Uncertainty and learning in expert systems

Uncertainty and Learning in Expert Systems

Incertitude et apprentisage dans les systèmes experts

Ungewissheit und Lernen in Expertensystemen

Enrique CASTILLO
Professor
Univ. of Cantabria
Santander, Spain

Elena ALVAREZ
Assistant Professor
Univ. of Cantabria
Santander, Spain

SUMMARY
The paper discusses some of the problems associated with conventional uncertainty propagation methods, as those based on independent probabilities, certainty factors, belief or possibility functions, and shows, by giving examples, the importance of associated errors. Then, alternative methods, based on log-linear, regression and casual networks or influence diagram models are discussed. Finally, their structural and parametric learning possibilities are analyzed.

RESUME
Le but de ce article est de montrer quelques problèmes associés aux méthodes conventionelles de propagation d'incertitude, tels que les dérivés des probabilités indépendantes, les coefficients de vraisemblance et les fonctions de possibilité. Nous montrons avec des exemples l'importance des erreurs associées à ces méthodes. Quelques méthodes alternatives, fondées sur des modèles logarithmiques linéaires, de regression et de réseaux ou diagrammes d'influence sont discutés. Finalement nous présentons leurs possibilités d'apprentissage paramétriques et structurales.

ZUSAMMENFASSUNG
Die vorliegende Arbeit behandelt einige Probleme, im Zusammenhang mit den konventionellen Fehlerfortplanzungsmethoden wie: unabhängige Wahrscheinlichkeiten, Gewissheitsfaktoren sowie Glaubens- oder Möglichkeitsfunktionen. Anhand von Beispielen wird die Bedeutung der aus den Ansatzhypothesen entstandenen Fehler gezeigt. Einige alternative Methoden, die auf Regressions — und linear-logarithmischen Modellen, sowie auf Kausalnetzen und Einflussdiagrammen beruhen, werden anschliessend vorgeschlagen. Zuletzt wird die Möglichkeit eines strukturellen und parametrischen Erlernens analysiert.

1.- INTRODUCTION

In classical logic any statement is either true or false; however, when working with uncertain implications, statements must be understood as possible rather than certain. Thus an uncertainty measure is necessary. The oldest measure of uncertainty and the most intuitive is probability. However, other measures are utilized in the field of expert systems, such as certainty factors, the measures of evidence theory and the possibility functions of fuzzy logic.

2.- UNCERTAINTY PROPAGATION

The main problem of coherence arises when propagation of uncertainty is involved. Some propagation formulas without an axiomatic basis have been proposed and accepted by the Artificial Intelligence community [3]. Many of the propagation formulas used are no better than the oft-criticized, independence assumption. When we deal with single evidence units, the calculation of uncertainty measures is straight forward, but what happens when we need to combine several single evidence units to get a mixed evidence?. In this section we shall analyze this question.

The problem of propagation of uncertainty in the case of probability can be reduced to the calculation of probabilities conditioned by all units of information [2]. In order to illustrate the problem we give the following example.

Example 1.- Let us assume that an engineer suspects the presence of problem E and that, based on some available data, he has arrived to a prior probability for E of 0.8. Because 0.8 is not high enough to make a decision (note that making a decision at this moment implies a probability 0.2 of error), he decides to obtain more information. Thus, he makes use of the following information, which is shown in figure 1a, where the shadowed area refers to hystorical cases with problem E and the white area to those without E, the symbols S_1, S_2 and S_3 refer to three symptoms related to E and the figures are frequencies (the knowledge base).

From figure 1a, the following information (prior probabilities and likelihoods) can be derived:

P(E) = 0.80	P(no E) = 0.20	$P(S_1 / E) = 0.70$	$P(S_1 / \text{no E})=0.10$
$P(S_2/ E)=0.80$	$P(S_2/ \text{no E})=0.20$	$P(S_3/ E)=0.60$	$P(S_3 / \text{no E})=0.30$

This figure will allow us to illustrate the failures of the assumption of independence and to see how completely erroneous results can be obtained by using this assumption. It is important to indicate that the above information (prior probabilities and likelihoods) is not sufficient to completely define a probability. In other words, there are many different probability or frequency distributions having as prior probabilities and likelihoods the above values. In figure 1 we show two of them.

Let us assume that the engineer receives items of information in the following order: 1.-initial data, 2.-presence of S_1, 3.-presence of S_2 and 4.-absence of S_3. Table 1 gives the

updated probabilities of E after the four steps indicated above for the two cases in figure 1. It is interesting to point out that for case (b), the real probability of $P(E/S_1, S_2$ and no $S_3)$ is zero, while that obtained from the independence assumption is 0.989. This suggests that care must be taken with the indiscriminated use of independence.

