Combining expert knowledge with data based on belief function theory: an application in waste water treatment

Sebastien POPULAIRE[1,2] and Thierry DENŒUX[2]

(1) Information Technology Division
Technical and Research Center, Ondeo Services,
F-60471 Compiègne, France
sebastien.populaire@hds.utc.fr

(2) UMR CNRS 6599, Heudiasyc
Université de Technologie de Compiègne
B.P. 20529, F-60205 Compiègne, France
thierry.denoeux@hds.utc.fr

Abstract— **This paper presents a methodology for combining expert knowledge with information from statistical data, in classification and prediction problems. The method is based on (1) a case-based approach allowing to predict a quantity of interest from past cases in the form of a belief function, (2) Bayesian networks for modelling expert knowledge and (3) a tuning mechanism allowing to optimally discount information sources by optimizing a performance criterion. This methodology is applied to the prediction of chemical oxygen demand solubility in wastewater. The approach is expected to be useful in situations where both small databases and partial expert knowledge are available.**

Keywords— **Belief functions, Dempster-Shafer theory, evidential reasonning, Bayesian networks, environmental engineering.**

I. Introduction

The task of building a classifier in a given domain is a difficult problem when only few data are available and, consequently, a lot of the domain cases are not listed. This task is even more difficult when data are uncertain. A way to handle with uncertainty is the use of the Demspster-Shafer theory of evidence [15], which allows to represent any degree of partial knowledge ranging from full knowledge to complete ignorance. In particular, Smets [16] developped a coherent and axiomatically justified interpretation of this theory called the Tranferable Belief Model (TBM). Belief functions encompass probability and possibility measures as special cases, thus constituting a very general and flexible framework for uncertainty representation.

A way to complement information given by poor quality data is to use the knowledge of one or several domain experts. Bayesian Networks (BN's) provide a convenient formalism to encode expert knowledge, and more generally, relations between domain variables [13]. BN's can be constructed in three differents ways: from expert knowledge, from data only when a comprehensive data collection is available, or using both sources of information (see [6] for a review of these techniques). But it can be risky to combine information from different sources in one single network [8].

In this paper, we propose a method for the fusion of expert knowledge with data. Section 2 briefly recalls the background of this study, i.e., the theory of belief functions, evidential case-based classification, and BN's. Section 3 addresses in a litte more detail the issue of building BN's from expert knowledge. We then show in Section 4 how these knowledge sources expressed in the belief function formalism can be optimally combined. Finally, Section 5 describes a real-world application of the above tools to the prediction of chemical oxygen demand (COD) solubility, a parameter of interest for the design of wastewater treatment plants.

II. Background

A. Belief function Theory

In this section, we briefly review the basis of the belief function theory and the TBM (see [16] for further informations). In Dempster-Shafer theory, a problem is represented by a set Θ of mutually exclusive hypotheses, called the frame of discernement [15]. A basic belief assignement (bba) is a function $m : 2^\Theta \mapsto [0, 1]$ verifying

$$\sum_{A \subseteq \Theta} m(A) = 1.$$

A bba verifying $m(\emptyset) = 0$ is said to be normal, but this condition is not necessarily imposed in the TBM. The quantity $m(A)$ can be interpreted as a measure of the belief that is commited exactly to A, given the available evidence. Two evidential functions derived from the bba are the credibility function Bel and the plausibility function Pl, defined as

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

and

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B),$$

for all $A \subseteq \Omega$. $Bel(A)$ quantifies the total amount of justified specific support given to A, whereas $Pl(A)$ corresponds to the total amount of potential specific support given to A.

Two useful operations that play a central role in the manipulation of belief functions are discounting, and Dempster's rule of combination. The discounting operation is used when a source of information provides a bba m, but one knows that this source has probability $1 - \alpha$ of being reliable. The bba m is then discounted

by a factor α, resulting in a new bba defined as:

$$m^{\alpha}(A) = (1-\alpha)m(A) \quad \forall A \subset \Theta \tag{1}$$
$$m^{\alpha}(\Theta) = \alpha + (1-\alpha)m(\Theta) \tag{2}$$

Consider now two pieces of evidence on the same frame Θ represented by two bbas m_1 and m_2. The joint bba quantifying the combined impact of these two pieces of evidence is obtained through the conjunctive combination rule as follows:

$$(m_1 \cap m_2)(A) = \sum_{\{B,C \subseteq \Theta : B \cap C = A\}} m_1(B)m_2(C) \tag{3}$$

The conjunctive combination followed by a normalization step is known as Dempster's rule of combination [15]. It is noted $m_1 \oplus m_2$.

