempts to adopt the strengths of Eshragh and Mamdani's approach and Bonissone's approach.

Several problems associated with this research are identified.

(1) In the proposed scheme to a linguistic approximation, we need to derive a possibility distribution from the probability distribution, and vice versa. In these conversions, we need to solve the nonlinear equations. The determination of the analytical solutions is very time consuming, thus can become the bottleneck in terms of computational time.

(2) The proposed scheme to a linguistic approximation employs the hedges. As discussed earlier, the sole reason

to employ the hedges is to increase the granularity of a term set. It is ideal to select the hedges so that the terms are evenly spread out in the closed interval  $[0,1]$ . However, the lack of the hedges in which the definitions are consistent or they are well-defined, restricts the selection of the hedges.

In the future, further research needs to be done on the following problems:

- (1) In a fuzzy number, more research needs to be done in regard to the flattening. Note that the flattening needs to be applied before the operations on fuzzy numbers are performed.
- (2) More research needs to be done in the hedges. That is, more hedges need to be defined and also their definitions need to be consistent among the researchers.
- (3) The development of the analytical solutions in deriving a possibility distribution from the probability distribution, and vice versa, is necessary to make a proposed scheme to a linguistic approximation more practical. This is due to the long computational time in determining the analytical solutions.

## REFERENCES

- Adamo, J.M. (1980a). Fuzzy decision trees, Fuzzy Sets and Systems, 4, 207-219.
- Adamo, J.M. (1980b). L.P.L. a fuzzy programming language: 1. syntactic aspects, Fuzzy Sets and Systems, 3, 151-179.
- Adamo, J.M. (1980c). L.P.L.-a fuzzy programming language: 2. semantic aspects, Fuzzy Sets and Systems, 3, 261-289.
- Adams, J.B. (1984). Probabilistic reasoning and certainty factors, in Rule-based Expert Systems, Buchanan, Bruce G. and Shortliffe, Edward H. (eds.), Addison-Wesley.
- Baldwin, J.F., and S.Q. Zhou (1984). A fuzzy relational inference language, Fuzzy Sets and Systems, 14, 155-174.
- Baldwin, J.F. (1986). Support logic programming, in Fuzzy Set Theory and its Applications, Zimmermann, H.J. (eds.), D. Reidel Publishing Company, 133-170.
- Barnett, Jefferey A. (1981). Computational methods for a mathematical theory of evidence, Proceedings of the seventh international joint conference on Artificial Intelligence, 868-875.
- Berenji, Hamid R. (1987). An integrated approach to reasoning under uncertainty, Proceedings of 18th annual modeling and simulation conference, Pittsburgh.
- Beyth-Marom, R. (1982). How probable is probable? a numerical taxonomy translation of verbal probability expressions, Journal of Forecasting, 1, 257-269.
- Bhatnagar Raj K., and L. N. Kanal (1986). Handling uncertain information: a review of numeric and non-numeric methods, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Black, Paul K. (1987). Is Shafer general Bayes?, Proceedings of the Third AAAI Workshop on Uncertainty in Artificial Intelligence, University of Washington, Seattle.

- Bonissone, P.P. (1979). A pattern recognition approach to the problem of linguistic approximation in system analysis, Proceedings of the IEEE International Conference on Cybernetics and Society, Denver, October, 793-798.
- Bonissone, P.P. (1982). A fuzzy sets based linguistic approach: theory and applications, in Approximate Reasoning in Decision Analysis, Gupta, M.M., and E. Sanchez (eds.), 329-339, North-Holland, New York. Also in Proceedings of the 1980 Winter Simulation Conference, Oren, T.I., C.M. Shub, and P.F. Roth (eds.), 99-111, Orlando, December 1980.
- Bonissone, Piero P. (1987). Summarizing and propagating uncertain information with triangular norms, International Journal of Approximate Reasoning, 1, 71-101.
- Bonissone, Piero P., and Keith S. Decker (1986). Selecting uncertainty calculi and granularity: an experiment in trading-off precision and complexity, in Uncertainty in Artificial Intelligence, Kanal, L.N. and J.F. Lemmer (eds.), North-Holland.
- Borisov, Arkady and Oyar Krumberg (1983). A theory of possibility for decision making, Fuzzy Sets and Systems, 9, 13-23.
- Buchanan, Bruce G., and Edward H. Shortliffe (1984a). The context of the MYCIN experiments, in Rule-based Expert Systems, Buchanan, Bruce G., and Edward H. Shortliffe (eds.), Addison-Wesley.
- Buchanan, Bruce G. and Edward H. Shortliffe (1984b). A model of inexact reasoning in medicine, in Rule-based Expert Systems, Buchanan, Bruce G., and Edward H. Shortliffe (eds.), Addison-Wesley.
- Buchanan, Bruce G. and Edward H. Shortliffe (1984c). Uncertainty and evidential support, in Rule-based Expert Systems, Buchanan, Bruce G. and Edward H. Shortliffe (eds.), Addison-Wesley.
- Cheeseman, Peter (1983). A method of computing generalized Bayesian probability values for expert systems, Proceedings of International Joint Conference on Artificial Intelligence, Karlsruhe, 198-202.
- Cheeseman, Peter (1985). In defense of probability, Proceedings of the Ninth International Joint Conference on Artificial Intelligence, 1003-1009.

