Uncertainty in Artificial Intelligence

PH.D. COMPREHENSIVE EXAM

Mohammad Shayganfar - mshayganfar@wpi.edu

May, 26 2015

1 Introduction to Uncertainty in AI

Problems with a large number of variables require maintaining large joint distributions to compute posterior probabilities based on evidence.

2 Theories of Uncertainties

2.1 Bayesian Belief Networks

A *Bayesian Belief Network* is a directed acyclic graph consisting of nodes and edges which provides a graphical model for reasoning under uncertainty. Each node in the network represents a random variable from the domain. The state of each node is called *belief* which based on the prior evidence reflects the posterior probability distribution of the other values associted with that node. Each node also has an associated *Conditional Probability Table* (CPT) which represents the coniditonal probability of the variable given the value of its parents in the graph. Each individual edge between two variables represents the relation or conditional dependence between those two variables. Also, the explicit directions represented by arrows as directinoal edges are the notion of the causality in the network (see Figure 1). They are always drawn from cause nodes to effect nodes, indicating dependencies between variables [3]. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node.

Constructing a belief network can be divided into two different subtasks: a) specifying the causal structure among the existing variables in the netwrok, and b) specifying the prior and conditional probabilies for these variables.

2.1.1 Network's Structure

The structure, or topology, of the network captures qualitative relationships between variables (see Figure 1). The first step in building the Bayesian

network's structure is to determine what are the nodes/variables to represent in the structure, and what are their possible values? For instance, nodes with discrete values can have boolean (to represent that a proposition is true or false), ordered (e.g., enumeration), and integral values (e.g., height of a person). Then, one should determine the exisiting causality between nodes, i.e., to determine which node (parent) influences the other (child) and conncet them through directed edges [7].

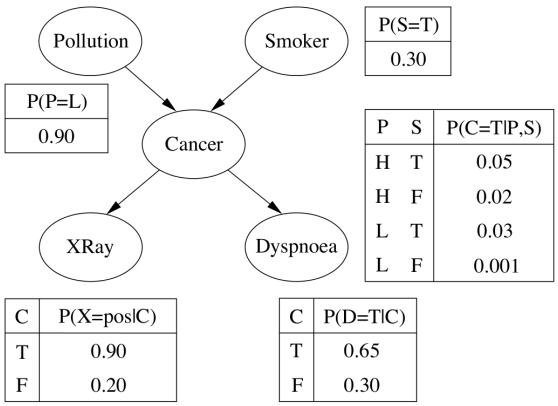

Figure 1: A Belief Network for a lung cancer problem [7].

2.1.2 Conditional Probability Table

As we mentioned earlier, after specifying the structure of a Bayesian Network, the next step is to quantify the relationships between connected nodes by specifying the conditional probability distribution for each node. These conditional probability distributions appear as the *Conditional Probability Tables* (CPT) if we consider discerete variables in the structure. To calculate values in CPTs, for each node, we need to think about all possible combinations of values of parent nodes. Each row in a CPT will contain the value of a conditional probability of a node for each case of the possible combination of values for the parent node. Clearly, a node has many parents or a node with the parents taking a large number of values, can cause creating very large CPTs. The size of the CPT is exponential in the number of parents. For instance, if the nodes of a network are boolean, a variable with n parents requires a CPT with 2^{n+1} probabilities. The probabilities in a CPT are typically acquired from experts of the subject, but they can also be learned automatically using machine learning approaches. Figure 1

shows variables, their relations, and associated CPTs for diagnosis of a lung cancer problem taken from [7].

2.1.3 Markov Property

In Bayesian Networks, each variable is independent of its non-descendants given its parent variables. Therefore, there are no direct dependencies in the system being modeled other than those already explicitly shown via edges. Meaning, there is no hidden connection between variables. This is called *Markov property* in a Bayesian Network. If Bayesian Networks do not adhere to Markov property, there will be redundant edges that connect independent variables together. Consequently, the network will not represent a minimal model.