A probability function $BetP$ on Θ can be derived from the m function in order to make decisions. It is called the pignistic probability function, obtained by applying the pignistic transformation, defined by

$$BetP(A) = \sum_{B \subseteq \Theta} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)}, \quad \forall A \subseteq \Theta \tag{4}$$

B. Case-based evidential classification

A cased-based evidential classification procedure was introduced by Denœux in [2], and subsequently refined in [19] and [3]. This approach allows to consider a learning set $\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1...N\}$ where $\mathbf{x_i}=(x_{i,1}, x_{i,2}, ...x_{i,p})$ is a p-dimensional vector of measurements related to object i, and m_i is a bba over a set $\Theta = \{\theta_1, \ldots, \theta_K\}$ of classes, which represents the partial knowledge about the class c_i of object i. The learning task considered is to determine the bba of a new case, given the observed value $\mathbf{x}$ of the measurement vector.

The method described in [2] consists in calculating the distance $d(\mathbf{x_i}, \mathbf{x})$ between $\mathbf{x}$ and each $\mathbf{x_i}$ of the learning set, using a suitable distance measure. Each bba m_i is then discounted by a factor α_i defined as an increasing function of the distance $d(\mathbf{x_i}, \mathbf{x})$. The discounted bba's are then finally combined using either the conjunctive rule, or Dempster's rule of combination. A learning scheme for optimizing the function relating the discounting factors to the distances is decribed in [19].

C. Bayesian Networks

A BN (also called probabilistic network or causal network) is a probabilistic expert system in which knowledge can be divided into two parts: a qualitative part and a quantitative part. The qualitative part encodes the causal relations among the domain variables in the form of a directed acyclic graph. Each node of the graph represents a random variable and each arc, a causal dependence between variables. The quantitative part of a BN consists of prior probability distributions over the variables on which no arc is directed, and conditional probability distributions over the variables that have predecessors. A BN is a representation of the joint probability distribution over all the variables.

BN's allow to calculate the conditional probabilities of the unobserved nodes in the network given the values of some observed nodes. The main advantage of BN's is to considerably reduce the amount of numbers that are necessary to describe the entire joint distribution. Some good introductions to Bayesian networks can be found in [11], [13] and [1].

III. Building a Bayesian network

A. Graph structure

The task of building the network structure is the first step in BN construction [10]. The objective is to describe causal relationships between variables of the domain. A method to elicit this knowledge from experts is described in [5]. The first step is building the model from the focus of the study. The expert is asked about the variables that could have an influence on the cited focus, or that could be influenced by it. Then, the same questions are asked, for every new variable given by the first step until the answer is no. So, a graph is built, for which variables are causally linked by arcs. To complete this information, the expert is also asked to sign the interactions among variables in the model. The aim of this operation is to build a Qualitative Probabilistic Network (QPN). QPN is the qualitative abstraction of BN. QPN have the same structure than BN, but instead of quantifying them with probabilities, we just try to determine if, for example, the truth of proposition a makes a proposition b (linked to a by an arc) more or less probable. For further information about QPNs, see [18] and [4].

B. Quantifying the probabilities

Once the structure of the BN is built, the second step consists in fixing the probabilities that will quantify the network. In this part, we will describe the two main tools that has been used: Noisy-Or gates and the probability scale.

B.1 Noisy-Or Gates

Noisy-Or gates are very useful tools that allow to reduce considerably the number of probabilities that have to be elicited. Indeed, in our application, expert time was scarce and we had to find all the ways that could reduce the time that an expert had to spend for the elicitation of probabilities.

We give here a short explanation of Noisy-Or Gates principles. See [12] for a complete information. Noisy-Or Gates are generally used to quantify relationships between n causes $X_1, X_2...X_n$ and one effect Y. In the network, this relation is modelled by an arc that links nodes $X_1, X_2...X_n$ to the node Y. Then, two conditions have to be satisfied:

1. each cause X_i has the probability p_i of being sufficient to cause the effect Y when all other causes are absent;
2. the ability of each cause X_i being sufficient to cause Y is independent of all other causes.