- Cheeseman, Peter (1986). Probabilistic versus fuzzy reasoning, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Cohen, Paul R., and Milton R. Grinberg (1983a). A theory of heuristic reasoning about uncertainty, The AI Magazine, Summer, 17-24.
- Cohen, Paul R., and Milton R. Grinberg (1983b). A framework for heuristic reasoning about uncertainty, Proceedings of the International Joint Conference on Artificial Intelligence, 355-357.
- Cox, R. (1946). Probability, frequency and reasonable expectation, American Journal of Physics, 14(1), 1-13.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping, Annals of Mathematical Statistics, 38, 325-339.
- Doyle, Jon (1979). A truth maintenance system, Artificial Intelligence, 12, 231-272.
- Dubois, Didier and Henri Prade (1980a). Fuzzy Sets and Systems: Theory and Applications, Chapter 2, Academic Press.
- Dubois, Didier and Henri Prade (1980b). New results about properties and semantics of fuzzy set-theoretic operators, in Fuzzy Sets: Theory and Applications to Policy Analysis and Information Systems, Wang, Paul P., and S.K. Chang (eds.), Plenum Press, New York.
- Dubois, Didier and Henri Prade (1982). A class of fuzzy measures based on triangular norms, International Journal of General Systems, 8, 43-61.
- Dubois, Didier and Henri Prade (1986). On the unicity of Dempster rule of combination, International Journal of Intelligent Systems, I, 133-142.
- Eshragh F., and E.H. Mamdani (1979). A general approach to linguistic approximation, Int. J. Man-Machine Studies, 11, 501-519.
- Etherington, David W. (1987). Formalizing nonmonotonic reasoning systems, Artificial Intelligence, 31, 41-85.
- Fox, John (1986). Three arguments for extending the framework of probability, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.

- Fung, Robert M., and Chee Yee Chong (1986). Metaprobability and Dempster-Shafer in evidential reasoning, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Gallaire, Herve and Jack Minker (1978). Logic and Databases, Plenum Press, New York.
- Garvey, Thomas D., John D. Lowrance, and Martin A. Fischler (1981). An inference technique for integrating knowledge from disparate sources, Proceedings of Seventh International Joint Conference on Artificial Intelligence, 319-325.
- Genesereth, Michael R., and Nils J. Nilsson (1987). Reasoning with uncertain beliefs (chapter 8), in Logical Foundations of Artificial Intelligence, Morgan Kaufman.
- Ginsberg, Matthew L. (1984). Non-monotonic reasoning using Dempster's rule, Proc. AAAI.
- Gordon, Jean and Edward H. Shortliffe (1984). The Dempster-Shafer theory of evidence, in Rule-Based Expert Systems, Buchanan, Bruce G., and Edward H. Shortliffe (eds.), Addison-Wesley.
- Gordon, Jean and Edward H. Shortliffe (1985). A method for managing evidential reasoning in a hierarchical hypothesis space, Artificial Intelligence, 26, 323-357.
- Grosof, Benjamin N. (1986a). Evidential confirmation as transformed probability: on the duality of priors and updates, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer, North-Holland.
- Grosof, Benjamin N. (1986b). An inequality paradigm for probabilistic knowledge: the logic of conditional probability intervals, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer, North-Holland.
- Gupta, M.M., R.K. Ragade, and R.R. Yager (1979). Report on the IEEE symposium on fuzzy set theory and applications, Fuzzy Sets and Systems, 2, 105-111.
- Hadley, G. (1961). Linear Algebra, Addison-Wesley.