2.1.4 Joint Probability Distribution

In many applications of probability, there are more than one random variable to be measured over the same sample space (e.g., existence of multiple causes for a lung cancer). A Bayesian Network provides a complete description of the domain. Once we identify random variables and their probabilistic relationships, the values in a joint probability distribution can then be obtained from the probabilities relating the random variables. Therefore, all the entries in the full joint probability distribution can be calculated from the information in the network. There is also a fundamental assumption that in the underlying structure of the problem being modeled by a Bayesian Network, not every single node is connected to every other one [7]. Therefore, if there is such a problem structure, then Bayesian Network can provide a compact representation of a model for that problem. In the following formula $P(x_1, x_2, \ldots, x_n)$ is an abbreviation for the conjuction of $\boldsymbol{n}$ assignments to each variable. Hence, the following formula gives the value of each variable:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

where $parents(X_i)$ denotes the specific values of the variables in $Parents(X_i)$. As we see, given Markov property, the product of only the appropriate elements (parent nodes) of the CPTs in the network represents the value of each individual entry in the joint probability distribution. The following provides an example based on the network provided in Figure 1.

$$P(X = pos \wedge D = true \wedge C = false \wedge P = high \wedge S = true)$$

$$= P(X = pos|D = true, C = false, P = high, S = true)$$
$$\times P(D = true|C = false, P = high, S = true)$$
$$\times P(C = false|P = high, S = true) \times P(P = high|S = true) \times P(S = true)$$

$$= P(X = pos|C = false) \times P(D = true|C = false) \times P(C = false|P = high, S = true) \times P(P = high) \times P(S = true)$$

2.1.5 Reasoning in Bayesian Networks

Reasoning in Bayesian Networks is the process of updating beliefs in face of the evidences. In other words, it is the process of efficiently deducing the belief distribution over a particular subset of random variables given that we know the states of some other variables in the network. Bayesian Networks can be conditioned upon any subset of their variables, supporting any direction of reasoning. Figure 2 shows four different types of reasoning using the network shown in Figure 1. These four types are reasoning are [7]:

Diagnostic reasoning – This is the reasoning from symptoms (effects) to cause. For instance, a doctor updates her belief about a patient's cancer when she checks the X-ray results.

Predictive reasoning – This is the reasoning based on new information about the causes to new beliefs about the corresponding effects. For instance, if the patient tells his doctor the information about the polluted area he lives, the doctor's belief about the patient having cancer increases, even without assessing patient's symptoms.

Intercausal reasoning – This is the reasoning about the mutual causes of a common effect. For instance, suppose that there are two different causes for lung cancer, smoking and pollution (see Figure 1). Initially these two causes are independent of each other; i.e., the patient smoking or not, does not change the parobability of the patient being subject to pollution. However, as soon as the patient is diagnosed with cancer, the probability of smoking or living in a polluted area increases. Now, if the doctor discovers that her patient is a smoker, then the probaility of him living in polluted area dicreases. Therefore, the presence of one explanatory cause for the cancer lowers the probability of the alternative cause, even though that they both were independent causes. In other words, the first explanatory cause *explains away* the alternative one.

Combined reasoning – Sometimes the reaosning does not fit to one of the explained types. Thus, any of these reasoning types can be combined to

solve a problem.

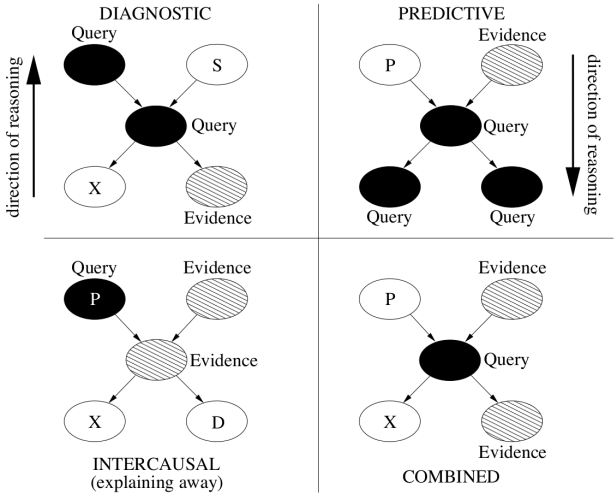

Figure 2: Types of Reasoning [7].

