A formal approach for the representation and the combination of imperfect data

Jérôme Dantan[1,2], Yann Pollet[2], Salima Taïbi[1]

[1] *Institut Polytechnique LaSalle Beauvais-Esitpa, Agri'terr research unit*
[2] *CNAM Paris, CEDRIC laboratory*
E-mail : *jdantan@esitpa.fr*
E-mail : *staibi@esitpa.fr*
E-mail : *yann.pollet@cnam.fr*

Abstract
Nowadays, the sustainability of human activities is a major worldwide concern. Indeed, the problem is no longer to evaluate only the efficiency of human activities, but also sustainability along many axes that can be of various kinds: economic, social, environmental, etc. Such assessments are a major challenge for today's society.
Because of the exponential development of means of data recording and storage ("big data" buzz word), and on the basis of Volume, Variety and Velocity properties of big data, scientists need to compute large amounts of data and so do not necessarily have time to clean them. In this context, they compute all available data whose types of imperfections are heterogeneous.
Actors in several domains have to cope with such data, especially to assist humans in their decisions by merging them from many data sources (e.g. measurements, sensors, observations) to model behaviours of complex systems. Mathematical approaches to model imperfect data are well known and established in various scientific areas today, such as both probability based and possibility based calculus. Decisions of experts from various fields have to handle rigorous computations and aggregations of both data and their associated uncertainty. We propose a rigorous model to handle uncertainty on the attributes of objects, and a way to rigorously aggregate discrete data, whose imperfections nature are covered either by the classical probability theory (randomness), either by the possibility theory (fuzziness) thanks to the Dempster-Shafer theory.

Keywords: Probability, Possibility, Dempster Shafer theory

1. Introduction

Agricultural systems (e.g. both crop and livestock systems) depend on data that may be uncertain, e.g. either imprecise (e.g. several possible values for discrete data, area allocated to different crops, especially in developing countries where they are not always accurately known, etc.), or random (e.g. crop forecasts and prices, the weather, etc.). However, such sources of uncertainty do not follow the same axioms and may not be taken into account with the same mathematical methods (possibility and probability theory). In addition, in today's context, the goals of farms are not only to be profit-able, but also to preserve ecosystems, to reduce waste, etc. Farmers have to produce better, to both contribute to sustainable development and optimize farms resources and costs, while limiting the impact of their agricultural activities on the environment.

That’s why the field of decision support in agricultural information systems is promising; its potential contributions are going to deeply impact the future. Decision support tools aim to assist the experts of agriculture in their decisions on specific issues such as what crops or how much input should be used, etc. For this, the stakeholders of the agriculture domain have to cope with increasing amounts of information, to solve issues from basic production evaluations to complex sustainability indicators. Data experts have to manage data that are based on both heterogeneous and increasing sources and formats (such as agricultural field samples, production cost and income, soil heavy metal rates, etc), e.g. (1) data from international research programs collected by research / consulting companies such as INRA[1], FAO[2], chambers of agriculture, GIS data, etc. and (2) data from multiple sensors, intelligent soil analyzing augers, GPS coordinates linked with weight measurements of crops at harvest, milking robots, all in real time.

Our approach integrates the knowledge we have about imprecise data in addition to data suitable for classical statistics analysis. Indeed, the first one focuses on the concept of membership and the second one focuses on distributional differences. We propose a uniform formal model able to deal with uncertain data, and to combine them. This article includes the following sections: (2) a state of the art, (3) the presentation of our approach, (4) an example that illustrates our approach and (5) our conclusions and a brief opening to our future work.

2. State of the art

Nowadays the scientific problems require reflect uncertainty characteristics, including incompleteness and inaccuracy of information. In this section, we present the mathematical modeling approaches and handling of imperfect information relating to possibility theory and probability theory. Finally, we introduce a way for mixing heterogeneous quantities.

2.1 Probability theory

For a long time, uncertainty modelling remained addressed by the probability theory, which is the mathematical study of phenomena characterized by randomness and uncertainty. The probability of membership is measured by the accumulation of the most favourable case, i.e. the probability of an event disjunction is equal to the sum of the occurrence probabilities of these events. Probability distributions characterize random phenomena: they either describe probabilities of each event of a discrete random variable, or probabilities that the continuous random variables belong to an arbitrary interval. Statistics are to collect, process, interpret and present sets of empirical data. From probability distributions the natural variability of phenomena can be modelled, e.g. the concentration of soil pollution, crops over several years, etc.). However, this approach is little suitable for total ignorance representation, and objective interpretations of probabilities, assigned to such events remain difficult when handled knowledge are no longer linked to random and repeatable phenomena (Dubois D., Prade H., 1988). As against, it is possible to model uncertainty thanks to the possibility theory.