- Heckerman, David (1986). Probabilistic interpretation for MYCIN's certainty factors, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland, Also in Proceedings of AAAI Workshop on Uncertainty and Probability in Artificial Intelligence, August 14-16, 1985, Los Angeles, California.
- Hillier, F.S., and G.J. Lieberman (1980). Introduction to Operations Research, third edition, Holden-Day.
- Hisdal, Ellen (1980). Developments in the wake of the theory of possibility, in Fuzzy Sets: Theory and Applications to Policy Analysis and Information Systems, Wang, Paul P., and S.K. Chang (eds.), Plenum Press, New York.
- Horvitz, Eric J. (1987). Reasoning about beliefs and actions under computational resource constraints, Proceedings of the Third AAAI Workshop on Uncertainty in Artificial Intelligence, Seattle, WA, July, 10-12.
- Horvitz, Eric and David Heckerman (1986). The inconsistent use of measures of certainty in artificial intelligence Research, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Horvitz, Eric J., David E. Heckerman, and Curtis P. Langlotz (1986). A framework for comparing alternative formalisms for plausible reasoning, Proceedings of the fifth AAAI National Conference on Artificial Intelligence, 210-214.
- Hunter, Daniel (1987). Dempster-Shafer vs. probabilistic logic, Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence, University of Washington, Seattle.
- Jaynes, Edwin T. (1979). Where do we stand on maximum entropy?, in The Maximum Entropy Formalism Levine and Tribus (eds.), MIT Press.
- Kandel, Abraham (1986). Fuzzy Mathematical Techniques with Applications, Addison-Wesley.
- Kim, Jin H. and Judea, Pearl (1983). A computational model for causal and diagnostic reasoning in inference systems, Proceedings of IJCAI, 190-193.
- Kim, Ki Hang (1982). Boolean Matrix Theory and Applications, Marcel Dekker Inc., New York and Basel.
- Kim, Ki Hang and Fred W. Roush (1980). Generalized fuzzy matrices, Fuzzy Sets and Systems, 4, 293-315.

- Kuang, P.S. (1986a). Fuzzy logic approach to-- hydraulic system diagnostics and troubleshooting: part 2-- fuzzy relational matrix method, The BFPR Journal, 19, Oklahoma State University, Stillwater, OK.
- Kuang, P.S. (1986b). Fuzzy logic approach to-- hydraulic system diagnostics and troubleshooting: part 3-- self-teaching program for manipulating a fuzzy relational matrix, The BFPR Journal, 19, Oklahoma State University, Stillwater, OK.
- Kuang, P.S. (1986c). Fuzzy logic approach to-- hydraulic system diagnostics and troubleshooting: part 4-- fuzzy theory applications, The BFPR Journal, 19, Oklahoma State University, Stillwater, OK.
- Kunz, John C., Thomas P. Kehler, and Michael D. Williams (1984). Applications development using a hybrid AI development system, AI Magazine, Fall, 41-54.
- Kyburg, Henry E., Jr. (1987). Bayesian and non-Bayesian evidential updating, Artificial Intelligence, 31, 271-293.
- Lemmer, John F. (1986). Confidence factors, empiricism and the Dempster-Shafer theory of evidence, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Loui, Ronald P. (1986). Interval-based decisions for reasoning systems, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Maiers, J., and Y.S. Sherif (1985). Applications of fuzzy set theory, IEEE Transactions on Systems, Man, and Cybernetics, January-February, 175-189.
- Mamdani, E.H. (1976). Advances in the linguistic synthesis of fuzzy controllers, International Journal of Man Machine Studies, 8, 669-678.
- McCain, Roger A. (1983). Fuzzy confidence intervals, Fuzzy Sets and Systems, 10, 281-290.
- Minsky, Marvin. (1975). A framework for representing knowledge, in The Psychology of Computer Vision, edited by Patrick H. Winston, McGraw-Hill, New York.
- Mood, A.M., F.A. Graybill, D.C. Boes (1963). Introduction to the Theory of Statistics, third edition, McGraw-Hill.

- Moral, S. (1986). Construction of a probability distribution from a fuzzy information, in Fuzzy Sets Theory and Applications, Kaufmann, Jones, A., and H-J, D. Zimmermann (eds.), Reidel Publishing Company.
- Newell, Allen, and Herbert A. Simon. (1972). Human problem solving, Prentice-Hall, Englewood Cliffs, NJ.
- Nilsson, Nils J. (1986). Probabilistic logic, Artificial Intelligence, 28, 71-87.
- Pearl, Judea (1982). Reverend Bayes on inference engines: a distributed hierarchical Approach, Proceedings of the National Conference on Artificial Intelligence, August 18-20, Pittsburgh, PA, 133-136.
- Pearl, Judea (1985a). A constraint-propagation approach to probabilistic reasoning, Proceedings Workshop on Uncertainty and Probability in AI, Los Angeles, CA (1985) 31-42, Also in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer, North-Holland, 1986, 357-369.
- Pearl, Judea (1985b). Bayesian network: a model of self-activated memory for evidential reasoning, Seventh Annual Conference of the Cognitive Science Society, 329-334.
- Pearl, Judea (1985c). Learning hidden causes from empirical data, Proceedings of IJCAI, 567-572.
- Pearl, Judea (1986a). Fusion, propagation, and structuring in belief networks, Artificial Intelligence, 29, 241-288.
- Pearl, Judea (1986b). On the logic of probabilistic dependencies, Cognitive Systems Laboratory, Computer Science Department, UCLA, Technical Report R-55-II, CSD-860062, April.
- Pearl, Judea (1986c). On evidential reasoning in a hierarchy of hypotheses, Artificial Intelligence, 28, 9-15.
- Phillips, L. and W. Edwards (1966). Conservatism in a simple probability inference task, Journal of Experimental Psychology, 72, 346-354.
- Pinter, C.C. (1971). Set Theory, Addison-Wesley.