2.1.6 Conditional Independence

Bayesian networks which satisfy the Markov property (see Section 2.1.3) explicitly express conditional independencies in probability distributions [7]. Therefore, since a Bayesian Network is based on joint probability distribution of a set of random variables, the knowledge about conditional independence of these random variables is important for understanding of reasoning based on conditinal probailities.

Two random variables A and B are *conditionally independent* given another variable C, if $p(A, B|C) = p(A|C).p(B|C)$, therefore:

$$p(A|B, C) = \frac{p(A,B|C)}{p(B|C)} = \frac{p(A|C).p(B|C)}{p(B|C)} = p(A|C)$$

And similarly, $p(B|A, C) = p(B|C)$. Figure 3(a) shows a *causal chain* between A and B and C. For instance, the fact that being a smoker can cause lung cancer which causes shortness of breath, in our example. This kind of causal chains can cause a conditional independence which can be described as: $P(C|A, B) = P(C|B)$. This means that if one already knows that C has occurred, knowing that A occurred doesnt make a difference to one's beliefs about C. Figure 3(b) shows that both variables A and C have a *common cause* called B. For instance, based on our example, lung cancer is a

common cause for a positive x-ray and dyspnoea in patient. This kind of common causes can also cause a conditional independence which , again, can be described as: $P(C|A,B) = P(C|B)$. This means that if one already knows about B, then an additional information that A provides, will not give more information about the chances of C. Figure 3(c) shows that one variable has two causes. *Common effect* produces the opposite conditional independence to that of common causes and causal chains. This means that parents are independent until the common effect provides new information. This kind of common effects can cause a conditional dependence which can be described as: $P(A|B,C) \neq P(A|B)$. In other words, if one knows about B (the effect), then finds out that for example A (one of two causes) is absent, this increases the probability of C (alternative cause).

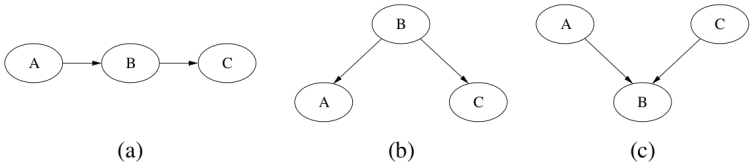

Figure 3: (a) causal chain, (b) common cause, and (c) common effect [7].

2.1.7 d-Separation

The concepts of conditional dependencies and independencies, discussed above, can apply not only between pairs of nodes, but also between sets of nodes. In general, it is possible to determine wheter two sets of nodes X and Y are independent, if there is a set of evidence nodes E, given the Markov property. If the two sets of nodes X and Y are *d-separated* (directional-dependent separation) by evidence set of nodes E, then (given the Markov property) the two sets of nodes X and Y are conditionally independent given E. d-separation is a topological criterion for Bayesian Networks [8]. Figure 4 shows how the evidence set of nodes E in three different conditions is blocking the two sets of nodes X and Y. In a graph, a path is blocked given a set of nodes E, if there is a node Z on the path for which at least one of the three conditions discussed in Section 2.1.6 holds.

Analogous to our example, based on this definition the pollution and smoking variables are d-separated from x-ray and dyspnoea (blocking condition 1), x-ray is d-separated from dyspnoea (blocking condition 2), and if cancer and x-ray or dyspnoea are not observed, then smoking variable would

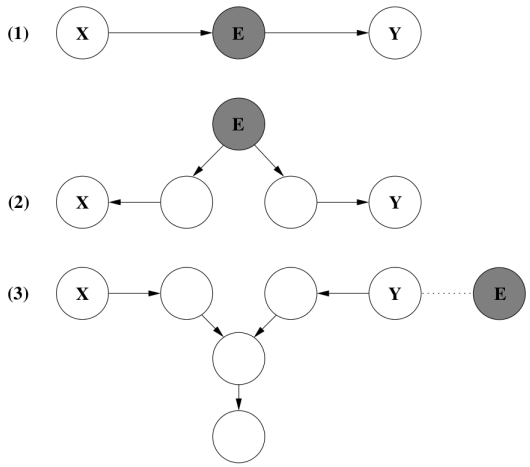

Figure 4: Three types of situations in which the path from X to Y can be blocked, given evidence E. In each case, X and Y are d-separated by E [7].

have been d-separated from pollution (blocking condition 3).