[1] French National Institute of Agronomic Research

[2] Food and Agriculture Organization of the United Nations

2.2 Possibility theory

The possibility theory (Dubois D., Prade H., 1988), (Zadeh L.A., 1978) removes the strong probability additive constraint and associates the events of Ω to a possibility measure denoted Π and a necessity measure denoted N, that are both applications from Ω to $[0,1]$, respectively satisfying: $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$ and $N(A \cap B) = \min(N(A), N(B))$. This approach also allows representing imprecision using notions of fuzzy sets and distributions of possibilities.

2.2 Dempster-Shafer theory

The Dempster-Shafer theory is usually related to neither a probability model nor a possibility model. But, in some particular cases, the belief or plausibility deduced from the bbm (basic belief masses) may follow either a probability distribution or a possibility distribution, as stated in the following theorems (Gacôgne L., 1997):
Theorem 1: the focal elements are totally ordered by inclusion iff Bel and Pl are respectively a measure of necessity and possibility.
Nota bene: A possibility distribution is not a belief distribution because the sum of Π is generally NOT 1. Possibility distributions need to be converted into bbm distributions in this way: if fuzzy intervals $I_1, I_2, \dots, I_i, \dots, I_n$ are discrete, there is a simple relationship between the masses $m_1, \dots, m_i, \dots, m_n$ and alpha-cuts $\alpha_1 \geq \ \dots \geq \alpha_i \geq \cdots \geq \alpha_n$ (with $\alpha_1 = 1$):

$$m_1 = \alpha_1 - \alpha_2 \; ; \; m_2 = \alpha_2 - \alpha_3 \; ; \dots ; \; m_i = \alpha_i - \alpha_{i+1} \; ; \dots ; \; m_n = \alpha_n$$

Theorem 2: a belief on a finite set is a probability iff the focal elements are singletons.
In this case, P is a probability, i.e. an application from Ω to $[0; 1]$, satisfying $P(\Omega) = 1$ and $P(\cup_i A_i) = \sum_i P(Ai)$ where $\{A_i\}_{i \in \{1,\dots,n\}}$ are a finite family of disjoint events of Ω, and $P(A) = 1 - P(\bar{A})$.

3. Approach

The provided approach provides a formalism for both representing and manipulating rigorously quantities which may have a finite number of possible or probable values with their interdependencies.
Let Ω be a universe, with both a possibility measure π and a probability measure P, each having a finite number of values. These values may belong either to a division ring K (e.g. $\mathbb{R}$) or a semigroup G (i.e. a set with an associative internal composition law).

3.1 Possibilistic and probabilistic bases

Let E be a vector space over K, infinite but countable dimension. B_I are B and B^J two base sets with:

- $B_I = \{X_{I,\,i} \; ; i = 1, \dots\}$ (I fixed), we call bases Possibilists .
- $B^J = \{X^{J,\,j} \; ; j = 1, \dots\}$ (J fixed), which we call probabilistic bases.

We define an internal product on vectors: Z = X.Y, that check the following properties:

- X.X = X (idempotency)
- $X.0 = 0_v$, where 0_v is the null vector (absorbing element)
- $i_1 \neq i_2$ implies $X_{I,\,i1}.\; X_{I,\,i2} = 0_v$

- $j_1 \neq j_2$ implies $X^{I,\, i1}.\ X^{I,\, i2} = 0_v$
- $X_{I1,\, i1}.\ X_{I2,\, i2} \neq 0_v$
- $X_{I1,\, i1}.\ X^{I2,\, i2} \neq 0_v$
- $X^{I1,\, i1}.\ X^{I2,\, i2} \neq 0_v$

We call $D_p(G)$ the set generated by the probability vector defined above. $D_p(G)$ is the finite set of probability values on G defined on Ω.
We call $D_\pi(G)$ the set generated by the possibilistic vectors defined above. $D_\pi(G)$ is the finite set of values Possibilists on G defined on Ω.

