Probabilistic reasoning with uncertain evidence[⋆]

Jiří Vomlel

Laboratory for Intelligent Systems, University of Economics,
Ekonomická 957, 148 01 Praha 4 - Kunratice, Czech Republic

Institute of Information Theory and Automation,
Pod vodárenskou věží 4, 182 08 Praha 8 - Libeň, Czech Republic
vomlel@utia.cas.cz

Abstract. Bayesian networks became a popular framework for reasoning with uncertainty. Efficient methods have been developed for probabilistic reasoning with new evidence. However, when new evidence is uncertain or imprecise different methods have been proposed. The original contribution of this paper are guidelines for the treatment of different types of uncertain evidence, the rules for combining evidence from different sources, and the model revision with uncertain evidence.

Keywords: Probabilistic models, Bayesian networks, belief revision, model revision

1 Introduction

One of rapidly developing areas of Artificial Intelligence is managing uncertainty. It is not surprising that intelligent systems are expected to be able to exploit uncertain or vague information since also humans often have to reason and decide without having precise and certain information. One can distinguish between two basic types of uncertainty: one caused by uncertain or vague information, another by unknown, imprecise, or stochastic relations between variables that are part of a model of a reality. Diverse frameworks were proposed to tackle the challenging problem of reasoning with uncertain and vague information: Dempster-Shafer theory of evidence [3, 11], theory of imprecise probabilities [14], possibility theory [16], fuzzy set theory [15], etc.

In this paper we deal with the standard probabilistic inference. It means that all uncertainties considered in this paper are treated in the sense of randomness and will be quantified and processed by the means and tools of classical probability. For updating we use the standard Bayes rule. The knowledge of a modeled domain is represented by a probability distribution $P(V)$ defined for all combinations of values of variables from a set V. The initial probability model $P(V)$ is usually built using data measured for a population of individuals, repeated events, etc. The built model can be used for individual case analysis -

[⋆] This work was supported by the Grant Agency of the Czech Republic through grant nr. 201/02/1269.

for example in the role of an expert system. In this paper we deal mainly with probability distributions defined by a Bayesian network. Among the first successful applications of Bayesian networks in the role of expert systems in diagnosis belong Munin [8] and the Quick Medical Reference (QMR) system [12]. For an introduction to Bayesian networks we refer to [7].

One of the basic tasks in probabilistic modeling is *belief revision* with new evidence $\boldsymbol{e}$. For example, assume we have created a model describing certain properties of a population of individuals. Than, we perform a number of observations and tests of one individual. Now, in the light of new information, we would like to update our beliefs about the unobserved properties of that individual. Assume we are interested in a variable A. In the probabilistic framework this corresponds to computing conditional probability distribution of variable A given the observed evidence $\boldsymbol{e}$, $P(A \mid \boldsymbol{e})$. The posterior probability is computed using the Bayes rule. For each state a of A we compute

$$P(A = a \mid \boldsymbol{e}) = \frac{P(A = a, \boldsymbol{e})}{P(\boldsymbol{e})} = \frac{P(A = a, \boldsymbol{e})}{\sum_a P(A = a, \boldsymbol{e})} \ . \tag{1}$$

However, in practice, observations or tests yield uncertain results. The question we deal with in this paper is how should we revise our model in the light of uncertain evidence.

Our main motivation is to provide a knowledge engineer appropriate methods for updating a probability model in the light of uncertain evidence that can take different forms. In contrast to different ad hoc approaches we stand firmly within the standard probability framework. A knowledge engineer gets an evidential statement and one or more numbers[1] specifying uncertainty about the provided evidence. The fundamental question is: how should the model be updated and the probabilities revised using the provided numbers. The handling of the provided numbers depends on what they actually mean, i.e., on their semantics.

Two main methods for probabilistic reasoning with uncertain evidence were proposed - the virtual evidence method [9] and Jeffrey's rule [6]. In [10] the fundamental difference between the methods is illustrated using an example. The basic principle underlying Jeffrey's rule is the principle of *probability kinematics*, which can be viewed as a principle that aims at minimizing belief change in the model. In [1] it is shown that also the method of virtual evidence commits to this principle and that the difference between these two methods is in the way uncertain evidence is specified. We will see that Pearl's method is based on sensitivity and specificity of a test (or of an observation) while when using the Jeffrey's rule we specify the resulting effect.

There is a fundamental difference between *belief revision* and *model revision*. The posterior probabilities we get after *belief revision* inform about the properties of one tested individual, one performed event, etc. On the other hand the posterior probabilities we get after *model revision* still correspond to the properties of the tested population of individuals, events, etc. We extend the analysis from belief revision to model revision using a standard method from statistics -

[1] We assume they are real numbers from interval $\langle 0, 1 \rangle$.

the maximum likelihood estimation of probability distributions given observed data.

The main original contribution of this paper are guidelines for the treatment of different types of uncertain evidence, the rules for the combination of evidence from different sources, and the model revision with uncertain evidence.

We begin the paper with a discussion of criteria that can be used to evaluate reliability of information sources (Section 2). We discuss belief revision based on sources' reliability in Section 3 and belief revision based on summary statistics in Section 4. In Section 5 belief revision is applied to a complex probabilistic model. The revision of model parameters is discussed in Section 6. It is generalized to model revision in Section 7. We conclude the paper with general recommendations for dealing with uncertain evidence.