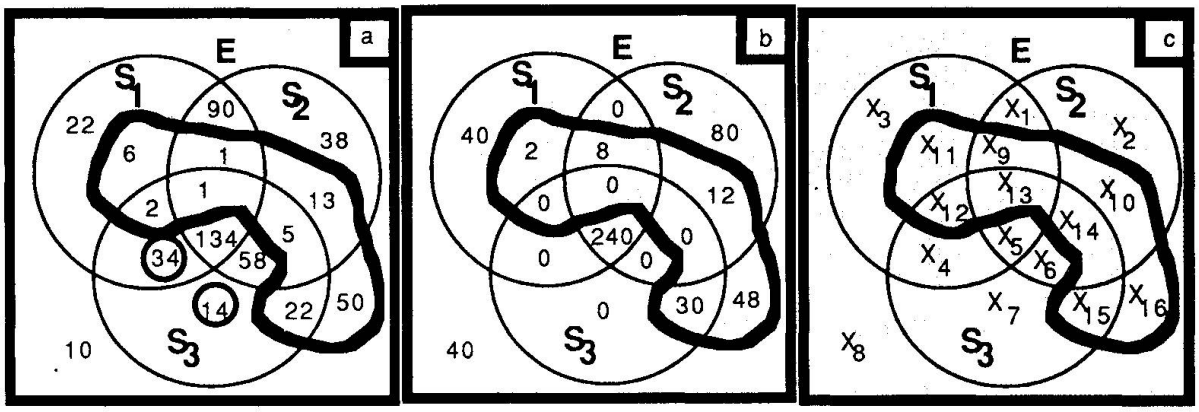

Figure 1.- Two different solutions with the same prior probabilities and likelihoods and notation

	P(E)	$P(E / S_1)$	$P(E / S_1, S_2)$	$P(E / S_1, S_2,$ no $S_3)$	
				real	independence
case a	0.80	0.966	0.994	0.989	0.989
case b	0.80	0.966	0.968	0.000	0.989

Table 1.- Updating of probabilities

The existence of many probabilities with given prior probabilities and likelihoods, suggests the method of calculating lower and upper bounds of desired probabilities under these constraints. In this way, an interval $[P_{min}(A), P_{max}(A)]$, which measures ignorance and uncertainty, can be obtained. An example is now given.

Example 2.- Let us consider the case of example 1. If we call X_1 to X_{16} the frequencies shown in figure 1.c, fixing the values of prior probabilities and likelihoods, as in example 1, is equivalent to using the constraints:

$$P(E) = 0.8 \leftrightarrow X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 = 400$$

$$P(\text{no } E) = 0.2 \leftrightarrow X_9 + X_{10} + X_{11} + X_{12} + X_{13} + X_{14} + X_{15} + X_{16} = 100$$

$$P(S_1 / E) = 0.7 \leftrightarrow 3X_1 - 7X_2 + 3X_3 + 3X_4 + 3X_5 - 7X_6 - 7X_7 - 7X_8 = 0$$

(1) $P(S_1 / \text{no } E) = 0.1 \leftrightarrow 9X_9 - X_{10} + 9X_{11} + 9X_{12} + 9X_{13} - X_{14} - X_{15} - X_{16} = 0$

$$P(S_2 / E) = 0.8 \leftrightarrow 2X_1 + 2X_2 - 8X_3 - 8X_4 + 2X_5 + 2X_6 - 8X_7 - 8X_8 = 0$$

$$P(S_2 / \text{no } E) = 0.2 \leftrightarrow 8X_9 + 8X_{10} - 2X_{11} - 2X_{12} + 8X_{13} + 8X_{14} - 2X_{15} - 2X_{16} = 0$$

$$P(S_3 / E) = 0.6 \leftrightarrow -6X_1 - 6X_2 - 6X_3 + 4X_4 + 4X_5 + 4X_6 + 4X_7 - 6X_8 = 0$$

$$P(S_3 / \text{no } E) = 0.3 \leftrightarrow -3X_9 - 3X_{10} - 3X_{11} + 7X_{12} + 7X_{13} + 7X_{14} + 7X_{15} - 3X_{16} = 0$$

The probability of any set, A, can be written as $P(A) = \sum_{i=1}^{16} a_i X_i$ where the coefficients a_i (i=1,2,...,16) are ones or zeros depending on whether or not the set associated with X_i belongs to the set A.