2.2 Dempster-Shafer Theory

In [4], Dempster proposed a probabilistic framework based on lower and upper bounds on probabilities. In [10], 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) [5]. 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 [3]. 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 [8], and it does not require the assumption that $\mathit{Belief}(\mathrm{A}) + \mathit{Belief}(\neg\mathrm{A}) = 1$. Dempster-Shafer theory deals with the possible values of an unknown variable, just as deos the theory of probability [12].

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 hy-

potheses. 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 [6]. Moreover, it is required that all hypotheses are unique, not overlapping and mutually exclusive.

2.2.1 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.

2.2.2 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:

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

2.2.3 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:

$Belief(A) = 1$ - $Plausible(\neg A)$ and $Plausible(A) = 1$ - $Belief(\neg A)$,

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:

$Belief(A) + Belief(\neg A) \leq 1$, 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 $Belief(A) = Plausible(A)$ [2].

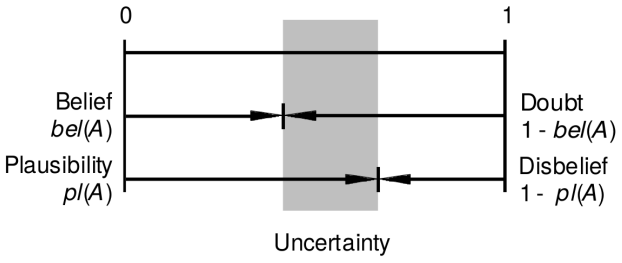

Figure 5: Measures of belief and plausibility. The uncertainty interval is shaded gray. [6].

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 $Belief(A) \leqslant Plausible(A)$ [13], and it is also called as *belief interval*. Figure 5 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 5, 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.

2.2.4 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 [11].

2.3 Fuzzy Logic Theory

Fuzzy Logic, introduced by Zadeh in 1965 [15], provides a mathematical framework to capture uncertainty. There are different kinds of uncertainties in the real world. For instance, randomness is one kind, which is typically modeled using probability theory. Fuzziness manipulates uncertainty by dealing with the boundaries of a set that are not clearly defined. 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 [8]. Logics as bases for reasoning can be distinguished essentially by three items: truth values, operators, and reasoning procedures (e.g., tautologies) [16]. 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 and including "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 6: Fuzzy Logic algorithm.

Figure 6 shows the Fuzzy Logic algorithm. It begins with initialization of linguistic varibales (see Section 2.3.4) and constructing appropriate membership functions (see Section 2.3.3) and rule-base of the fuzzy system (see Section 2.3.6). The constructed membership functions transform the input data to fuzzy values (see Section 2.3.7). 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 transformed to a non-fuzzy (crisp) value (see Section 2.3.9).

2.3.1 Probability vs Possibility

The theory of possibility is analogous and yet conceptually different from the theory of probability. Based on the Fuzzy Logic theory of Zadeh [15] there is a difference between possibility of an event happening and the probability of that. The follwing example by Zimmermann in [16] shows the differnce. Consider the statement "Hans ate X eggs for breakfast", where $X \in U = \{1, 2, \ldots, 8\}$. We may associate a probability p by observing *Hans* eating breakfast for 100 days,

$$\begin{array}{lcl} U & = & [\ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8\ ] \\ p & = & [\ .1 \quad .8 \quad .1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ] \end{array}$$

A fuzzy set expressing the degree to which *Hans* can eat X eggs in breakfast may be the following possibilty distribution π,

$$\begin{array}{lcl} U & = & [\ 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8\ ] \\ \pi & = & [\ 1 \quad 1 \quad 1 \quad 1 \quad .8 \quad .6 \quad .4 \quad .2\ ] \end{array}$$

where the possiibility of $X = 3$ is 1, while the probability of *Hans* eating 3 eggs for breakfast is only 0.1. Therefore, as the example shows, a possible event does not necessarily imply that it is probable too. However, if the event is probable it must also be possible [16].

