Uncertainty in Artificial Intelligence

PH.D. COMPREHENSIVE EXAM

Mohammad Shayganfar - mshayganfar@wpi.edu

May, 26 2015

1 Introduction to Uncertainty in AI

2 Sources and Types of Uncertainties

-Sources:

+ Noise (imprecise observation)
+ Uncertain Change (unprediatable or stochastis behavior of the world)
+ Incompleteness or igonorance (missing information)

3 Theories of Uncertainties

3.1 Bayesian Networks

3.1.1 Types of Reasoning

3.1.2 Probability Theory

3.2 Dempster-Shafer Theory

In [3], Dempster proposed a probabilistic framework based on lower and upper bounds on probabilities. In [8], Shafer developed a formalism for reasoning under uncertainty which uses some of Dempster's mathematical expressions with different interpretation. Based on Shafer's formalism, each piece of evidence may support a subset containing several hypotheses. This is a generalization of the pure probabilistic framework in which every finding corresponds to a value of a variable (a single hypothesis) [4]. Therefore, Dempster-Shafer theory is the generalization of the Bayesian theory of subjective probability to combine accumulative evidence or to change prior opinions in the light of new evidence [2]. Dempster-Shafer theory is designed to deal with the distinction between uncertainty and ignorance. Rather than computing the probability of a proposition, it computes the probability that the evidence supports the proposition [6], and it does not require the assumption that $Belief(\mathrm{A}) + Belief(\neg \mathrm{A}) = 1$. Dempster-Shafer theory deals

with the possible values of an unknown variable, just as deos the theory of probability [10].

There are three basic functions in the Dempster-Shafer theory that we need to understand for modeling purposes, *mass function, belief function*, and *plausibility function*. Let $\Theta = \{h_1, h_2, \ldots, h_n\}$ be a finite set of hypotheses. This set of hypotheses is also called *frame of discerntment*. The hypotheses represent all the possible states of the system considered. The set of all subsets of Θ is its *power set*: 2^Θ. A subset of these 2^Θ sets may consist of a single hypothesis or of a conjunction of several hypotheses (e.g., a snowy day and a dry day). The pieces of evidence are events that occurred or may occur (e.g., high pressure shown by a barometer, or low temprature). One piece of evidence can be related to a single hypothesis or a set of hypotheses. However, it is not allowed to have different pieces of evidence lead to the same hypothesis or set of hypotheses. In fact, the relation between a piece of evidence and a hypothesis corresponds to a cause-effect chain, i.e., a piece of evidence implies a hypothesis or a set of hypotheses [5]. Moreover, it is required that all hypotheses are unique, not overlapping and mutually exclusive.

3.2.1 Mass Function

A *Basic Probability Assignment* (BPA) or *mass function* is a function $m : 2^\Theta \rightarrow [0, 1]$ such that:

$$m(\emptyset) = 0, \text{ and } \sum_{x \in 2^\Theta} m(x) = 1.$$

The value 0 indicates no belief and the value 1 indicates total belief, and any value between these two indicate partial belief. As you see the mass function uses the notion of 2^Θ to be able to use all possible subsets of the *frame of discernment* Θ. All of the assigned probabilities sum to unity. There is no belief in empty set. Any subset x of the frame of discernment Θ for which m(x) is non-zero is called a *focal element* and represents the exact belief in the proposition depicted by x. Thus, any subset is proposition and vice versa. Other elements in Dempster-Shafer theory are defined by mass function.

3.2.2 Belief Function

Now, we can define another important notion in Dempster-Shafer theory, the *belief function* (sometimes called a *support function*). It is the measure

of total belief committed to $A \subseteq \Theta$ that can be obtained by simply adding up the mass of all the subsets of A. In other words, given the frame of discernment Θ and $A \subseteq \Theta$, the belief in A, denoted $Belief(A)$, is a number in the interval [0, 1]. Belief in a set of elements, say A, of a frame Θ, represents the total belief that one has based on the evidence obtained. Unlike probability theory, $Belief(A) = 0$ represents lack of evidence about A, while $p(\mathrm{A}) = 0$ represents the impossibility of A. However, $Belief(A) = 1$ represents certainty, that is A is certain to occur, similar to $p(\mathrm{A}) = 1$, which also represents the certainty that A is true. A belief function defined on a space Θ must satisfy the following three properties:

$$Belief(\emptyset) = 0$$

$$Belief(\Theta) = 1$$

$$Belief(A_1 \cup \ldots A_n) \geq \sum_i Belief(A_i) \text{ - } \sum_{i<j} Belief(A_i \cap A_j) + \ldots + (-1)^{n+1} Belief(A_i \cap \ldots \cap A_n)$$

