What Does a Belief Function Believe In ?

Andrzej Matuszewski, Mieczysław A. Kłopotek

Institute of Computer Science, Polish Academy of Sciences

e-mail: klopotek@ipipan.waw.pl

June 9, 2017

ABSTRACT

The conditioning in the Dempster-Shafer Theory of Evidence has been defined (by Shafer [15] as combination of a belief function and of an "event" via Dempster rule. On the other hand Shafer [15] gives a "probabilistic" interpretation of a belief function (hence indirectly its derivation from a sample). Given the fact that conditional probability distribution of a sample-derived probability distribution is a probability distribution derived from a subsample (selected on the grounds of a conditioning event), the paper investigates the empirical nature of the Dempster- rule of combination. It is demonstrated that the so-called "conditional" belief function is not a belief function given an event but rather a belief function given manipulation of original empirical data.

Given this, an interpretation of belief function different from that of Shafer is proposed. Algorithms for construction of belief networks from data are derived for this interpretation.

1 Introduction

The Dempster-Shafer (DS) Theory (DST) or the Theory of Evidence [14], [3] is considered by many researchers as an appropriate tool to represent various aspects of human dealing with uncertain knowledge, especially for representation of partial ignorance [17], though this view has been challenged by various authors (compare the presentations and discussions in International Journal of Approximate Reasoning (IJAR), special issues in Vol. 1990:4 No. 5/6 and Vol. 1992:6 No.3; see also [6], [4].

This paper is intended to shed some light onto the dispute over adequacy of the DST from the technical point of view. The authors of this paper have been engaged in a project having as its goal the implementation of an expert system dealing with uncertainty via DST methodology mixed with the Bayesian approach [7]. Knowledge is represented by a belief network to enable application of the reasoning system based on the work of Shenoy and Shafer [16] (Shenoy-Shafer's axiomatic framework encompasses both DST and Bayesian-like reasoning scheme). It has been an ultimate goal of the developers of that expert system to support also knowledge acquisition from data. Literature provides with many methods of recovery of Bayesian belief network from data (and in certain cases from additional appropriate hints of an expert), if the belief network should be a tree [1], a polytree [13], or a general-type (usually sparse) network [2], [18]

As "generalized probability" is a term frequently used to characterize the DS belief function, it seems plausible to try to generalize Bayesian methods onto recovery of Dempster-Shafer belief networks from data. However, as the above-mentioned discussion in IJAR demonstrates, the relationship between empirical frequencies and DS belief functions seems to be far from being clear.

We agree with Smets [17] that fundamental deviation of DST from any probabilistic measure of uncertainty lies in the DS rule of combination ($\oplus$) which serves as a way of conditioning the overall DS distribution on some event (see [15]).

In this paper we assume that DST notions like basic probability assignment or mass function m, belief function Bel, pseudo-mass and pseudo-belief functions, combination $\oplus$, marginalization $\downarrow$ and empty extension $\uparrow$ operations are well understood [14], [16].

2 Nature of Conditioning in DST

Let us consider for a moment how "empirical" conditioning may be viewed in the probability theory. Let a probability distribution be defined as relative frequency over a (large) population. Let us want to condition on an event, say $\{\omega|\alpha(\omega)\}$ where α is a predicate - a logical expression in variables describing the population. Then we select all the objects ω of the population which match the predicate $\alpha(\omega)$ and the conditional distribution $P(.|\alpha)$ will be the relative frequencies within this subpopulation.

Let us try to do the same with DS Bel's. Following Shafer [15] we may be tempted to interpret a valuation of an object ω of a population with a set A as a statement that our variable of interest takes for this object one of the values mentioned in A, but we do not know which one of them (and are not ready even to select any proper subset of A). A will be treated as our most specific commitment to the value of the variable. Under these circumstances, the basic probability assignment function m may be understood as the probability distribution (read: relative frequencies) of such commitments within our population. This is, in fact, the way as the "generalized probability" in [6], or families of probability distributions in [10] may be understood. Now let us go over to conditioning on an event $\{\omega|\alpha(\omega)\}$ (after [15]). It is the beauty of the DST that there exists always a set B that corresponds exactly to such an event. If Bel_B is the simple support function capturing the evidence of the set B (that is $m_B(B) = 1$, $m_B(A) = 0$ for any $A \neq B$, [15]) then by definition $Bel(.|B) = Bel \oplus Bel_B$ is the belief of Bel conditioned on the event B [15]. But how does this definition "run" on a population of objects ? It has been demonstrated [8] that we can view it the following way: We take the predicate α