- Prade, Henri M. (1980). Operations research with fuzzy data, in Fuzzy Sets: Theory and Application to Policy Analysis and Information Systems, Wang, Paul P., and S.K. Chang (eds.), Plenum Press, New York.
- Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. (1986). Numerical Recipes: the art of scientific computing, chapter 11, Cambridge University Press.
- Quillian, M.R. (1968). Semantic memory, in Semantic Infor Processing, Minsky, M. (ed), MIT Press, Cambridge, MA.
- Quinlan, J.R. (1983). Inferno: a cautious approach to uncertain inference, The Computer Journal, 26(3), 255-269.
- Rolston, David W. (1988). Principles of Artificial Intelligence and Expert systems Development, McGraw Hill.
- Ruspini, Enrique H. (1986). The logical foundations of evidential reasoning, Technical Note 408, SRI International, December 20.
- Schank, R.C., and Abelson, R.P. (1979). Scripts, plans, goals, and understanding, Lawrence Erlbaum, Hillsdale, NJ.
- Self, Matthew and Peter, Cheeseman (1987). Bayesian prediction for artificial intelligence, Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence, University of Washington, Seattle, 61-69.
- Shafer, G. (1976). A Mathematical Theory of Evidence, Princeton University Press.
- Shore, John E. and Rodney W. Johnson (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, IEEE Transactions on Information Theory, IT-26(1), 26-37.
- Spiegelhalter, David J. (1986). Probabilistic reasoning in predictive expert systems, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Stallings, William (1977). Fuzzy set theory versus Bayesian statistics, IEEE transactions on Systems, Man, and Cybernetics, March, 216-219.

- Szolovits, P. and S.G. Pauker (1978). Categorical and probabilistic reasoning in medical diagnosis, Artificial Intelligence, 11, 115-144.
- Taha, H.A. (1982). Operations Research: An Introduction, third edition, MacMillan Publishing Co.
- Tribus, Myron (1980). Fuzzy sets and Bayesian methods applied to the problem of literature search, IEEE Transactions on Systems, Man, and Cybernetics, SMC-10(8), 501-502.
- Trillas, E., and T. Riera (1978). Entropies in finite fuzzy sets, Information Sciences, 15, 159-168.
- Wenstop, Fred (1980). Quantitative analysis with linguistic values, Fuzzy Sets and Systems, 4, 99-115.
- Winston, Patrick Henry (1984). Artificial Intelligence, Second Edition, Addison-Wesley.
- Wise, Ben P. (1986). An Experimental Comparison of Uncertainty Inference Systems, Ph.D dissertation, Department of Engineering and Public Policy, Carnegie Mellon University.
- Wise, B.P. and Max Henrion (1986). A framework for comparing uncertain inference systems to probability, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F.Lemmer (eds.), North Holland.
- Yager, R.R. (1977). Multiple objective decision-making using fuzzy sets, Int J. Man-Machine Studies, 9, 375-382.
- Yager, R.R. (1984). Approximate reasoning as a basis for rule-based expert systems, IEEE Transactions on Systems, Man, and Cybernetics, SMC-14(4), 636-643.
- Zadeh, Lofti A. (1965). Fuzzy sets, Information and Control, 8, 338-353.
- Zadeh, Lotfi A. (1972). A fuzzy-set-theoretic interpretation of linguistic hedges, Journal of Cybernetics, 2(3), 4-34.
- Zadeh, Lofti A. (1973). Outline of a new approach to the analysis of complex systems and decision processes, IEEE Transactions on systems, Man, Cybernetics, SMC-3(1), 28-44.