Determination of the interval $[P_{min}(A), P_{max}(A)]$ can be reduced to solving the following two linear programming problems:

$$\text{Minimize} \sum_{i=1}^{16} a_i X_i \text{ subject to (1)} \quad \text{and} \quad \text{Maximize} \sum_{i=1}^{16} a_i X_i \text{ subject to (1)}$$

If what we desire is an interval of conditional probabilities, the above problems are equivalent to the following two non-linear programming problems:

$$\text{Minimize} \quad \sum_{i=1}^{16} a_i X_i \Big/ \sum_{i=1}^{16} b_i X_i \quad \text{and} \quad \text{Maximize} \quad \sum_{i=1}^{16} a_i X_i \Big/ \sum_{i=1}^{16} b_i X_i \quad \text{subject to (1)}$$

which are equivalent to sequences of linear problems:

$$\min_{\lambda} \left[\text{Min} \sum_{i=1}^{16} a_i X_i / \lambda \quad \text{subject to (1)} \quad \text{and to} \sum_{i=1}^{16} b_i X_i = \lambda \right]$$

and

$$\max_{\lambda} \left[\text{Max} \sum_{i=1}^{16} a_i X_i / \lambda \quad \text{subject to (1)} \quad \text{and to} \sum_{i=1}^{16} b_i X_i = \lambda \right]$$

where the coefficients b_i (i=1,2,...,16) are also zeros or ones.

The propagation of the belief and unbelief measures and of the certainty factors, CF, is usually carried out by means of the well known Dempster's formula. In order to illustrate some of the problems associated with this formula, Table 2 shows the exact values and those resulting from it.

	$CF(E ; S_1)$	$CF(E ; S_1, S_2)$	$CF(E ; S_1, S_2, \text{no } S_3)$	$CF(E ; S_1, S_2, \text{no } S_3)$ Propagation formulas
case a	0.873	0.970	0.945	0.989
case b	0.828	0.839	-1.000	0.972

Table 2.- Updating of certainty factors

Note the extremely large difference between the exact and the calculated certainty factor $CF(E ; S_1, S_2, \text{no } S_3)$ in case b. This result proves that the above propagation formula is not satisfactory in this case and should warn the user against its uncontrolled use.

Similar errors result from evidence theory or fuzzy logic if standard propagation formulas are used.

3.- STATISTICAL MODELS IN EXPERT SYSTEMS

Most of the problems mentioned above come from the fact that uncertainties of composed events cannot be derived from uncertainties of single events. Thus, a precise uncertainty propagation technique requires models to include frequencies of composed events as well as those of single events. In this section we describe log-linear, regression and causal network models. They are three of the most useful techniques to solve the above problems.

3.1.- Log-linear models

The most general log-linear model is of the form [1]:

$$\log m_{ijk..r} = u + u_1(i) + .. + u_s(r) + ... + u_{(s-1)s}(qr) + ... + u_{12..s}(ij...r)$$

where $m_{ijk...}$ denote the frequency of the class defined by the i-th problem, j-th level of the first symptom, k-th level of second symptom, and so on, parameters must satisfy some additional constraints and indexes vary from 1 to the number of levels for each symptom.

Example 3.- If the above model is fitted to data in figure 1.a we get the log-linear model

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il)$$
$$u = 2.6429 ; u_1(1) = 0.938 ; u_2(1) = -0.337 ; u_4(1) = -0.110$$
$$u_{12}(1,1) = 0.761 ; u_{13}(1,1) = 0.693 ; u_{14}(1,1) = 0.313$$

			case a		case b	
S_1	S_2	S_3	E	no E	E	no E
YES	YES	YES	134(134.4)	1(0.6)	240 (239.7)	0 (0)
YES	YES	NO	90(89.6)	1(1.4)	0 (0.3)	8 (8)
YES	NO	YES	34(33.6)	2(2,4)	0 (0)	0 (0)
YES	NO	NO	22(22.4)	6(5.6)	40 (40)	2(2)
NO	YES	YES	58(57.6)	5(5.4)	0(0)	0 (0)
NO	YES	NO	38(38.4)	13(12.6)	80 (80)	12(12)
NO	NO	YES	14(14.4)	22(21.6)	0 (0)	30(30)
NO	NO	NO	10(9.6)	150(50.4)	40(40)	48(48)

Table 3.- Real values and predictions for frequencies in figure 1

Similarly, for data in figure 1.b, we get the model

$$\log m_{ijkl} = u + u_1(i) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{24}(jl) + u_{34}(kl) + u_{123}(ijk)$$
$$u = 4.004 ; u_1(1) = 0.188 ; u_3(1) = 0.117; u_4(1) = 1.534 ;$$
$$u_{12}(1,1) = -0.241 ; u_{13}(1,1) = -0.515; u_{23}(1,1) = -0.342 ;$$
$$u_{24}(1,1) = 1.137; u_{123}(1,1,1) = -1.035 ; u_{34}(1,1) = 0.633 ;$$

These models have 7 and 10 degrees of freedom, respectively, implying a saving of 9 and 6 parameters with respect to the general model.Table 3 shows the real values of frequencies and those given by the above models (in brackets).