and check the population object by object. Some of them deny the predicate. We reject them as the frequentist model of conditional probability does. Some of them meet the predicate - and we select them for our subpopulation - as conditional probability model does. But there are some objects for which we are actually unable to decide whether they meet the predicate or not (as our commitment is not specific enough). And what do we have to do in this case to meet numerically the DS rule of combination ? We have to accept them ! But (Alas!) this is not enough. We have to change our commitment - we have to make our commitment for a particular object more specific so that it meets the predicate α. So even if our commitment to the value of the attribute for this object may have been correct prior to conditioning (that is the variable of interest took for the object one of the values mentioned in the prior commitment), it may be not correct after the conditioning. That is, after the conditioning we work with a subpopulation with partially incorrect valuations (not corresponding with empirical reality), and we combine evidence further ..... . (Classical!) probability theory does not do things like this - it assumes that measuring sequence has no impact on the value of a variable (Of course, there are several non-classical probability theories which took into account possibility of disagreement between various observations if the sequence of making observations is changed, but obviously most frequentist interpretations of DST didn't consider them).

We feel that this (essentially numerical) argument explains most of apparent contradictions derived from frequentist interpretations of the DST. But the DST will not be helped with if left with the impression of telling us lies about the population (we have to say, plausible lies, because by definition we are unable to check by observation if strengthening of a commitment for an object is in fact correct or not, because if we were able to carry out such an observation then our prior commitment would have been more specific). So instead of saying that the variable takes for the object one of the values in A, we could say that the variable is set-valued and takes for the object all the values in A (this would be a kind of random set interpretation [11]). In this case conditioning could be viewed as rejecting some of the values of the object which are not of interest. Then after conditioning the object would have a commitment corresponding to the values it really takes, though some other values were ignored as not of interest. However, this view would contradict the usage of Bel's to represent material implication $P(\omega) \rightarrow Q(\omega)$ in form of a set $\{(P(\omega), Q(\omega)), (\neg P(\omega), Q(\omega)), (\neg P(\omega), \neg Q(\omega))\}$ because it would lead to the impression that at the same time $P(\omega) \wedge Q(\omega)$ and $\neg P(\omega) \wedge Q(\omega)$ hold which is counterintuitive as $P(\omega) \wedge \neg P(\omega) = FALSE$.

3 An Alternative View of DST

Hence an alternative view of DST is required. One has been developed in [8]. We present it here informally. Instead of saying that the set A expresses that "the variable of interest takes one of the values in A" as well as instead of saying that the set A expresses that "the variable of interest takes all of the values in A" a compromise is proposed: it is assumed that the variable cannot be observed directly, but only via some measurement procedure (with some special properties ensuring consistence), and the set A expresses that "the measurement procedure yielded TRUE when testing if $X = a_i$ (X - the variable of interest) for all the $a_i \in A$ and for no $a_i \notin A$". If we make conditioning, the objects are labeled, and the measurement method takes into account the labeling of objects by refraining from carrying out tests on variable values outside of the label. In this way the following is achieved:

- before and after every conditioning the interpretation of the commitment A is the same for not rejected objects: "the measurement procedure yielded TRUE when testing if $X = a_i$ (X - the variable of interest) for all the $a_i \in A$ and for no $a_i \notin A$".

- the impact of conditioning onto measurement results is taken into account - via labeling
- any logical contradictions resulting from random set interpretation are avoided: we do not say ”X takes the value” but that ”X has been measured to be”, and contradictions resolve in imprecision of measurement method.

The importance of such an interpretation is not to be underestimated: a way is paved towards experimental studies of populations with DS belief distributions.

To demonstrate this a development of a method of a tree/polytree factorization of a joint DS belief distribution for purposes of Shenoy/Shafer uncertainty propagation [16] is briefly outlined.

We define *mk-conditional belief function* $Bel^{X|X_i}(A)$ as any pseudo-belief function solving the equation $Bel = Bel^{\downarrow X_i} \oplus Bel^{X|X_i}$ Notice, that in general this equation has no unique solution, and a solution being a proper belief function does not always exist.