3.2 Discrete possibilistic and probabilistic quantities

In (Dantan et al., 2015), we defined the canonical form of a purely possibilistic $D_\pi(G)$ a is the following expression: $a = \sum_{i=1}^{n} a_i/\alpha_i.X_{I,i}$, with a_i are the possible values of a, α_i are the possibilities, associated to each value a_i (one of which at least is equal to 1) and $X_{I,i}$ (I fixed) is the partition of the universe Ω corresponding to values of quantity a.
The canonical form of a purely probabilistic $D_p(G)$ (Dantan et al., 2015) b is the following expression: $b = \sum_{i=1}^{n} b_i/\beta_i.X^{J,i}$, with b_i are the probable values of b, β_i are the probabilities, associated to each value bj (the sum of β_i is equal to 1) and $X^{J,i}$ (J fixed) is the partition of the universe Ω corresponding to values of quantity b.

3.3 Combination of possibility and probability distributions

In this section, we present a particular Dempster-Shafer interpretation of combining probabilistic and possibilistic quantites as defined previously. We define the two internal compositions laws + and * on mixed operands:

$$a + b = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (a_i + b_j)/((\alpha_i.X_{I,i}).(\beta_j.X^{J,j}))$$

$$a * b = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (a_i * b_j)/((\alpha_i\,.X_{I,i}).(\beta_j.X^{J,j}))$$

We define φ, as: || $(\alpha_{1,i}.\ X_{1,i}).\ (\beta_{2,j}\ .X^{2,j})$ || = φ $(\alpha_{1,i},\ \beta_{2,j}).\ X_{1,i}.X^{2,j}$. φ is a given interpretation of "mixed" possible and probable values. For each computation / combination of quantities from E, φ $(\alpha_{1,i}\ ,\ \beta_{2,j})$ values are updated, from $X_{i1}.X_{i2}....X_{in}$ axes on the one hand and $X^{j1}.X^{j2}....X^{jm}$ on the other.

By hypothesis, one can select the interpretation φ $(\alpha_{1,i},\ \beta_{2,j}) = \alpha_{1,i}\ .\ \beta_{2,j}$, which is equivalent to consider the possibility as a bound of a probability measure and redefine on that basis, i.e. overload the independence between a possible value and a probable value.

Remark: there may be another composition of $\alpha_{1,i}$, $\beta_{2,j}$, which would be another interpretation of mixing possible and probable values. Indeed, the cross terms $(\alpha_{1,i}.\ X_{1,i}).\ (\beta_{2,j}\ X^{2,j})$ are assessed at the end of computations.

The couple (Π, P) is associated with a Dempster-Shafer Measure (DSM) defined by the following focal elements (this is actually the orthogonal sum of Π and P considered as DSM):

$$D_{i,j} = a'_i \text{ X } a_j \text{ where } A'_\iota = \bigcup_{\kappa=1,..,\iota} a_{\chi(\kappa)}$$

Where χ is a permutation on {1, 2, …, N} such as sequence $a_{\chi(1)}, \ldots, a_{\chi(N)}$ is decreasing (but not necessary strictly decreasing), with the associated focal masses:

$$m_M (D_{i,j}) = q_i \,.\, p_j$$

Where q_i is defined by:

- $q_N = \pi_{\chi(N)}$
- $q_i = \pi_{\chi(i)} - \pi_{\chi(i-1)}$ with $N > i \geq 1$ (positive or null because the sequence is increasing)

Permutation χ is not necessarily unique (if multiple values are as much possible) so, in fact, we define a family of DSM which immediately lead to the same plausibility values and credibility for any subset. It is therefore possible to consider that the defined DSM is one of those equivalent measures (which is representative of the equivalence class "gives the same values"). This DSM, defined as the Dempster-Shafer measure product of Π and P, is denoted:

$$M = \Pi \,.\, P$$

Intuitively, the definition of M = Π . P corresponds to the hypothesis that knowledge respectively expressed by Π and P are independent. This hypothesis is formulated as follows:

Definition: let Π be a distribution of possibility, and P a distribution of probability. Π and P are called independent iff:

$\forall A \neq \emptyset, \forall B \neq \emptyset, A \cap B \neq \emptyset$ and Plausibility$(A \cap B) = \Pi(A).P(B)$

Indeed, Π (A) defines the upper bound of a probability interval. This is equivalent to saying that it is true for the corresponding focal elements of Π. It is immediate that this assumption leads to compute new weights associated with M as products of Π . P.