3.2.- Regression models

The log-linear models in the previous section are useful for symptoms or variables of a discrete type, but not for continuous symptoms unless they are made discrete by subdivision into a finite number of intervals. With the aim of solving this problem, regression models are developed. The model to be described in this section is of logistic type [2]:

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{r} \beta_j A_{ij} \;;\; \beta_j \equiv u_{i_1 i_2 \ldots i_s}(j_1 j_2 \ldots j_s) \;;\; 1 \le j_k \le l_k \;;\; A_{ij} = f_{i_1 i_2 \ldots i_s}(x_1 x_2 \ldots x_t)$$

where p_i is the probability of the disease conditioned by the given symptoms, l_j is the number of levels of the j-th discrete symptom and the functions $f(x_1,x_2,\ldots,x_t)$, which are given, can be constant, take on value one, if the term represents the influence of a group of discrete symptoms only, or be null. Analogously, the "u" parameters can degenerate to unity if the term reflects the influence of continuous symptoms alone.

If we have enough data of patients with their diseases and symptoms, the parameters of the regression model can be easily estimated by the maximum likelihood method ([2],[5]).

Example 4.- Assume that one engineer is interested in distinguishing the following 4 problems in a nuclear power plant based on the following 4 symptoms:

1.- Recirculation line large break	X_1=Reactor pressure (RP)
2.- Loss of vacuum condenser	X_2=Vessel water level (VWL)
3.- Loss of offside power	X_3=Drywell pressure (DP)
4.- Main steam line small break	X_4=Closed main steam valve (CMSV) (-1=no, 1=yes)

and that he has tha data shown in Table 4.

In order to make the distinction between those different problems he decides to fit 4 regression models (one for each problem) such that given the four symptoms indicated in Table 4, the probability of not having each problem can be calculated.

A very general logistic model is the following

$$\log\left(\frac{p_i}{1-p_i}\right) = u_0 + u_1 X_1 + u_2 X_2 + u_3 X_3 + u_4 X_4 + u_5 X_1 X_2 + u_6 X_1 X_3 + u_7 X_1 X_4 + u_8 X_2 X_3 +$$
$$+ u_9 X_2 X_4 + u_{10} X_3 X_4 + u_{11} X_1 X_2 X_3 + u_{12} X_1 X_2 X_4 + u_{13} X_1 X_3 X_4 + u_{14} X_2 X_3 X_4 + u_{15} X_1 X_2 X_3 X_4$$

where p_i is the probability of not having problem i-th and the u coefficients are constants to be estimated. From the data above, stepwise regression (a method for selecting which symptoms are and which are not relevant for the distinction) leads to the following models:

$$\log\left(\frac{p_1}{1-p_1}\right) = -0.21 + 10.776\, X_4 \quad ; \quad \log\left(\frac{p_2}{1-p_2}\right) = 472.86 - 6.472\, X_1$$

$$\log\left(\frac{p_3}{1-p_3}\right) = -4.436 + 0.378\,X_2 - 11.34\,X_4 \quad ; \quad \log\left(\frac{p_4}{1-p_4}\right) = 20.74 - 2.1\,X_2 X_3$$

DATA #	PROBLEM	RP	VWL	DP	CMSV	DATA #	PROBLEM	RP	VWL	DP	CMSV
1	1	69	10	0.2	-1	2	1	71	12	0.25	-1
3	1	70	6	0.26	-1	4	1	69	8	0.3	-1
5	1	68	5	0.18	-1	6	1	70	14	0.24	-1
7	1	72	16	0.17	-1	8	1	69	10	0.23	-1
9	1	71	12	0.25	-1	10	1	71	11	0.17	-1
11	2	75	74	0.07	1	12	2	74	75	0.08	1
13	2	77	76	0.08	1	14	2	76	76	0.07	1
15	2	76	75	0.06	1	16	2	77	77	0.07	1
17	2	77	74	0.08	1	18	2	75	74	0.08	1
19	2	76	76	0.07	1	20	2	76	73	0.06	1
21	3	70	15	0.20	1	22	3	71	16	0.17	1
23	3	72	12	0.30	1	24	3	70	17	0.19	1
25	3	69	11	0.25	1	26	3	70	15	0.21	1
27	3	70	15	0.26	1	28	3	70	14	0.20	1
29	3	70	16	0.21	1	30	3	69	13	0.18	1
31	4	68	75	0.20	1	32	4	69	73	0.21	1
33	4	72	76	0.26	1	34	4	70	75	0.19	1
35	4	70	77	0.30	1	36	4	70	74	0.20	1
37	4	69	73	0.25	1	38	4	69	75	0.18	1
39	4	71	74	0.21	1	40	4	71	76	0.21	1

Table 4.- Nuclear power plant data

The above models are surprisingly simple, but they differentiate very well the four problems (see Table 5 where the probabilities p_1, p_2, p_3 and p_4, have been calculated for the above models). The engineer is temptated to use more complicated models (at least with two non-constant terms) because he knows the symptoms associated with those problems (see Table 6) but they are not necessary to the above aim.