A belief function is a function $Belief : 2^{\Theta} \rightarrow [0, 1]$ and is defined by:

$$Belief(A) = \sum_{B \subseteq A} m(B) \qquad \text{for all } A \subseteq \Theta$$

3.2.3 Plausibility Function

Plausibility in a set, say A of a frame Θ consisting of a mutually exclusive and exhaustive set of elements, represents the maximum possibility that a set A is true given all the evidences. A plausibility function *Plausible* defined on a space Θ must satisfy the following three properties:

$$Plausible(\emptyset) = 0$$

$$Plausible(\Theta) = 1$$

$$Plausible(A_1 \cap \ldots A_n) \leq \sum_i Plausible(A_i) \text{ - } \sum_{i<j} Plausible(A_i \cup A_j) + \ldots + (-1)^{n+1} Plausible(A_i \cup \ldots \cup A_n)$$

A *plausibility* measure is a function $Plausible : 2^{\Theta} \rightarrow [0, 1]$, and is defined by:

$$Plausible(A) = \sum_{B \cap A \neq \emptyset} m(B) \qquad \text{for all } A \subseteq \Theta$$

Plausible(A) in a subset A is defined to be the sum of all mass functions for the subsets B that have non-zero intersections with A, and it represents the extent to which we fail to disbelieve A. In other words, it corresponds to the total belief that does not contradict A. The plausibility and belief functions are related to one another, and we can represent this relation as:

$$Belief(A) = 1 - Plausible(\neg A) \quad \text{and} \quad Plausible(A) = 1 - Belief(\neg A),$$

where $\neg A$ is A's complement. Also, $Belief(\neg A)$ is often called the *doubt* in A. It is noteworthy to mention that Dempster-Shafer theory allows the representation of *ignorance* since $Belief(A) = 0$ does not imply $Belief(\neg A) > 0$ even though $Belief(\neg A) = 1$ implies $Belief(A) = 0$. Other notable relations are:

$$Belief(A) + Belief(\neg A) \leq 1, \text{ and}$$

$$Plausible(A) + Plausible(\neg A) \geq 1.$$

Here, we also note that in the case of each of the focal elements being singletons then we return back to traditional Bayesian analysis incorporating normal probability theory, since in this case $Belief(A) = Plausible(A)$ [1].

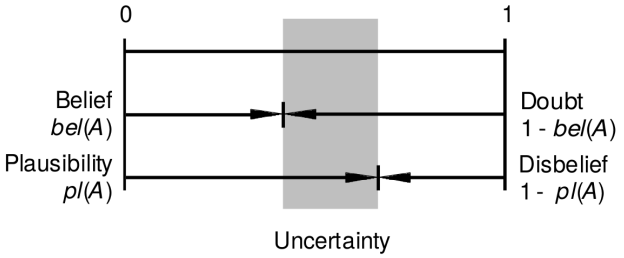

Figure 1: Measures of belief and plausibility. The uncertainty interval is shaded gray. [5].

Collectively the above measures provide Dempster-Shafer theory with an explicit measure of ignorance about A and its complement. All the above measures of confidence and the BPA are equivalent, in the sense that each of them can be expressed as a function of any one of the rest. The *uncertainty* measure is defined as the length of the interval $[Belief(A), Plausible(A)]$ where $Belief(A) \leqslant Plausible(A)$ [11], and it is also called as *belief interval*. Figure 1 illustartes a graphical representation of the belief, plausibility, and

doubt measures which we defined above. As it is shown and said earlier, the difference between plausibility and belief describes the evidential interval range which represents the uncertainty concerning the set A. Also, as we see in Figure 1, lack of belief does not imply disbelief, since the complements of belief and plausibility are doubt and disbelief, respectuvely. Furthermore, the mass assigned to Θ can be interpreted as the global ignorance, since the level of mass value is not discernible among the hypotheses.