4. Use case

Let a and b be respectively a probabilistic and a possibilistic numbers.

- $a = 2\ (probability\ 0.7)\ or\ 3\ (probability\ 0.3)$
- $b = 1\ (possibility\ 1)\ or\ 2\ (possibility\ 0.5)$

a and b are expressed by:

$$a = 2/0.7 + 3/0.3$$

$$b = 1/1 + 2/0.5$$

By applying theorem 1, we may convert the possibilistic distribution b to a bbm distribution, which implies: $m(\{1\}) = 0.5\,; m(\{1{,}2\}) = 0.5$

$$a = 2/0.7.\ X^{I,1} + 3/0.3.\ X^{I,2}$$

$$a = 1/1.\ X^{J,1} + 2/0.5.\ X^{J,2}$$

The general case is expressed as follows:

$$a + b = 3/0.7.\ X^{I,1}.1.\ X^{J,1} + 4/0.7.\ X^{I,2}.0.5.X^{J,1} + 4/0.3.\ X^{I,1}.1.\ X^{J,2} + 5/0.3.\ X^{I,2}.0.5.\ X^{J,2}$$

The Dempster-Shafer particular interpretation is expressed as follows:

$a + b = 3/ \| 0.7.\ X^{I,1} .1.\ X^{J,1} \| + 4/ \| 0.3.\ X^{I,1} .1.\ X^{J,2} \| + 4/ \| 0.7.\ X^{I,2} .0.5.X^{J,1} \| + 5/ \| 0.3.\ X^{I,2} .0.5.\ X^{J,2} \|$

$a + b = 3/\varphi\ (07, 1).\ X^{I,1} .\ X^{J,1} + 4/\varphi\ (0.3, 1)\ X^{I,1} .\ X^{J,2} + 4/\varphi\ (0.7, 0.5).\ X^{I,2} .X^{J,1} + 5/\varphi\ (0.3, 0.5) .\ X^{I,2} .\ X^{J,2}$

$a + b = 3/0.7.\ X^{I,1} .\ X^{J,1} + 4/0.3.\ X^{I,1} .\ X^{J,2} + 4/0.35.\ X^{I,2} .X^{J,1} + 5/0.15.\ X^{I,2} .\ X^{J,2}$

$a + b =$ may have three possible values: 3, 4 or 5 with their respective plausibilities: 0.7, 0.65 (0.3+0.35) et 0.15.

4. Conclusion

The provided approach provides a formalism for mixing quantities which may have possible or probable values with their interdependencies. The algebraic structure we defined operates on chained computations on such quantities with properties similar to a vector space.
We have extended our formalism on continuous quantities. In special cases such as trapezoids for possibilities and normal distributions for probabilities), some algebraic properties have been maintained. However, combinations of continuous quantities, including probabilistic ones require additional assumptions that imply non rigorous mathematic computations.
The next steps are to compute mixed continuous quantities, by considering either the trapezoidal possibility distributions as intervals of cumulative distribution functions (Destercke S., Dubois D., 2009) or probability-possibility transformations (Dubois et al, 2004).

References

Dantan, J., Pollet, Y., Taïbi, S. 2015. Combination of imperfect data in fuzzy and probabilistic extension classes. Journal of Environmental Accounting and Management. 3 (2), 123-150. DOI: 10.5890/JEAM.2015.06.004.

Destercke S., Dubois D., 2009. The role of generalised p-boxes in imprecise probability models. In proceedings of 6. International Symposium on Imprecise Probability (p. 179-188). Presented at ISIPTA '09, Durham, GBR (2009-07-14 - 2009-07-18).

Dubois D., Foulloy L., Mauris G., Prade H., 2004. Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. Reliable Computing. August 2004, Volume 10, Issue 4, pp 273-297.

Dubois, D. and Prade, H., 1988. Théorie des possibilités, Application à la représentation des connaissances en informatique. Masson 1988. (In French).

Gacôgne L., 1997. Éléments de logique floue. CNAM, Institut d'informatique d'Entreprise, p. 47, may 1997.

Zadeh L.A., 1978. Fuzzy Sets as a basis for a Theory of Possibility. Fuzzy Sets and Systems, 1, 1978.