- Zadeh, Lofti A. (1975a). The concept of a linguistic variable and its application to approximate reasoning, Information Sciences, Part I: 8, 199-249; Part II: 8, 301-357; Part III: 9, 43-80.
- Zadeh, Lofti A. (1975b). Fuzzy logic and approximate reasoning, Synthese, 30, 407-428.
- Zadeh, Lofti A. (1978a). Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems, 1, 3-28.
- Zadeh, L.A. (1978b). PRUF-a meaning representation language for natural languages, International Journal of Man-Machine Studies, 10, 395-460.
- Zadeh, Lofti A. (1983a). A fuzzy-theoretic approach to the compositionality of meaning: propositions, dispositions and canonical forms, Cognitive Science Program, Institute of Cognitive Studies, University of California at Berkeley, December.
- Zadeh, Lofti A. (1983b). A computational approach to fuzzy quantifiers in natural languages, Com. & Maths. with Appls. 9(1), 149-184.
- Zadeh, Lofti A. (1983c). The role of fuzzy logic in the management of uncertainty in expert systems, Fuzzy Sets and Systems, 11, 199-227.
- Zadeh, Lofti A. (1983d). Linguistic variables, approximate reasoning and dispositions, Med. Inform., 8(3), 173-186.
- Zadeh, Lofti A. (1986a). A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination, The AI Magazine, Summer, 85-90.
- Zadeh, Lofti A. (1986b). Is probability theory sufficient for dealing with uncertainty in AI: a negative view, in Uncertainty in Artificial Intelligence, Kanal, L.N., and J.F. Lemmer (eds.), North-Holland.
- Zimmer, A.C. (1985). The estimation of subjective probabilities via categorical judgments of uncertainty, Proceedings of the Workshop on Uncertainty and Probability in Artificial Intelligence, UCLA, Los Angeles, CA, August 14-16, 217-224.
- Zimmermann, H. -J. (1986). Fuzzy set theory and mathematical programming, in Fuzzy Set Theory and Applications, Jones, A. et al., R. Reidel Publishing company, 99-114.

## APPENDIX

### FLOW CHARTS OF THE PROGRAMS

Two programs are to be implemented, namely, simulation program and a program performing a linguistic approximation, in this dissertation. A simulation program is used to select a set of operations appropriate for the linguistic probabilities. The flow chart of a simulation program is presented first, followed by the description of the flow chart of a program performing a linguistic approximation.

Because the actual coding is generally straightforward, the listings of the entire programs are not shown here. However, a program to determine the weights of four parameters is listed here, because it seems worth being listed. A program to determine the weights of four parameters is adopted in a program performing a linguistic approximation.

The flow Chart of the simulation model is shown in Figure 15 and the flow chart for a program performing a linguistic approximation is shown in Figure 16.