DATA #	p_1	p_2	p_3	p_4	DATA#	p_1	p_2	p_3	p_4
1 to 10	0.00	1.00	1.00	1.00	11 to 20	1.00	0.00	1.00	1.00
21to 30	1.00	1.00	0.00	1.00	31 to 40	1.00	1.000	1.00	0.00

Table 5.- Calculated values of p_1 to p_4 using logistic models

PROBLEM	SYMPTOMS
Recirculation line large break	high drywell pressure ($\geq$ 0.14 Kg/cm^2) and low vessel water level (< 18 cm)
Loss of vacuum condenser	high reactor pressure (> 73.5 Kg/cm^2) and closed main steam valve
Loss of offside power	low vessel water level (< 18 cm), high drywell pressure ($\geq$ 0.14 Kg/cm^2) and closed main steam valve
Main steam line small break	high drywell pressure ($\geq$ 0.14 Kg/cm^2) and closed main steam valve

Table 6.- Symptoms associated with the above four problems

The knowledge base in the case of expert systems based on log-linear or regression models consists of their structures and associated parameters, and the inference engine

consists of a program or procedure able to calculate conditional probabilities of problems given certain symptoms by means of the model. As information available is normally incomplete, it is necessary to add all the frequencies associated with the partial given information. Rough estimates can be obtained based on mean values.

3.3.- Causal network models

In this section we describe one modified version of the Lauritzen and Spiegelhalter [4] model, which is one of the methods based on causal networks. In order to illustrate the method, we shall analyze in detail the following pedagogical example.

Example 5.- Figure 2.a shows the security mechanism of a room which is composed of two subsystems. The first, C, consists of a video-camera which transmits the image to a computer for analysis. After the analysis, the computer decides whether or not to activate a relay which closes an electric circuit with a battery activating an alarm. The second, G, consists of a photoelectric cell, D, which closes another electric circuit with an alarm F activated by a battery E. Figure 2.b shows the rules associated with the alarm system. Note that the first system has been simplified to hardware plus software, and that rules are interpreted in a weak sense (conclusions are very likely but not sure).

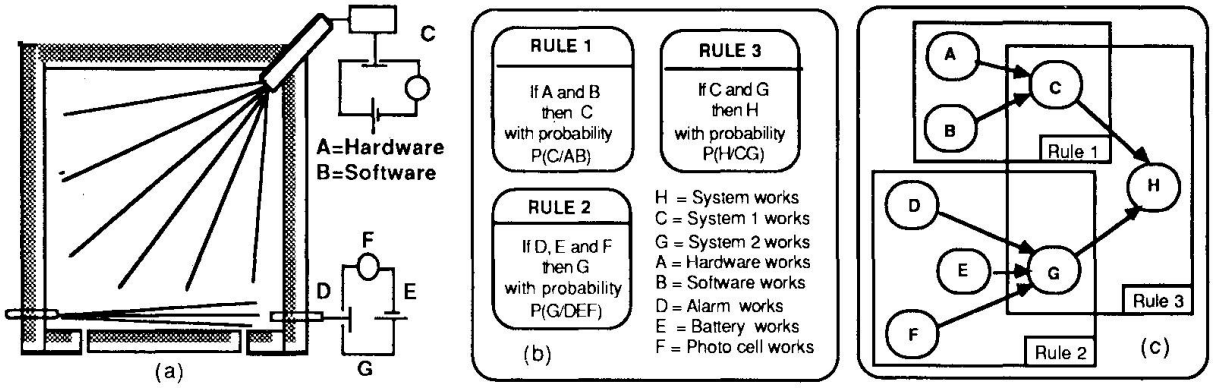

Figure 2.- Security system: rules and influence diagram

The idea of Lauritzen and Spiegelhalter consists of utilizing a probabilistic structure such that propagation of uncertainty can be carried out accurately, quickly and without the need for an excessive number of parameters. To this aim, they assume that the knowledge can be represented by means of an "influence diagram", which is a set V of nodes and a set of oriented edges between pairs of nodes (see figure 2.c). An oriented edge between nodes "A" and "B" can be represented by means of the notation $A \rightarrow B$ and then we say that the node "A" is a father of node "B" and that node "B" is a son of node "A".

A set of nodes **C** is said to be "complete" if there are edges between all pairs of nodes and we say that it is a "clique" if it is maximal, that is, it cannot be extended to another complete set. The set of all extremes of edges of a given node A is called the boundary of that node and is denoted by Bd(A). Numbering of nodes is called "perfect" if the set of nodes $Bd(i) \cap \{1,2,...,i-1\}$ is complete.