A *DS Belief network* be [9] a pair (D,Bel) where D is a dag (directed acyclic graph) and Bel is a DS belief distribution called the *underlying distribution*. Each node i in D corresponds to a variable X_i in Bel, a set of nodes I corresponds to a set of variables X_I and x_i, x_I denote values drawn from the domain of X_i and from the (cross product) domain of X_I respectively. Each node in the network is regarded as a storage cell for any distribution $Bel^{\downarrow\{X_i\}\cup X_{\pi(i)}|X_{\pi(i)}}$ where $X_{\pi(i)}$ is a set of nodes corresponding to the parent nodes $\pi(i)$ of i. The underlying distribution represented by a DS belief network is computed via:

$$Bel = \bigoplus_{i=1}^{n} Bel^{\downarrow\{X_i\}\cup X_{\pi(i)}|X_{\pi(i)}}$$

Let, after [5] $I(J,K|L)_D$ denote *d-separation* of J from K by L in a directed acyclic graph D, where J,K and L are three disjoint sets of nodes in this dag D. We shall then define [9]

If X_J, X_K, X_L are three disjoint sets of variables of a distribution Bel, then X_J, X_K are said to be *conditionally independent* given X_L (denoted $I(X_J, X_K|X_L)_{Bel}$ iff

$$Bel^{\downarrow X_J \cup X_K \cup X_L|X_L} \oplus Bel^{\downarrow X_L} = Bel^{\downarrow X_J \cup X_L|X_L} \oplus Bel^{\downarrow X_K \cup X_L|X_L} \oplus Bel^{\downarrow X_L}$$

$I(X_J, X_K|X_L)_{Bel}$ is called a *(conditional independence) statement*

THEOREM 1 *[9] Let $Bel_D = \{Bel|$(D,Bel) is a DS belief network$\}$. Then:*

$$I(J, K|L)_D$$

iff

$$I(X_J, X_K|X_L)_{Bel}$$

for all $Bel \in Bel_D$.

Many authors have connected causality with the notion of statistical dependence or non-independence. We parallel here [18] in formulating the following principles, while understanding independence as defined above

Let **V**be a set of random DS variables with a joint DS-belief distribution. We say that variables X,Y ∈**V**are *directly causally dependent* if and only if there is a causal dependency between X,Y (either the value of X influences the value of Y or the value of Y influences the value of X or the value of a third variable not in **V**influences the values of both X and Y) that does not involve any other variable in **V**.

Principle I: For all X,Y in **V**, X and Y are directly causally dependent if and only if for every subset **S**of **V**not containing X or Y, X and Y are not statistically independent conditional on **S**.

We say that *B is directly causally dependent on A* provided that A and B are causally dependent and the direction of causal influence is from A to B.

Principle II: if A and B are directly causally dependent and B and C are directly causally dependent, but A and C are not, then: B is causally dependent on A, and B is causally dependent on C if and only if A and C are statistically dependent conditional on any set of variables containing B and not containing A or C.

Principle III: A directed acyclic graph represents a DS-belief distribution on the variables that are vertices of the graph if and only if
for all vertices X,Y and all sets **S**of vertices in the graph (X,Y $\notin$ **S**), **S**d-separates X and Y if and only if X and Y are independent conditional on **S**.

THEOREM 2 [9] *Let Bel be a DS-belief distribution represented by an acyclic directed graph G according to Principle III. Then G is an orientation (G has the undirected structure) of the undirected graph U that represents Bel according to Principle I.*

THEOREM 3 [9] *Principle III implies Principle II.*

THEOREM 4 [9] *Let* Γ *be the set of directed graphs that represent DS-belief distribution Bel according to Principle III. Then* Γ *is also the set of directed graphs obtained from P by Principles I and II.*

4 Belief Networks from Data under New Interpretation

Based on these purely theoretical considerations it was tried to develop some practical algorithms for recovery of belief network structure from data for some limited classes of belief networks. It was started with the most successful structures of Bayesian networks: the tree and the polytree structures. In these efforts, corresponding Bayesian algorithms were exploited as general frameworks, though details had to be elaborated anew. It is also worth mentioning, that, unlike in probabilistic case, a randomized generation of a belief function possessing given belief network structure is not a trivial task due to the data-changing nature of DS combination.