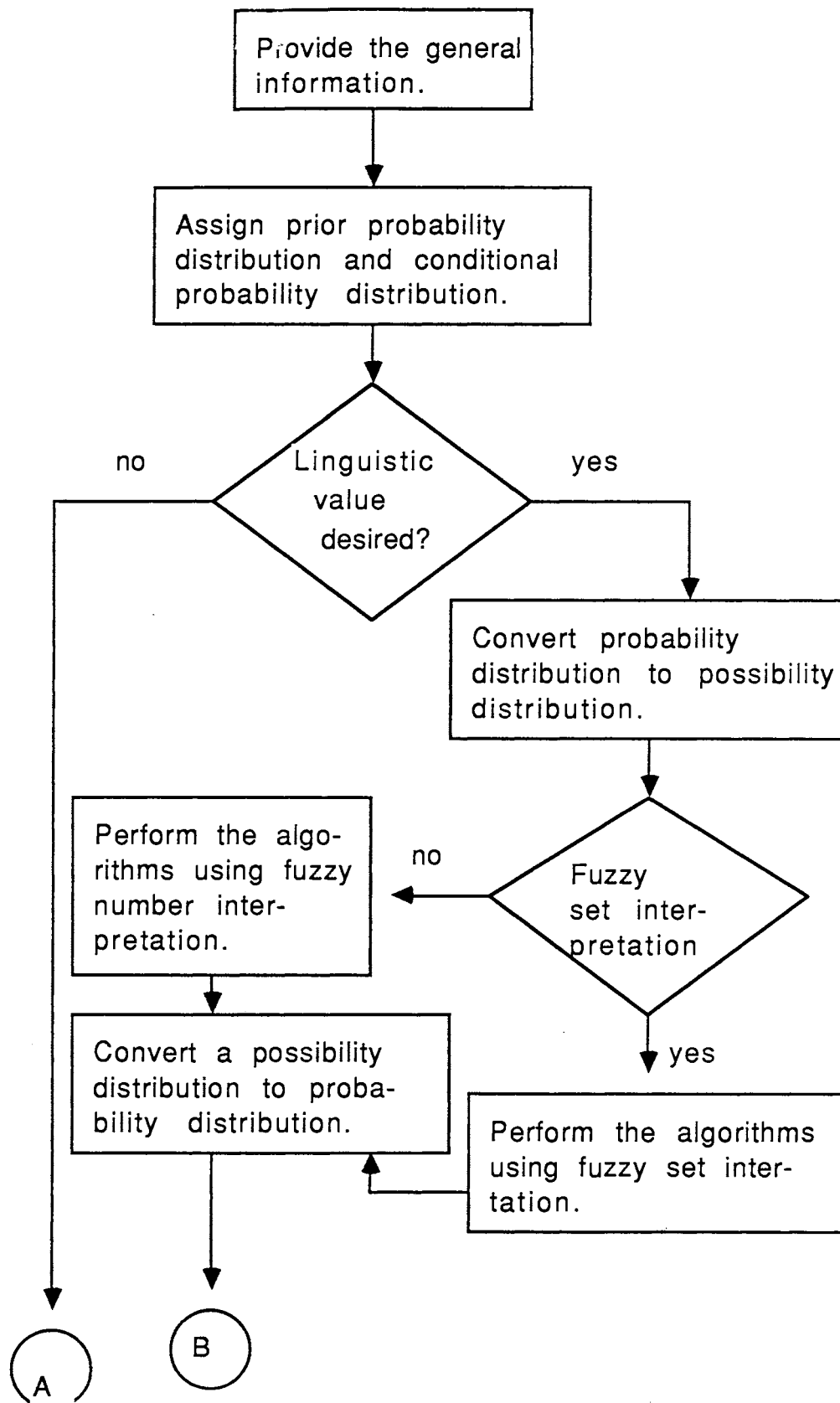
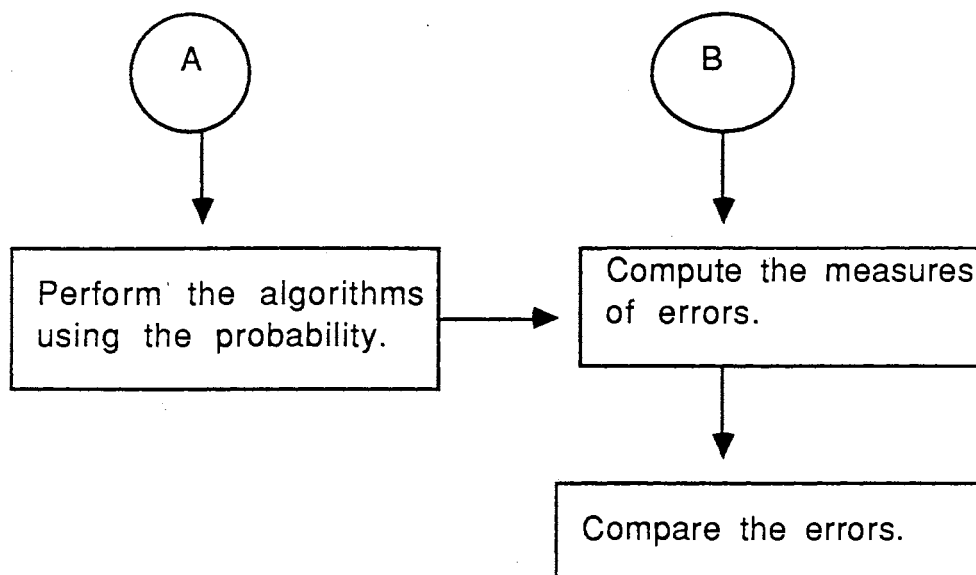


Figure 15. Flow Charts of a Simulation Model



Figure 15. (Continued)