The nodes of the graph represent the objects, which can take on a finite number of values. As starting data, "conditional probability tables" are given. These tables contain the probabilities of each node taking each of its possible values for any of the possible combinations of values for its parents. In addition they assume that if we know the values of parents, Π_A, of a node A whose value is currently unknown, then no other knowledge (except concerning descendants of A) will influence our opinion concerning the true value of A (Markov property), that is:

$$P(A / B, U) = P(A / B) \quad ; \quad \forall A, U \subset V - \Pi_A \; ; \; B \subset \Pi_A$$

This implies that the joint probability function of all nodes can be written as the product of conditional probabilities in the above tables.

Sometimes it is easier to use a representation of the joint probability distribution as the product of functions ("evidence potentials", which are denoted as ψ) defined on cliques (the set of cliques will be denoted as Δ). In this case the joint distribution does not need to be known exactly, but can be expressed as a proportional function, which can be normalized if needed. Thus, we have: $P(V) = \prod_{i=1}^{q} \psi(C_i) / Z$ where Z is a normalization constant.

Lauritzen and Spiegelhalter [4] utilize a third form of representing the joint probability of nodes by means of a "set chain" (of cliques) having the running intersection property, in that the nodes of one clique also contained in previous cliques are all members of one previous clique. This property facilitates the calculation of the joint probability functions on cliques. In fact, the above chain is such that:

$$P(V) = \prod_{i=1}^{q} P(R_i / S_i) \quad ; \quad S_i = C_i \cap (C_1 \cup C_2 \cup \ldots \cup C_{i-1}) \; ; \; R_i = C_i - S_i$$

Sets S_i and R_i are called clique separators and residuals, respectively.

From evidence potentials the marginal probability of a given set $U \subset V$ can be easily obtained (when doing this we say that we are marginalising over $\bar{U}$):

$$P(U) = \sum_{U} P(U, \bar{U}) = \sum_{U} Z^{-1} \prod_{A \in \Delta} \psi(A) = Z^{-1} \prod_{A \in \Delta_1} \psi(A) \sum_{U} \prod_{A \in \Delta_2} \psi(A) = Z^{-1} \phi(B) \prod_{A \in \Delta_1} \psi(A)$$

where

$$\Delta_1 = \{A \in \Delta / A \cap U = \phi\} \text{ and } \Delta_2 = \Delta - \Delta_1 \; ; \; B = \bigcup_{A \in \Delta_2} A - U \; ; \; \phi(B) = \sum_{U} \prod_{A \in \Delta_2} \psi(A)$$

Thus, if $\bar{U} = R_q$ then the initial evidence potentials transform to

$$(2)\qquad \bar{\psi}(A)=\left\{\begin{array}{lll}\psi(A)\phi(B) & \text{if} & A=C_{q-1}\\ 1 & \text{if} & A=C_q\\ \psi(A) & \text{otherwise} & \end{array}\right\}\ ;\ \phi(B)=\sum_{R_q}\psi(C_q)$$

where A_1 is an element of Δ_1 such that $B \subset A_1$, and the normalization constant Z is unchanged in this operation.

In addition, for the last clique we can write:

$$(3)\ P(R_q / S_q) = P(R_q / C_1\, C_2 \ldots C_{q-1}) = P(V) / P(C_1\, C_2 \ldots C_{q-1}) = \psi(C_q) / \sum_{R_q} \psi(C_q)$$

Thus, progressive marginalization and expression (3) allow probabilities $P(R_i/S_i)$ (i=1,2,...,q) to be obtained. In fact, $P(R_q/S_q)$ is directly obtained from (3). Then, we marginalise over R_q, using (2), and once again we use (3) to calculate $P(R_{q-1}/S_{q-1})$. Then, we marginalise over R_{q-1} and calculate $P(R_{q-2}/S_{q-2})$, and we repeat the same process until $P(R_1/S_1)$ is obtained.

If the value of node i is known and we want to know how the joint probability distribution P(V) is affected by that information, evidence potentials are modified in the following way:

$$\psi^*_{A^*}(.)=\left\{\begin{array}{ll}0 & \text{if value of node i is contrary to inf ormation}\\ \psi_{A^*}(.) & \text{otherwise}\end{array}\right\}\ ;\ \psi^*_A(.)=\psi_A(.)$$

where A^* is the first clique containing node i and ψ^* are the new evidence potentials.

Initially, conditional probabilities are obtained from the human expert and/or the knowledge engineer or the data base and evidence potentials are obtained from them (see example 6).

Example 6.- As an illustration of the above method, we apply it to the case of example 1 (influence diagram in Figure 2.c and 3). The undirected edges (A,B), (D,E), (D,F), (E,F) and (C,G) have been included to take into account that sets {A,B,C}, {D,E,F,G} and {C,G,H} define the 3 rules for systems 1 and 2 to work, respectively.