Let us present briefly these new algorithms:

The algorithm of Chow and Liu [1] for recovery of tree structure of a probability distribution is well known and has been deeply investigated, so we will omit its description in this paper. To accommodate it for the needs of DST one needs to introduce a definition of distance between variables. Regrettably, no such definition having the nice properties of the Chow and Liu exists, so a similar one has been elaborated: Let p be a mass function and x be a pseudo-mass function. Let $f(x;p) = \sum_{A;p(A)>0} p(A) \cdot \ln x(A)$, where the assumption is made that natural logarithm of a non-positive number is minus infinity. The values of f in variable x with parameter p have range:$(-\infty, f(p;p)]$.Let $g(x;p) = \frac{f(x;p)}{f(p;p)}$.The values of g in variable x with parameter p range:$[1, +\infty)$.Let $a(x;p) = e^{1-g(x;p)}$. The values of a in variable x with parameter p range:$[0, 1]$

By the ternary joint distribution of the variables X_1, X_2 with background X_3 we understand the function:

$$m^{\downarrow X_1 \times X_2 [X_3]} =$$

$$= (m^{\downarrow X_1 \times X_3 | X_3} \oplus m^{\downarrow X_2 \times X_3 | X_3} \oplus m^{\downarrow X_3})^{\downarrow X_1 \times X_2}$$

By the distance (for use with Chow/Liu algorithm) $DEP(X_1, X_2)$ we understand the function:

$$DEP0_{DS}(X_1, X_2) = 1 - \max(a(m^{\downarrow X_1} \oplus m^{\downarrow X_2}; m^{\downarrow X_1 \times X_2}), \max_{X_3; X_3 \in V - X_1, X_2}$$

$$a(m^{\downarrow X_1 \times X_2 [X_3]}; m^{\downarrow X_1 \times X_2}))$$

with **V** being the set of all variables.

For randomly generated tree-like DS belief distributions, if we were working directly with these distributions, as expected, the algorithm yielded perfect decomposition into the original tree. For random samples generated from such distributions, the structure was recovered properly for reasonable sample sizes (200 for up to 8 variables). Recovery of the joint distribution was not too perfect, as the space of possible value combination is tremendous and probably quite large sample sizes would be necessary. It is worth mentioning, that even with some departures from truly tree structure a distribution could be obtained which reasonable approximated the original one.

A well known algorithm for recovery of polytree from data for probability distributions is that of Pearl [12], [13], we refrain from describing it here. To accommodate it for usage with DS belief distributions we had to change the dependence criterion of two variables given a third one.

$$Criterion(X_1 \to X_3, X_2 \to X_3) = (1 - a(m^{\downarrow X_1 \times X_2 [X_3]}; m^{\downarrow X_1 \times X_2})) -$$
$$-(1 - a(m^{\downarrow X_1} \oplus m^{\downarrow X_2}; m^{\downarrow X_1 \times X_2}))$$

If the above function *Criterion* is positive, we assume head-to-head meeting of edges X_1, X_3 and X_2, X_3. The rest of the algorithm runs as that of Pearl.

For randomly generated polytree-like DS belief distributions, if we were working directly with these distributions, as expected, the algorithm yielded perfect decomposition into the original polytree. For random samples generated from such distributions, the structure was recovered properly only for very large sample sizes (5000 for 6 variables), with growing sample sizes leading to spurious indications of head-to-head meetings not present in the original distribution. Recovery of the joint distribution was also not too perfect, due to immense size of space of possible value combinations.

Other distance and dependence measures than those mentioned above have been tried but no clear winner could have been decided so far.

Though we are still far away from our goal of developing an efficient algorithm for recovery of general DS belief network structure from empirical data, our efforts demonstrated, that there exists at least one way of connecting the formalism of the Dempster-Shafer Theory with frequencies from empirical data (though this may not be the one the creators of this theory had in mind). At the same time our view of the nature of the DS belief functions was shifted from traditional frequentist view to one with accepting changing valuation of objects while running the conditioning process. This proved helpful when comparing results of reasoning of the inference engine with the empirical distribution given by data.One of the consequences of this changing valuation of data during a reasoning process is the crucial difference between probabilistic and DS belief networks: In probabilistic networks the conditioning of a whole distribution on a set of variables has exactly the same meaning as conditionality contained in a node of a network. That is, if the variable X_n represented by a node n depends on the set of variables $X_{\pi(n)}$ then if we calculate the conditional probability $P(X_n|X_{\pi(n)}$ on a whole network e.g. via Shenoy/Shafer algorithm [16], then the result will be exactly the same as is the valuation attached to the node n of the network. The situation is entirely different in case of DS networks: the Shaferian $Bel(X_n|X_{\pi(n)})$ calculated from the overall network is (and usually must be) in general distinct from the valuation (mk-conditioning) $Bel^{\downarrow\{X_n\}\cup X_{\pi(n)}|X_{\pi(n)}}$ we attach to a node of the network. Clearly, the Shenoy/Shafer uncertainty propagation algorithm [16] is fully unaffected by the lack of identity between these two notions of conditioning, and in fact a node valuation neither in probabilistic nor in DS case is required to have anything to do with any notion of conditionality. But attachment of conditionality to a node of a belief network is important for understanding the contents of a belief network which was invented as a means of representing causal dependencies [18]. Our notion of mk-conditionality $Bel^{\downarrow\{X_n\}\cup X_{\pi(n)}|X_{\pi(n)}}$ gives a node in a DS belief network a local meaning: it can be estimated from data using only variables engaged, that is $\{X_n\}\cup X_{\pi(n)}$. Notably, this does not hold for the general view of belief networks (that is without reference to conditionality) presented by Shenoy and Shafer [16].To verify the validity of valuation of any node of a general form hypertree considered in [16] one may be forced to consider the entire hypertree at once.