3.2.4 Dempster's Rule of Combination

Suppose that we have two pieces of uncertain evidence relevant to the same frame of discerntment Θ. Dempster-Shafer theory also provides a method to combine the measures of evidence from different sources, using Dempster's rule of combination which combines two pieces of evidence into a single new piece. The rule assumes that the sources are independent. If m_1 and m_2 are the BPA's associated with Bel_1 and Bel_2 respectively and Bel_1 and Bel_2 are independent, then Dempster's rule of combination is as follows:

$$[m_1 \oplus m_2](y) = \begin{cases} 0, & y = \emptyset \\ \dfrac{\sum\limits_{A \cap B = y} m_1(A) m_2(B)}{1 - \sum\limits_{A \cap B \neq \emptyset} m_1(A) m_2(B)}, & y \neq \emptyset \end{cases}$$

The numerator, i.e., $\sum_{A \cap B = y} m_1(A) m_2(B)$, represents the accumulated evidence for the sets A and B, which supports the given hypothesis y. The denominator in the Dempster's rule of combination, i.e., $1 - \sum_{A \cap B \neq \emptyset} m_1(A) m_2(B)$, is an important normalization factor denoted by $\mathcal{K}$ which can be interpreted as a measure of conflict between the sources [9].

3.3 Fuzzy Logic Theory

Fuzzy Logic provides a mathematical framework to capture uncertainty, and it was introduced by Zadeh in 1965 [12]. Basically, Fuzzy Logic is a multivalued logic, that allows intermediate values to be defined between conventional evaluations like "true" and "false". Fuzzy Logic's ultimate goal is to provide foundations for approximate reasoning using imprecise propositions based on fuzzy set theory. In order to deal with such imprecise inference, Fuzzy Logic allows the imprecise linguistic terms such as: fuzzy

predicates (e.g., old, expensive), fuzzy quantifiers (e.g., many, little), and fuzzy truth values (e.g., unlikely false or unlikely true). Fuzzy Logic is a method for reasoning with logical expressions describing membership in fuzzy sets [6]. Logics as bases for reasoning can be distinguished essentially by three items: truth values, operators, and reasoning procedures (e.g., tautologies) [13]. For instance, in dual logic, truth values can be "true" (1) or "false" (0), operators can be defined using the truth tables, and modus ponens or contrapositions can be considered as tautology. In Fuzzy Logic, the truth values are no longer restricted to two values, but are expressed by the linguistic variables such as "true" or "false". In all forms of fuzzy reasoning, the implications can be modeled in different ways.

```
1. Define the linguistic variables and terms (initialization)
2. Construct the membership functions (initialization)
3. Construct the rule base (initialization)
4. Convert crisp input data to fuzzy values
   using the membership functions (fuzzification)
5. Evaluate the rules in the rule base (inference)
6. Combine the results of each rule (inference)
7. Convert the output data to non-fuzzy values (defuzzification)
```

Figure 2: Fuzzy Logic algorithm.

Figure 2 shows the Fuzzy Logic algorithm. It begins with initialization of linguistic varibales (see Section 3.3.2) and constructing appropriate membership functions (see Section 3.3.1) and rule-base of the fuzzy system (see Section 3.3.6). The constructed membership functions transfer the input data to fuzzy values (see Section 3.3.8). Then, the inference system evaluates the constructed rules with respect to the given input value, and merges the results obtained from each rule. Finally, the ovelall result will be transfered to a non-fuzzy (scalar) value (see Section 3.3.10).

3.3.1 Membership Functions

The difference between crisp (i.e., classical) and fuzzy sets is established by introducing a *membership function*. Membership functions are mathematical tools for indicating flexible membership to a set, modeling and quantifying the meaning of symbols. Membership functions are used in the fuzzification and defuzzification steps (see Sections 3.3.8 and 3.3.10) of a Fuzzy Logic system. A membership function is used to quantify a linguistic term (see Section 3.3.2). Therefore, the manipulation of fuzzy quantities

can be accomplished by manipulation of fuzzy set membership functions. Some of the manipulation includes set complement, intersection, and union as well as fuzzification and defuzzification (see Sections 3.3.8 and 3.3.10) [7].