In the case of binary variables, the above two conditions allow to define the conditional joint probability distributions with n numbers instead of 2^n. p_i represents the probability that Y is true when X_i is present and all

other causes $X_j, j \neq i$ are absent:

$$p_i = Pr(y \mid \overline{x_1}, \overline{x_2}, \ldots, x_i, \ldots, \overline{x_n}) \quad (5)$$

and for X_p, a subset of true variables $X_i s$, we obtain:

$$Pr(y|X_p) = 1 - \prod_{\{i:X_i \in X_p\}} (1 - p_i). \quad (6)$$

The full conditional probability distribution of Y given causes $X_1, X_2, ..., X_n$ can then be computed.

B.2 The probability scale

In the last step, probabilities have to be assessed by experts. This is often considered as a difficult part because experts are generally reluctant to express numerical probabilities. For an overview of methods that helps solving this problem, see [6] and [14]. For example, the probability scale described in [14] can be used (a real world application of BN in medicine using this scale can be found in [17]). This scale, shown in Figure 1, uses both words and numbers, in order to help experts who do not feel comfortable with numbers.

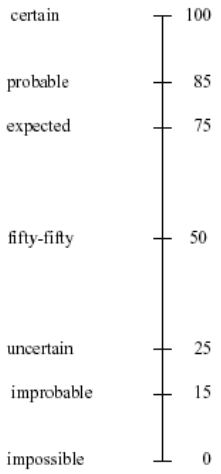

Fig. 1. The probability scale

IV. Combination of data information and expert knowledge

Let us now assume that we have, for a given classification problem, the following sources of information:
- a data set $\mathcal{E}$ related to N cases with partially known classification, and
- knowledge from one or several experts, modeled by B Bayesian Networks BN_b, $b = 1, \ldots, B$.

We want to build a classifier that efficiently combines these information sources. This can be achieved in the TBM framework, using the expert tuning technique introduced by Elouedi et al. [9]. For each object o_i in the data base, let $m\{o_i\}[\mathcal{L}^{-i}]$ denote the bba concerning the class of o_i, induced by the rest of the learning set (i.e., the N other learning cases), and constructed as explained in Section II-B. Let $P\{o_i\}[BN_b]$ denote the probability function regarding the class of o_i, computed using BN_b, $b = 1, \ldots, B$. For each object o_i in the learning set, we go through the following steps:
1. Apply a discounting factor α on bba $m\{o_i\}[\mathcal{L}^{-i}]$

$$m^{\alpha}\{o_i\}[\mathcal{L}^{-i}](A) = (1-\alpha_1)m\{o_i\}[\mathcal{L}^{-i}](A), \forall A \subset \Theta \quad (7)$$

$$m^{\alpha}\{o_i\}[\mathcal{L}^{-i}](\Theta) = \alpha_1 + (1-\alpha_1)m\{o_i\}[\mathcal{L}^{-i}](\Theta) \quad (8)$$

2. Apply a discounting factor β_b on each probability distribution $P\{o_i\}[BN_b]$

$$m^{\beta_b}\{o_i\}[BN_b](A) = (1-\beta_b)P\{o_i\}[BN_b](A), \forall A \subset \Theta \quad (9)$$

$$m^{\beta_b}\{o_i\}[BN_b](\Theta) = \beta_b \quad (10)$$

3. Combine the above $B+1$ discounted bba's using Dempster's rule, to obtain a combined bba $m\{o_i\}$.
4. Build the pignistic probabilities $BetP\{o_i\}$ based on $m\{o_i\}$.
5. Compute the distance $dist_i$ between $BetP\{o_i\}$ and the most probable class of object o_i, defined as:

$$dist_i = \sum_{k=1}^{K} (BetP\{o_i\}(\theta_k) - \delta_{i,k})^2 \quad (11)$$

where $\delta_{i,k} = 1$ if θ_k is the class of maximum pignistic probability according to m_i, and 0 otherwise.
Finally, we find the discounting factors $(\alpha, \beta_1, \ldots, \beta_b)$ that minimize the total error

$$TDist = \sum_{i=1}^{N} dist_i, \quad (12)$$

using a numerical optimization procedure.

V. Application: prediction of COD solubility

A. General

Pollution in wastewater can be decomposed into three parts: a settled part, a coagulable part and a soluble part. It is important for wastewater treatment specialists to determine the ratio of each of these parts, in order to adapt the treatment that will efficiently remove pollution. Pollution solubility has been analyzed in different countries and different cities. These studies have revealed an important variability from one place to an other: water pollution solubility is influenced by environmental conditions, sewage systems features and sample taking point.