References

[1] C.K. Chow, C.N. Liu: Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory*, Vol. IT-14, No.3, (May 1968), 462-467

[2] G.F. Cooper, E. Herskovits: A Bayesian method for the induction of probabilistic networks from data, *Machine Learning* 9, 309-347 (1992).

[3] A.P. Dempster: Upper and lower probabilities induced by a multi-valued mapping, *Ann. Math. Stat.* 38 (1967), 325-339

[4] R. Fagin, J.Y. Halpern: Uncertainty, belief, and probability, Comput. Intell. 7 (1991), 160-173

[5] D. Geiger, T. Verma, J. Pearl: d-Separation: From theorems to algorithms, in *: Uncertainty in Artificial Intelligence 5* (M.Henrion, R.D.Shachter, L.N.Kamal and J.F.Lemmer Eds), Elsevier Science Publishers B.V. (North-Holland), 1990, 139-148.

[6] J.Y. Halpern, R. Fagin: Two views of belief: belief as generalized probability and belief as evidence,*Artificial Intelligence* 54(1992), 275-317

[7] M.A. Kłopotek, M. Michalewicz, S.T. Wierzchoń: *Ekstrakcja wiedzy w sieci bayesowskiej*, [in:] M.Dąbrowski, M.Michalewicz (Eds): *Praktyczne aspekty sztucznej inteligencji*, pp. 88-99, (in Polish)

[8] M.A. Kłopotek: Dempster-Shafer Belief Function - A New Interpretation, submitted

[9] M.A. Kłopotek: Dempsterian-Shaferian Belief Network From Data, submitted

[10] H.E. Kyburg Jr: Bayesian and non-Bayesian evidential updating, *Artificial Intelligence* 31 (1987), 271-293.

[11] H.T. Nguyen: On random sets and belief functions, *J. Math. Anal. Appl.* 65, 539-542, 1978.

[12] J. Pearl: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Influence*, Morgan and Kaufmann, 1988

[13] G. Rebane, J. Pearl: The recovery of causal poly-trees from statistical data, [in] *Uncertainty in Artificial Intelligence* 3, Kanal L.N., Levit T.S., Lemmer J.F. Eds., Elsevier Science Publishers B.V., (North Holland) 1989, 175-182

[14] G. Shafer: *A Mathematical Theory of Evidence* , Princeton University Press, Princeton, 1976

[15] G. Shafer: Belief Functions. Introduction, [in]: G. Shafer, J. Pearl eds: Readings in Uncertain Reasoning, (Morgan Kaufmann Publishers Inc., San Mateo, California, 1990), 473-481, also therein: .G. Shafer, R. Srivastava: The Bayesian and Belief-Function Formalisms. A General Prospective for Auditing, 482-521.

[16] P.P. Shenoy, G. Shafer: Axioms for probability and belief-function propagation, in: Shachter R.D., Levitt T.S., Kanal L.N., Lemmer J.F. (eds): *Uncertainty in Artificial Intelligence 4*, Elsevier Science Publishers B.V. (North Holland), 1990,

[17] Ph. Smets: Resolving misunderstandings about belief functions, *International Journal of Approximate Reasoning* 1992:6:321-344.

[18] Spirtes P., Glymour C., Scheines R.: *Causality from probability*, [w:] G.McKee (Ed.): *Evolving knowledge in natural and artificial intelligence*, London: Pitman, 1990.