2.3.2 Fuzzy Sets

Fuzzy sets are a further development of the mathematical concept of a conventional or crisp sets. A fuzzy set is a class of objects with continuum of degrees of membership [15]. Following Zadeh [15] many sets have more than an either-or criterion for membership. For insatnce, the set young people which can contain people at diffrent ages. A fuzzy set A is defined by a membership function μ_A from the universe of discourse $\mathcal{X}$ to the closed unit interval [0,1]. We interpret $\mu_A(x)$ as the degree of membership of x in A (see Section 2.3.3). Zadeh proposed this degree of membership, such that the transition from membership to non-membership is gradual rather than abrupt. Therefore, the degree of membership for all its members describes a fuzzy set.

2.3.3 Membership Functions

Membership functions are the crucial part of the Fuzzy Logic theory. In fact, the difference between crisp (i.e., classical) and fuzzy sets is established by introducing membership functions. 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 2.3.7 and 2.3.9) of a Fuzzy Logic system. A membership function is used to quantify a linguistic term (see Section 2.3.4). 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 2.3.7 and 2.3.9) [9].

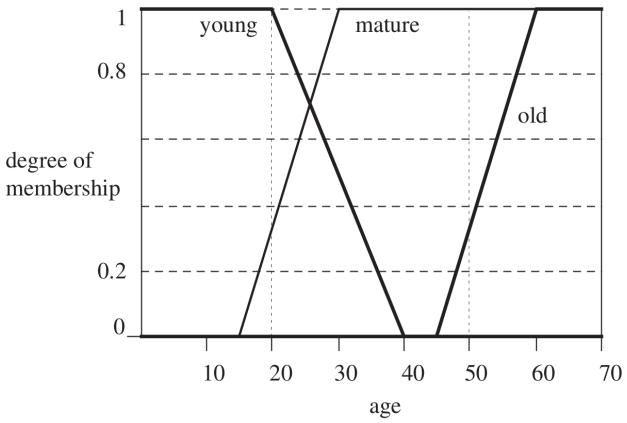

Figure 7: Membership functions for the concepts young, mature and old.

Let $\mathcal{X}$ be a crisp 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 (grade)* of x in A. The degree of membership is expressed by a real number in the interval [0, 1]. The degree of membership is a precise, but subjective measure that depends on the context.

Figure 7 shows membership functions for three linguistic terms of age variable (see also Section 2.3.4). 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 instance, 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 matures 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.

Membership functions can have different shapes and their shape can be determined arbitrarilty based on experience or sometimes by running statistical studies on data. They can be sigmoidal, heyperbolic, Gaussian or any other shape. The followings are some of the important properties of fuzzy sets and membership functions:

Height: The height of a fuzzy set A, denoted by $h(A)$, corresponds to the upper bound of the membership function. In other words, it is the largest membership degree obtained by any element in that set:

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

Support: The support of a fuzzy set 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. In other words, support is a set of all elements of $\mathcal{X}$ that have non-zero membership degrees in A.

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

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

core: The core of a fuzzy set A is the crisp set that contains all the elements of $\mathcal{X}$ that have the membership degrees of **one** in A.

2.3.4 Linguistic Variables

The concept of membership function discussed in Section 2.3.3 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. In fact, just like an algebraic variable takes numbers as values, a linguistic variable takes words or sentences as values [16]. 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, $Age(a) = \{\text{very-old, old, mature, young, very-young}\}$. The members of this decomposition set are called *linguistic terms* which can cover a portion of the overall values of people's age. In other words, the values that a linguistic variable can take is called its linguistic terms.