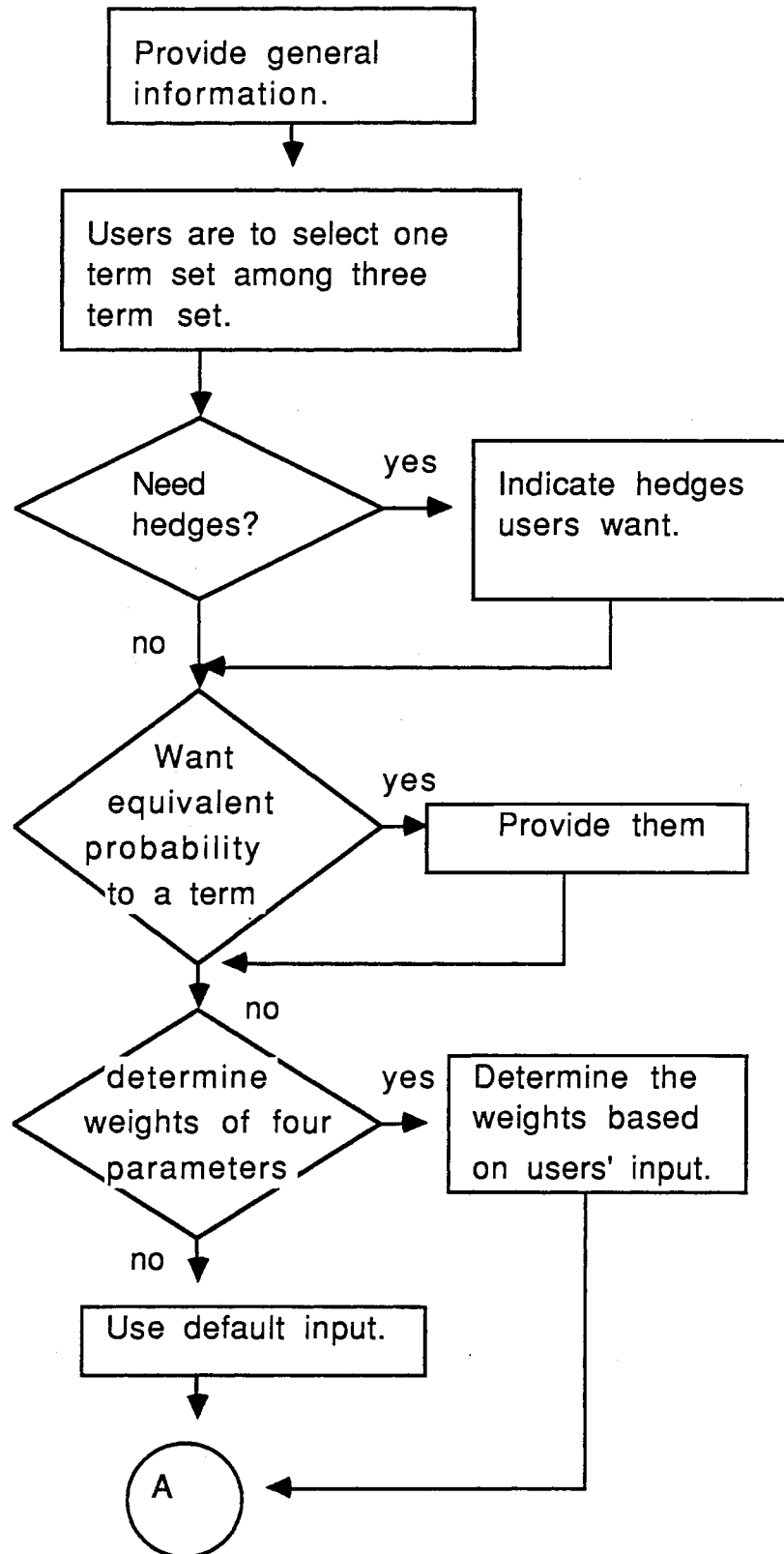
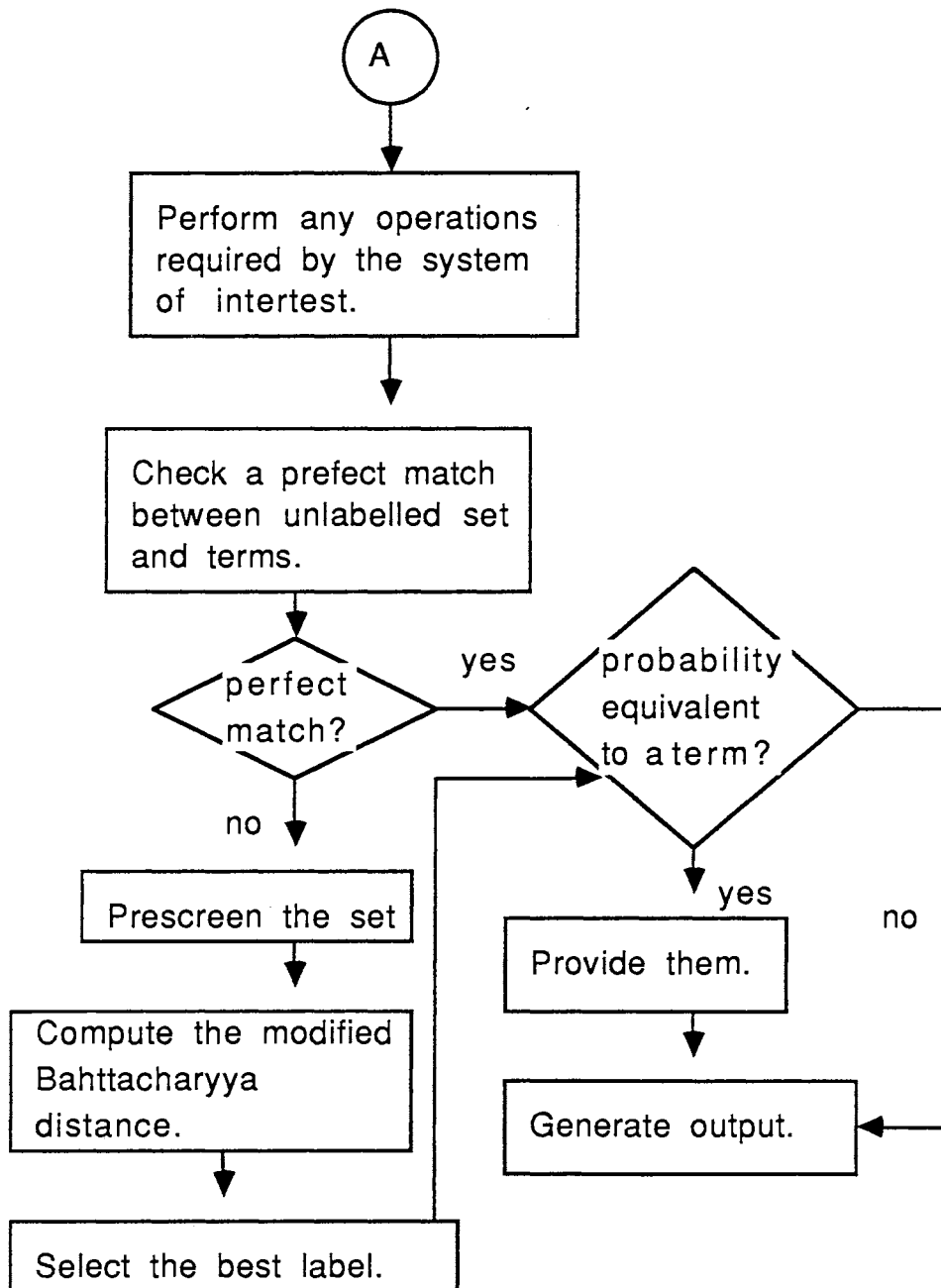