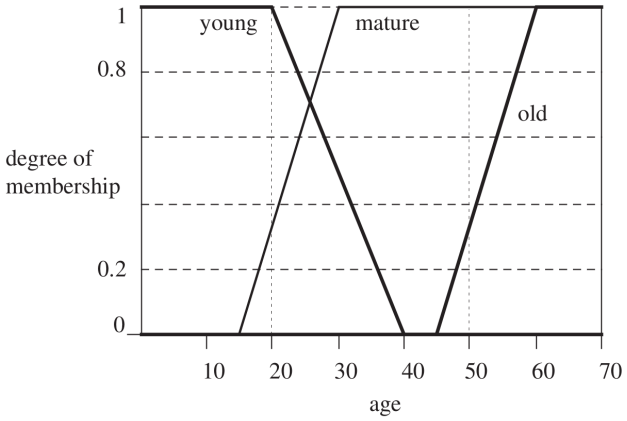

Figure 3: Membership functions for the concepts young, mature and old.

Figure 3 shows membership functions for three linguistic terms of age variable. It shows three examples of a membership functions in the interval 0 to 70 years. These three functions define the degree of membership of any given age in the sets of young, mature, and old ages. Note that, an important characteristic of fuzzy logic is that a numerical value does not have to be fuzzified using only one membership function. In other words, a value can belong to multiple sets at the same time. For insatnce, if someone is 20 years old her degree of membership in the set of young persons is 1.0 (maximum value), in the set of adults 0.35, and in the set of old persons 0.0 (minimum value). As another example, if someone is 50 years old the degrees of membership in the sets of young, mature, and old are 0.0, 1.0, 0.3 respectively.

The membership function is a graphical representation of the magnitude of participation of each input. It associates a weighting with each of the inputs that are processed, define functional overlap between inputs, and ultimately determines an output response.

Let $\mathcal{X}$ be a classical universal set. A fuzzy subset A of $\mathcal{X}$ is characterized by a *membership function*: $\mu_A : \mathcal{X} \rightarrow [0, 1]$. $\mu_A(x)$ is called the *membership degree* of x in A. The degree of membership is expressed by a real number in the interval [0, 1]. The followings define some of the important properties of membership function.

Height: The height of A, denoted by $h(A)$, corresponds to the upper bound of the membership function:

$h(A) = \sup\{\mu_A(x) | x \in \mathcal{X}\}$.

Support: The support of A is a set of all elements x of $\mathcal{X}$ for which (x, $\mu_A(x)) \in A$ and $\mu_A(x) > 0$ holds.

α-cut: An α-cut of A is the subset of elements with a membership degree greater than or equal to α. The α-cut is denoted by:

$\alpha\text{-cut}(A) = \{x \in \mathcal{X} | \mu_A(x) \geqslant \alpha\}$.

kernel:

Membership functions can have different shapes and their shape can be determined arbitrarilty based on experience or by running statistical studies on data. They can be sigmoidal, heyperbolic, Gaussian or any other shape.

The rules use the input membership values as weighting factors to determine their influence on the fuzzy output sets of the final output conclusion.

3.3.2 Linguistic Variables

The concept of membership function discussed in Section 3.3.1 allows us to define fuzzy systems in natural language. Linguistic variables are the input or output variables of the system whose values are words or sentences from a natural language, instead of numerical values. A linguistic variable is generally decomposed into a set of linguistic terms. For instance, for people's age, we usually use terms such as "old" or "young" which are called linguistic values of the age. Then, we can consider a set of decompositions for the linguistic variable age, *Height*(h) = very-old, old, mature, young, very-young. The members of this decomposition set are called linguistic terms which can cover a portion of the overall values of people's age.

3.3.3 Fuzzy Operators

In order to easily manipulate fuzzy sets, we are redefining the operators of the classical set theory to fit the specific membership functions of fuzzy logic for values strictly between 0 and 1.