If we assume that nodes can take values "true" or "false", we get the following conditional probability tables:

$$P(A,B)\ ,\ P(D,E,F)\ \ ,\ P(C/A,B)\ ,\ P(G/D,E,F)\ \text{ and } P(H/C,G)$$

Thus, the joint probability function of all nodes can be written

$$(4)\qquad P(V) = P(A,B,C,D,E,F,G,H)=P(A,B)P(C/A,B)P(D,E,F)P(G/D,E,F)P(H/C,G)$$

Due to the fact that the cliques are {(A,B,C), (D,E,F,G), and (C,G,H)}, the joint probability function of nodes as a function of evidence potentials becomes:

(5) $$P(V) = P(A,B,C,D,E,F,G,H) = \psi(A,B,C)\,\psi(D,E,F,G)\,\psi(C,G,H)/Z$$

Initially, we can make (see (4))

$$\psi(A,B,C) = P(A,B)\,P(C/A,B)\,,\ \psi(D,E,F,G) = P(D,E,F)\,P(G/D,E,F)\,,$$
$$\psi(C,G,H) = P(H/C,G) \text{ and } Z=1$$

A perfect numbering of nodes is shown in figure 3. From it, the following set chain representation can be obtained (see Table 7):

(6) $$P(V) = P(A,B,C,D,E,F,G,H) = P(A,B,C)\,P(G,H/C)\,P(D,E,F/G)$$

number i	clique C_i	residual R_i	separator S_i
1	ABC	ABC	F
2	CGH	GH	C
3	DEFG	DEF	G

Table 7.- Set chain decomposition

As an example, let us consider the conditional probability tables (we only give the conditional probabilities of the value "true", because those for "false" are their complements to one):

$P(a, b) = 0.90$	$P(c/a, b) = 0.96$	$P(d, e, f) = 0.90$	$P(\bar{d}, e, f) = 0.02$
$P(a, \bar{b}) = 0.05$	$P(c/a, \bar{b}) = 0.04$	$P(d, e, \bar{f}) = 0.02$	$P(\bar{d}, e, \bar{f}) = 0.01$
$P(\bar{a}, b) = 0.04$	$P(c/\bar{a}, b) = 0.02$	$P(d, \bar{e}, f) = 0.02$	$P(\bar{d}, \bar{e}, f) = 0.01$
$P(\bar{a}, \bar{b}) = 0.01$	$P(c/\bar{a}, \bar{b}) = 0.01$	$P(d, \bar{e}, \bar{f}) = 0.01$	$P(\bar{d}, \bar{e}, f) = 0.01$

$P(h/c, g) = 0.98$	$P(g/d, e, f) = 0.98$	$P(g/\bar{d}, e, f) = 0.02$
$P(h/c, \bar{g}) = 0.03$	$P(g/d, e, \bar{f}) = 0.02$	$P(g/\bar{d}, e, \bar{f}) = 0.01$
$P(h/\bar{c}, g) = 0.02$	$P(g/d, \bar{e}, f) = 0.02$	$P(g/\bar{d}, \bar{e}, f) = 0.01$
$P(h/\bar{c}, \bar{g}) = 0.01$	$P(g/d, \bar{e}, \bar{f}) = 0.01$	$P(g/\bar{d}, \bar{e}, \bar{f}) = 0.01$

where "a" means A = true and "$\bar{a}$" , A = false and we use analogous notation for the rest of the nodes.

This allows initial evidence potentials to be obtained as indicated, from which, using the process described after expression (3), terms in (6) can be obtained. Finally, marginal probabilities of cliques or nodes are calculated based on terms in (6). The first factor on the right hand side of (6) gives the marginal probability distribution of the clique {A,B,C}, from which, by marginalization (sum in the adequate set) we obtain the marginal of the nodes A, B and C. Multiplying P(C), which has already been obtained, by P(G,H/C) we obtain the marginal of the clique {G,H,C} and from it the marginal of G and H. Multiplying now P(G) by P(DEF/G) we get the joint probability of D,E, F and G, which allows the

marginal probabilities of D, E and F to be obtained. In this way, the marginal probabilities of nodes shown in figure 3 have been obtained.If now we know that "C = true" (C="c") we get the new evidence potentials

$$\psi^*(A\ B\ c) = \psi(A\ B\ c)\ ;\ \psi^*(A\ B\ \bar{c}) = 0\ ;\ \psi^*(DEFG) = \psi(DEFG)\ ;\ \psi^*(CGH) = \psi(CGH)$$

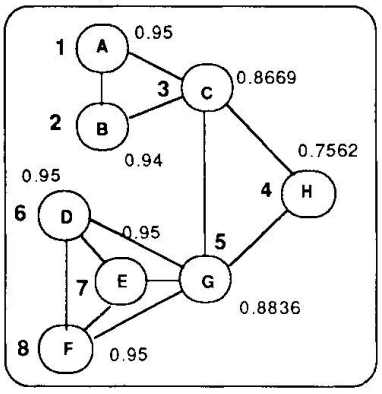

Figure 3.- Initial probabilities of nodes

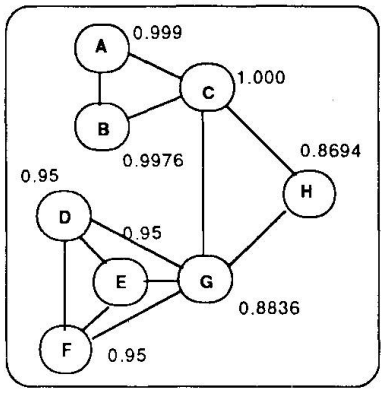

Figure 4.- Updated probabilities when C is true

By a similar process, the new marginal probabilities of nodes, shown in figure 4 have been obtained.