2.3.5 Fuzzy Operators

Fuzzy operators are used in order to manipulate fuzzy sets, and for being able to evaluate the constructed fuzzy rules (see Section 2.3.6), and ultimately to be able to combine the results of the individual rules. The operations on

fuzzy sets are different than the operations on classical sets. The definitions of operators on fuzzy sets are not the same and can be arbitrarily chosen. Zadeh in [15] defined the intersection (logical and), union (exclusive or), and complement (negation) operations for fuzzy sets as generalization of crisp sets and of crisp statements. Here are the operators for the complement (NOT), the intersection (AND) and union (OR) that are most commonly used:

The memebership function of the **Intersection** of two fuzzy sets A and B:

$$\mu_{A\cap B}(X) = \text{Min}(\mu_A(X), \mu_B(X)) \qquad \forall x \in X$$

The memebership function of the **union** of two fuzzy sets A and B:

$$\mu_{A\cup B}(X) = \text{Max}(\mu_A(X), \mu_B(X)) \qquad \forall x \in X$$

The memebership function of the **complement** of a fuzzy set A:

$$\mu_A(X) = 1 - \mu_A(X) \qquad \forall x \in X$$

These definitions were later extended by other reasearchers too, e.g., [14].

2.3.6 Fuzzy Rules

In a Fuzzy Logic system, a rule-base is constructed to determine and control the output variable. Fuzzy rules are simply comprised of IF-THEN rules which include two parts of condition and conclusion. A fuzzy rule is encoded in a statement in the following form:

IF (a statement of conditions is satisfied)
THEN (a set of consequences can be inferred).

The followings are two examples of fuzzy rules based on the age example depicted in Figure 7:

IF (age is *young*) **THEN** (run *command "talk"*)
IF (age is *old* **OR** *mature*) **THEN** (run *command "listen"*)

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

2.3.7 Fuzzification

For mapping the crisp values to fuzzy ones, we need to evaluate their membership degree using membership functions (see Seciton 2.3.3). This process is called fuzzification and it helps us to get one fuzzy value for each crisp input. Therefore, the fuzzification process is mainly used to transform a crisp set to a fuzzy set.

2.3.8 Reasoning in Fuzzy Logic

In order to draw conclusions from a rule-base, we need a mechanism that can produce an output from a collection of IF-THEN rules. Meaning, after evaluating the result of each rule with respect to the given input value(s), the results of the rules should be combined to obtain a final result. This process in Fuzzy Logic systems is called reasoning. Fuzzy reasoning includes two distinct parts: evaluating the IF part of the rule and applying the result to the consequent (the THEN part of the rule). In fuzzy systems the evaluation is slightly different than the classical rule-based systems. In fuzzy systems the IF part of the rule is a fuzzy statement which means all the rules fire at some extent. If the IF part of the rule is true in some degree of membership, then the consequent is also true in some degree. It is noteworthy to mention that the results of individual rules can be combined in different ways. There are different types of accumulation methods that can be used to combine the results in individual rules.

2.3.9 Defuzzification

After the reasoning step, the Fuzzy Logic system provides the overall result as a fuzzy value. Then, to obtain a final crisp output value, this fuzzy result should be defuzzified which is the purpose of the defuzzifier component of a Fuzzy Logic system. Defuzzification is performed according to the membership function of the output variable. Figure 8 shows the defuzzification step in a Fuzzy Logic system.

There are different algorithms for defuzzification step. One of the most widely used algorithms is called *centeroid*, or *Center of Area*, or *Center of Gravity* (COG). This method computes the center of area of the region under the curve dened by a fuzzy set. In this method, the defuzzified values tend to move smoothly in reaction to small changes, and it is relatively easy to compute the value. In the following formula, A is a fuzzy set and z_{COG} is the final single crisp output which in this case is obtained by COG method:

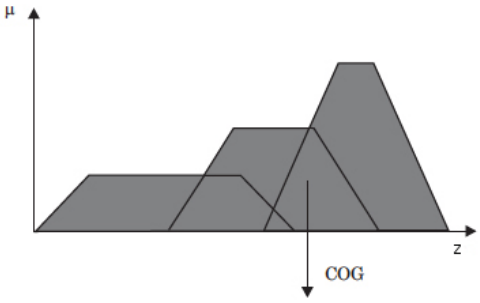

Figure 8: Defuzzification using Center of Gravity (COG) method [1].

$$z_{COG} = \frac{\sum_{i=1}^{n} \mu_A(z_i).z_j}{\sum_{i=1}^{n} \mu_A(z_i)}$$

In summary, Figure 9 (see also the fuzzy algorithm in Figure 6) generally 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 2.3.4), fuzzy linguistic terms and membership functions (see Section 2.3.3). This step is known as fuzzification (see Section 2.3.7). Afterwards, an inference is made based on a set of rules (see Section 2.3.8). Lastly, the resulting fuzzy output is mapped to a crisp output using the membership functions, in the defuzzification step (see Section 2.3.9).