Figure 16. Flow Chart of Proposed Scheme of a Linguistic Approximation

Figure 16. (Continued)



```
/******  
* This routine determine the eigenvector corresponding to*  
* a maximum eigenvalue of a given (square) matrix.      *  
******/  
{  
    float b[4],r[4],e[4]; /* e: elements of eigenvector */  
    float a[4][4]; /* a given matrix */  
    float sum,eigenvalue;  
    int i,j,cnt=0, found=0;  
    int n=4; /* size of matrix */  
  
    for (i=0; i < n; i++)  
        b[i]= 1.0/ ((float) n);  
    for (i=0; i < n; i++) {  
        sum=0.0;  
        for (j=0; j < n; j++)  
            sum += b[j]* a[i][j];  
        r[i]= sum;  
    }  
    sum=0.0;  
    for (i=0; i < n; i++)  
        sum += r[i];  
    for (i=0; i < n; i++)  
        e[i]= r[i]/ sum;  
}
```

VITA

Jong-Nam Lee

Candidate for the Degree of

Doctor of Philosophy

Thesis: PROPOSED ALGORITHMS FOR BAYESIAN NETWORKS BASED ON  
FUZZY SET THEORY

Major Field: Business Administration

Biographical:

Personal Data: Born in Seoul, Korea, January 24, 1955,  
the son of Nak-Joon Lee and Jung-Suk Suh, married  
with one child.

Education: Graduate from Chung-Dong High School, Seoul,  
Korea, in January 1973; received Bachelor of  
Science Degree in Industrial Engineering from  
Seoul National University, Seoul, Korea, in  
February 1977; received Master of Business  
Administration from State University of New York  
at Buffalo, Buffalo, NY, in June, 1981; completed  
requirements for the Doctor of Philosophy at  
Oklahoma State University in July, 1989

Professional Experience: Lecturer, Department of  
Business Administration, Kyung Hee University,  
Seoul, Korea, September 1981 to July 1983;  
Assistant Professor, Department of Business  
Administration, Kyung Hee University, Seoul,  
Korea, August 1983-August 1984; Graduate Teaching  
Associate, Department of Management, Oklahoma  
State University, August 1984-December 1988.