B. The database

The available database was composed of observations made on 76 wastewater treatment plants. Each case i was described by
- a vector $\mathbf{x_i}$ of 8 binary variables $(x_{i1} ... x_{i8})$ describing the main characteristics of the sewage system (about 15 % of the values were missing),
- n_i measurements of COD solubility (CODS) taken at different times.

The COD solubility variable was discretized in three classes: Low, Medium and High, and for each case i a bba m_i was constructed as follows. First, the class

c_i corresponding to the average of the n_i CODS measurements was determined. A bba m_i was then defined as:

$$m_i(\{c_i\}) = \xi_i \quad (13)$$
$$m_i(\Theta) = 1 - \xi_i \quad (14)$$
$$m_i(A) = 0 \quad \forall A \in 2^\Theta \setminus \{\Theta, \{c_i\}\} \quad (15)$$

with

$$\xi_i = \frac{n_i}{\max_i n_i + 1}$$

The bba m_i thus reflects the amount of knowledge related to case i: the more measurements are available, the more precise is m_i.

C. Expert knowledge

As explained in Section III-A, the first step has been the construction of the graph and the QPN. The obtained result can be seen in Figure 2.

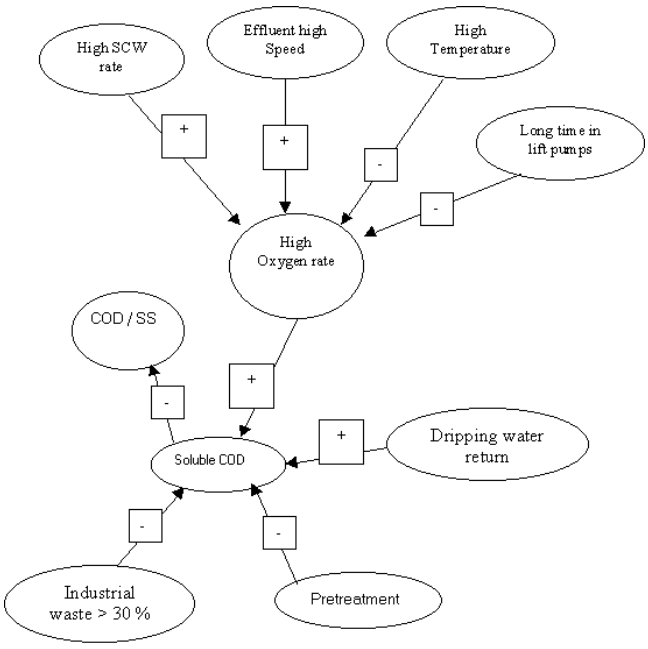

Fig. 2. The qualitative probabilistic network

In order to quantify the BN, a questionnaire containing 38 questions was prepared and submitted to two experts. Each question corresponded to one of the graph probability, and each expert had to give his answer using the scale shown in Figure 1. We thus obtained two distinct BN's, with the same structure but different probabilities.

To test the BN, we used the database described in Section V-B and we proceeded as follow. For each case of the database, we:

- entered the known values of the measured variables in the BN,
- obtained a probability for each class of CODS.

The GeNIe software [7] was used for the implementation.

D. Results

The confusion matrices of the cased-based classifier, the two BN's, and the optimized classifer using the three sources of information are reported in Tables I, II, III and IV, respectively. The cased-based and combined classifier error-rates were obtained using the leave-one-out method: the error for each case i was computed using a classifier built exclusively from a training set $\mathcal{L}^{-i}$ composed of the other $N-1$ training cases. Consequently, the reported error estimates are unbiased. Table V shows the the sum $TDist$ of squared distances between output probabilities and class indicator variables as defined in (12), for the four classification rules. The optimized discounted factors are

$$\alpha = 0 \quad \beta_1 = 0 \quad \beta_2 = 1$$

We can see that the BN obtained from expert 2, which performs rather poorly (Table III), is effectively removed from the combination using the discounting procedure. The probabilities produced by the BN of expert 1, although insufficient to provide reliable prediction (Table II), do improve the performances of the case-based classifier (Table V).