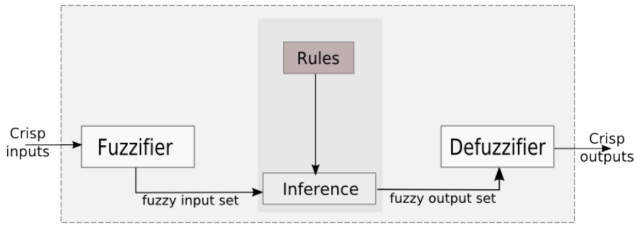

Figure 9: A Fuzzy Logic system.

2.4 Other approaches

3 Strengths and Weaknesses

In general, there is an increasing trend of computational complexity from using Fuzzy Logic to probabilistic approaches and Dempster-Shafer theory.

However, the representational power and precision increases in the same order and direction.

- 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 [8].

- 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 [8].

3.1 Advantages and Disadvantages of Belief Networks

- Transparent representation of causal relationships between system variables.

- BNs can facilitate learning about causal relationships between variables.

- The graphical nature of a BN clearly displays the links between different system components. Knowledge of the structure of a system can reveal the dependence and independence of variables and suggest a direction of causation. Bayesian Networks visually represent all the relationships between the variables in the system with connecting arcs.

- It is relatively easy to recognize the dependence and independence between various nodes.

- Bayesian networks can handle situations where the data set is incomplete since the model accounts for dependencies between all variables.

- A BN can be readily updated when new knowledge becomes available.

- Suitable for small and incomplete data sets.

- New evidence can be incorporated.

A major disadvantage of belief networks 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).

- Efficient reasoning

- It can be computationally intensive if the conditional independencies are not considered properly among the variables.

- Constructing a Bayesian Network involves identifying causal dependencies between different events, which might not be an easy task.

Modeling difficulites:

1. specifying joint probabilities is tedious.

2. Beliefs (held by human experts) are not really enforced.

Like any other computational formalism, belief network technology offers certain advantages and disadvantages. Advantages of belief networks include [3]:

- 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, 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 can be accurately captured.

- 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.

- 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.

- Difficult to design the network (required variables and their relations).

- Difficult to obtain the CPTs.

- It is challenging to get experts' knowledge in the form probability to build the network.

- No feedback loops. The BN structure has an acyclic nature which prevents typical feedback loops in design of its applications.

- All branches must be calculated in order to calculate the probability of any one branch.

3.2 Advantages and Disadvantages of Dempster-Shafer Theory

- Effectively representing the uncertainty on the correctness of the different hypotheses by means of two limiting values, belief and plausibility, overtaking the difficulties that experts encounter in characterizing their uncertainty using single probability values or probability density functions.

- Addresses the concept of possibility.

- 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:

- This allows one to state that certain prior and conditional probabilities cannot be assessed.

- 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 conditions of its estimation.

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

+ The main problem of the Dempster-Shafer theory in 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.

3.3 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

4 Applications of Bayesian Networks

5 Conclusion

References

[1] Antonio Claudio Marques Afonso, Andre Maciel Netto, and Wagner Estaquio de Vasconcelos. Fuzzy logic applied to the modeling of water dynamics in an oxisol in northeastern brazil. *Revista Brasileira de Ciencia do Solo*, 38(2):454–463, 2014.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence*. Taylor & Francis, 2003.

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

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

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

[11] 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.

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

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

[14] Ronald R. Yager, Carol L. Walker, and Elbert A. Walker. Generalizing leximin to t-norms and t-conorms: the lexit and lexis orderings. *Fuzzy Sets and Systems*, 151(2):327–340, 2005.

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

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