Unlike the definitions of the properties of fuzzy sets that are always the same, the definition of operators on fuzzy sets is chosen, like membership functions. Here are the two sets of operators for the complement (NOT), the intersection (AND) and union (OR) most commonly used:

3.3.4 Fuzzy Sets

A fuzzy set A is defined by a membership function μ_A from the universe of discourse X_i to the closed unit interval [0,1]. We interpret $\mu_A(x)$ as the degree of membership of x in A.

3.3.5 Fuzzy Relations

3.3.6 Fuzzy Rules

3.3.7 Fuzzy Propositions

3.3.8 Fuzzification

3.3.9 Reasoning in Fuzzy Logic

3.3.10 Defuzzification

3.3.11 Types of Fuzzy Functions

In summary, Figure 4 (see also Figure 2) shows the process of fuzzy logic. Firstly, a crisp set of input data are gathered and converted to a fuzzy set using fuzzy linguistic variables (see Section 3.3.2), fuzzy linguistic terms and membership functions (see Section 3.3.1). This step is known as fuzzification (see Section 3.3.8). Afterwards, an inference is made based on a set of rules (see Section 3.3.9). Lastly, the resulting fuzzy output is mapped to a crisp output using the membership functions, in the defuzzification step (see Section 3.3.10).

3.4 Other approaches

4 Strengths and Weaknesses

In general, there is an increasing trend of computational complexity Fuzzy Logic to probabilistic approaches and Dempster-Shafer theory. However, the representational power and precision increases in the same order and direction.

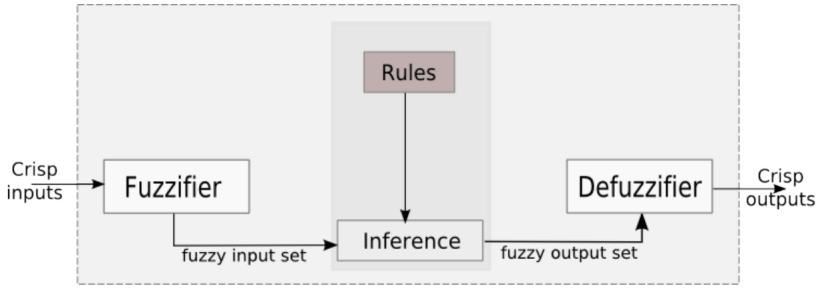

Figure 4: A Fuzzy Logic system.

- Locality in rule-based systems vs. using all evidences in probabilistic systems [R&N AI book p.524]

-Detachment in rule-based systems vs. requiring the source of evidence for subsequent probabilistic reasoning [R&N AI book p.524]

- Dempster-Shafer theory allows no definite decision in many cases, whereas probabilistic inference does yield a specific choice [6].

- In contrast to Dempster-Shafer theory, a complete Bayesian model would include probability estimates for factors that allow us to express the ignorance in terms of how our beliefs would change in the face of future informaiton gathering [6].

4.1 Advantages and Disadvantages of Belief Networks

Like any other computational formalism, belief network technology offers certain advantages and disadvantages. Advantages of belief networks include [2]:

- Sound theoretical foundation: The computation of beliefs using probability estimates is guaranteed to be consistent with probability theory. This advantage stems from the Bayesian update procedures strict derivation from the axioms of probability.

- Graphical models: Belief networks graphically depict the interdependencies that exist between related pieces of domain knowledge, enhancing understanding of the domain. The structure of a belief network captures the cause-effect relationships that exist amongst the variables of the domain. The ease of causal interpretation in belief network models typically makes them easier to construct than other

models, minimizing the knowledge engineering costs and making them easier to modify.

- Predictive and diagnostic reasoning: Belief networks combine both deductive/predictive and abductive/diagnostic reasoning. Interdependencies among variables in a network are accurately captured and speculative if-then type computation can be performed.

- Computational tractability: Belief networks are computationally tractable for most practical applications. This efficiency stems principally from the exploitation of conditional independence relationships over the domain. We have presented an efficient single-pass evidence propagation algorithm for networks without loops.

- Evidence handling: Evidence can be posted to any node in a belief network. This means that subjective evidence can be posted at an intermediate node representing an abstract concept.