4.- LEARNING

In this section we analyse several techniques for making possible the learning process. We differentiate between parametric and structural learning. The parametric learning refers to the acquisition of parameters in the knowledge base. Whether we work with rules or probabilities, the uncertainty models depend on parameters, which must be known with precision in order to get a reliable expert system. Mechanisms for progressively estimating improved parameters are the basis for the parametric learning subsystem. In order to ilustrate the learning process for probability based models we give the following example.

Example 7.-Let us assume that we are in the case of example 1 and that the engineer knows of the presence of problem E with symptoms S_1, no S_2 and S_3. Then, the updating of parameters (frequencies), that is, parametric learning, consists of adding one unit to the frequency associated to that combination of symptoms, obtaining the value 35 (34+1) (see Figure 1.a). But, what happens if we only know symptoms no S_2 and S_3, but we ignore whether or not S_1 is present. In this situation we do not know if the problem is in the same case as the above 34 or in the case of the 14 problems without S_1 (see Figure 1.a). Thus, we do not know to which of the frequencies the one should be added. The Solomonic solution consists of distributing that unit proportionally to the previously existing frequencies. So value 34 modifies to 34+34/(34+14) and value 14 changes to 14+14/(34+14). In this way we get fractional values, instead of integers, but we update information without any loss of information. With this parametric learning procedure, we can start using the expert system with an imperfect knowledge base and progressively improve its quality with experience. In the case of the log-linear or regression models parametric learning involves a new estimation of parameters, with inclusion of the new

data but without any modification of the model's structure. Any modification in the knowledge base structure leading to some improvement is known as structural learning. Among these, the most well-known variant is the inclusion of new symptoms (additional parameters). Some well known statistical techniques allow the selection, among a set of given parameters, of those which represent knowledge most adequately.

In order to test the adequacy of a probabilistic model relative to one of its extensions (more general models), it is sufficient to estimate, by the maximum likelihood method, parameters of both models and calculate the likelihood ratio. If M_1 and M_2 are two models with r_1 and r_2 parameters, respectively, M_2 being an extension of M_1, we calculate the ratio

$$V = \underset{M_2}{\text{Max}} \prod_{j=1}^{n} P(E_j \cap A_{1j} \cap A_{2j} \cap \ldots \cap A_{mj}) \; / \; \underset{M_1}{\text{Max}} \prod_{j=1}^{n} P(E_j \cap A_{1j} \cap A_{2j} \cap \ldots \cap A_{mj})$$

where n is the sample size (number of data items with known symptoms and associasted problems) and the maximization must be understood with respect to the set of parameters of models M_1 and M_2, respectively, and subject to their respective constraints. The significance level can be calculated by taking into account that the statistic -2logV converges in probability to a $\chi^2 (r_2 - r_1)$.

Structural learning with log-linear or regression models consists of choosing the simplest model reproducing the real frequencies up to acceptable levels of error. Thus, terms to be included in the model must be selected. In order to make this selection we have two procedures:

(1) Start from the saturated model (with the maximum number of parameters) and proceed to eliminate terms until the quality of the model is substantially affected by their removal

(2) Start from a simple model and add new terms until a substantial improvement is no longer obtained.

For stepwise selection of log-linear and regression models several statistical packages can be used as BMDP, SPSS, SAS, etc. Log-linear and regression models in examples 3 and 4 were selected by this method using the BMDP package.

5.- REFERENCES

1. Bishop, Y. M. M., Fienberg, S. A. and Holland, P. W. (1975). Discrete multivariate analysis: Theory and Practice. Cambridge, Mass., The MIT Press.

2. Castillo, E. and Alvarez, E. (1989). Introducción a los sistemas expertos. Aprendizaje e incertidumbre. Editorial Paraninfo. Madrid.

3. Klahr, P. and Waterman, D. A. (1986). Expert system techniques, tools and applications. Addison Wesley Publishing Co.

4. Lauritzen, S.L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society, Series B, 50, N0. 2, 157-224.

5. Luceño, A. (1988). Métodos de Estadística aplicada. Servicio de publicaciones. Universidad de Cantabria.

Leere Seite
Blank page
Page vide