TABLE I
CONFUSION MATRIX FOR THE CASE-BASED CLASSIFIER

	predicted class		
real class	High	Medium	Low
High	6	4	3
Medium	9	28	9
Low	3	7	7

TABLE II
CONFUSION MATRIX FOR THE BN OF EXPERT 1

	predicted class		
real class	High	Medium	Low
High	3	10	0
Medium	0	46	0
Low	0	17	0

TABLE III
CONFUSION MATRIX FOR THE BN OF EXPERT 2

	predicted class		
real class	High	Medium	Low
High	12	0	1
Medium	40	0	6
Low	17	0	0

VI. CONCLUSION

In this paper, we have described a method for building a classification system in the TBM framework, using both a case-based approach and expert knowledge

TABLE IV
CONFUSION MATRIX FOR THE OPTIMIZED COMBINED CLASSFIER

	predicted class		
real class	High	Medium	Low
High	5	7	1
Medium	9	36	1
Low	3	12	2

TABLE V
VALUE OF THE TDIST CRITERION FOR THE FOUR CLASSIFICATION RULES

Method	Criterion Value
case-based	82
BN - Expert 1	78
BN - Expert 2	92
Tuning	76

encoded as BN's. The optimal tuning of the different information sources is realized by minimizing an empirical error criterion. The main advantage of the method is the ability to combine effectively the information that can be inferred from small amounts of possibly poor quality data, with partial knowledge elicited from experts. The presented case-study in wastewater treatment shows that this methodology can be used in real-world applications.

ACKNOLEDGEMENTS

We would like to thank Joëlle Blanc, Philippe Ginestet, Jean-Marc Audic, Robert Turgis, Christian Fayoux and Albert Mpe A Guilikeng from Ondeo Services for their help in the knowledge elicitation process. We also would like to thank Marek Druzdzel, Silja Renooij and Javier Diez for sharing their knowledge of Bayesian networks.

REFERENCES

[1] E. Charniak. Bayesian networks without tears. *AI Magazine*, 12(4):50–63, 1991.

[2] T. Denœux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. Syst. Man. Cybern.*, SMC-25(5):804 – 813, 1995.

[3] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasonning. *Fuzzy Sets and Systems*, 122(3):47 – 62, 2001.

[4] M. Druzdzel. Qualitative verbal explanations in bayesian belief networks. *AI and Simulation of Behaviour Quarterly (special issue on Bayesian belief networks)*, 94:43–54, 1996.

[5] M. Druzdzel, A. Onisko, D. Schwartz, J. Dowling, and H. Wasyluk. Knowledge engineering for very large decision-analytic medical models. In *Proceedings of the 1999 Annual Meeting of the American Medical Informatics Association (AMIA–99)*, page 1049, Washington D.C., 1999.

[6] M. Druzdzel, L. van der Gaag, M. Henrion, and F. Jensen. Building probabilistic networks: where do the numbers come from. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):481 – 486, 2000.

[7] M.J. Druzdzel. Genie: a development environment for graphical decision-analytic models. In *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association (AMIA-1999)*, 1999.

[8] M.J. Druzdzel and F.J. Diez. Criteria for combining knowledge from different sources in probabilistic models. In *Working Notes of the UAI 2000 Workshop on Domain Knowledge with Data for Decision Support*, 2000.

[9] Z. Elouedi, K. Mellouli, and P. Smets. The evaluation of sensor's reliability and their tuning for multisensor data fusion within the transferable belief model. In *EC-SQARU 2001*, pages 305 – 361, 2001.

[10] M. Henrion. Some practical issues in constructing belief networks. In *Uncertainty in Artificial Intelligence 3*, pages 161–174, 1987.

[11] F. Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[12] Agnieszka Onisko, Marek J. Druzdzel, and Hanna Wasyluk. Learning bayesian networks parameters from small data sets : Application of noisy-or gates. In *Working Notes of the Workshop on 'Bayesian and Causal Networks : From Inference To Data Mining', 12th European Conference on Artificial Intelligence*, 2000.

[13] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, California, 1988.

[14] S. Renooij and C. Witteman. Talking probabilities: communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22:169 – 194., 1999.

[15] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[16] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191 – 234, 1994.

[17] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, and et al. Probabilities for a probabilistic network: A case-study in oesophageal carcinoma.

[18] Michael P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990.

[19] L. M. Zouhal and T. Denœux. An evidence-theoretic k-nn rule with parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, 28(2):263–271, 1998.