A major disadvantage of belief network technology is the high level of effort required to build network models. Although it is relatively easy to build a belief network structure with the help of subject matter experts, the model will require a significant amount of probability data as the number of nodes and links in the structure increase. The size of a CPT corresponding to a node with multiple parents can potentially be huge. For example, the number of independent entries in the CPT of a binary node (a node with two states) with 8 binary parent variables is 128.

Belief networks are also poor at handling continuous variables. Current software handles continuous variables in a very restrictive manner (for example, they must be Gaussian and can only be children). Lener et al. (2001) developed an inference algorithm for static hybrid belief networks, which are Conditional Linear Gaussian models, where the conditional distribution of the continuous variables assigned to the discrete variables is a multivariate Gaussian. Cob and Shenoy (2004) developed an inference algorithm in hybrid belief networks using Mixtures of Truncated Potentials. But these techniques are yet to be incorporated in commercial software.

4.2 Advantages and Disadvantages of Dempster-Shafer Theory

- Its ability to represent ignorance in a direct and straightforward fashion.
- Its consistency with classical probability theory.

- Its manageable computational complexity.
- Represents the actual state of belief more precisely
- Distinguishes randomness from missing information
- Prior probabilities not required.

Dis:

- Lack of assessment strategies: There is a necessity to assign precise numbers in Dempster-Shafer theory's applications to each subset $A \subseteq \Theta$ by the basic assignment m. Although, the precise degrees of the desired measures may exist, but it is perhaps too difficult to determine them with the necessary precision.

- Instability: Underlying beliefs may be unstable. Estimated beliefs may be influenced by the condi- tions of its estimation.

- Ambiguity: Ambiguous or imprecise judgement could not be expressed by the evidence measures.

- The main problem of the Dempster-Shafer theoryin its original formulation is that its computational complexity grows exponentially with the number of hypotheses.

- mathematically complex
- Has to be calculated over all possible sets of states

- A small modification of the evidence assignments may lead to a completely different conclusion.

- Can lead to misleading and counter-intuitive results.

4.3 Advantages and Disadvantages of Fuzzy Logic

- Easy to design
- Relatively intuitive rules
- Relatively robust controllers

Dis:

- Verification and validation of a fuzzy knowledge-based is typically expensive.

- Determining the exact fuzzy rules and membership functions is a hard task (it is difficult to determine or predict the required number of membership functions).

- Stability is an important concern for fuzzy systems.
- Longer inference chains can be problematic
- The order of inference steps matters

- After inference it can be difficult to exactly interpret the membership value

5 Applications of Bayesian Networks

6 Conclusion

References

[1] Malcolm Beynon, Bruce Curry, and Peter Morgan. The dempstershafer theory of evidence: an alternative approach to multicriteria decision modelling. *Omega, The International Journal of Management Science*, 28(1):37–50, 2000.

[2] Subrata Das. *Foundations Of Decision-Making Agents: Logic, Probability and Modality*. World Scientific Publishing Co., 2008.

[3] Arthur P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30(B):205–247, 1968.

[4] Francisco J. Diez and Marek J. Druzdzel. Reasoning under uncertainty. In L. Nadel, editor, *Encyclopedia of Cognitive Science*, pages 880–886. London: Nature Publishing Group, 2003.

[5] Rakowsky Uwe Kay. Fundamentals of the dempster-shafer theory and its applications to system safety and reliability modelling. *Reliability : Theory & Applications*, 3(4):173–185, 2007.

[6] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 2003.

[7] Robert J. Schalkoff. *Intelligent Systems: Principles, Paradigms and Pragmatics*. Jones Bartlett Learning, 2011.

[8] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[9] Rajendra P. Srivastava. An introduction to evidential reasoning for decision making under uncertainty: Bayesian and belief functions perspectives. *International Journal of Accounting Information Systems*, 12(2):126–135, 2011.

[10] Steven Tanimoto. *The elements of artificial intelligence: an introduction using LISP*. Computer Science Press, 1987.

[11] Ronald R. Yager. On the dempstershafer framework and new combination rules. *Information Science*, 41(2):93–137, 1987.

[12] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[13] Hans-Jürgen Zimmermann. *Fuzzy Set Theory and Its Applications*. Springer Science Business Media, 2001.