2 Reliability of information sources

For simplicity assume that a partially reliable source T reports about an event A that has two possible outcomes only: *yes* or *no*. The report can be also only *yes* or *no*. Four situations are possible:

- the source reports *yes* when event A actually happened,
- the source reports *no* when event A actually did not happen,
- the source reports *no* when event A actually happened, and
- the source reports *yes* when event A actually did not happen.

To evaluate reliability of an information source we can count how often the four situations discussed above happened. Generally we can create a 2×2 contingency table, see Table 1[2].

Table 1. Values of $n \cdot P(A, T)$

	$A = yes$	$A = no$
$T = yes$	tp	fp
$T = no$	fn	tn

Usually, these statistics are used to compute an evaluation criteria. The most common criteria are defined in Table 2. Sometimes, in different domains the criteria have different names. Note that these criteria can be generalized to variables with more than two outcomes.

Remark 1. The values of sensitivity and specificity need not be based on relative frequencies. Within a subjective probability framework they are subjective beliefs of an expert or a model designer.

[2] tp stands for the number of *true positive* reports, fp for *false positive*, fn for *false negative*, and tn for *true negative*.

Table 2.

accuracy	$P(A = T) = \frac{tp+tn}{tp+tn+fp+fn}$
positive predictive value (precision)	$P(A = yes \mid T = yes) = \frac{tp}{tp+fp}$
negative predictive value	$P(A = no \mid T = no) = \frac{tn}{fn+tn}$
true positive rate (recall or sensitivity)	$P(T = yes \mid A = yes) = \frac{tp}{tp+fn}$
true negative rate (specificity)	$P(T = no \mid A = no) = \frac{tn}{fp+tn}$
false positive rate	$P(T = yes \mid A = no) = \frac{fp}{fp+tn}$
false negative rate	$P(T = no \mid A = yes) = \frac{fn}{tp+fn}$

3 Belief revision based on sources' reliability

In this section we will use *sensitivity* and *specificity* to measure reliability of a test or an information source. Observe that these two criteria (together with their complements) define all values of the conditional probability distribution $P(T|A)$.

Let variable A denote the true state of a patient. It has only two states *yes* (meaning that the patient has an illness A) and *no* (which is just the complement to *yes*). Assume a prior probability of the illness A among the patients that are tested is 0.2. Assume the result of a medical test T_1 is positive ($T_1 = yes$) and both the sensitivity and the specificity of test T_1 is known to be 70%, i.e., $P(T_1 = yes \mid A = yes) = 0.7$ and $P(T_1 = no \mid A = no) = 0.7$. Then we can revise the probability of patient's state using the *Bayes rule*. Thus we compute the posterior probability as

$$P(A = yes \mid T_1 = yes) = c \cdot P(T_1 = yes \mid A = yes) \cdot P(A = yes) \qquad (2)$$

$$P(A = no \mid T_1 = yes) = c \cdot P(T_1 = yes \mid A = no) \cdot P(A = no) \ , \qquad (3)$$

where c is the normalization constant defined so that $P(A = yes \mid T_1 = yes) + P(A = no \mid T_1 = yes) = 1$. In our example

$$P(A = yes \mid T_1 = yes) = c \cdot 0.7 \cdot 0.2 = c \cdot 0.14 = \frac{7}{19} \qquad (4)$$

$$P(A = no \mid T_1 = yes) = c \cdot 0.3 \cdot 0.8 = c \cdot 0.24 = \frac{12}{19} \ . \qquad (5)$$

The method where the reported uncertainty is in the form of

- sensitivity of T_1,
- specificity of T_1, and
- an observed result $T_1 = t_1$

is called *virtual evidence method*. It was introduced by Pearl [9] as a method for belief revision with uncertain evidence.

If we were to decide whether the patient is sick or not we would choose the more probable value, i.e. $A = no$. Since there is quite uncertainty about A we decide to do a second test T_2 that has quite low sensitivity $P(T_1 = yes \mid A = yes) = 0.6$ but relatively high specificity $P(T_1 = no \mid A = no) = 0.9$. Assume the test result is positive again.

The question is how should we combine the results of two tests together. The problem is simplified if we can make the assumption of conditional independence of two tests given the patient's state, i.e. if there are no other interactions between the tests than those given by the presence or the absence of the tested illness. In such a case for all $(t_1, t_2, a) \in \{yes, no\}^3$:

$$\begin{aligned} & P(A = a, T_1 = t_1, T_2 = t_2) \\ & = P(T_1 = t_1, T_2 = t_2 \mid A = a) \cdot P(A = a) \\ & = P(T_1 = t_1 \mid A = a) \cdot P(T_2 = t_2 \mid A = a) \cdot P(A = a) \ . \end{aligned} \tag{6}$$

This relation can be visualized using a Bayesian network with structure given in Figure 1.

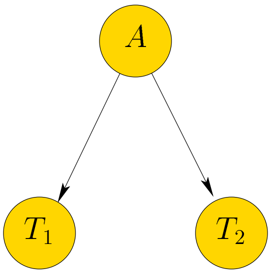

Fig. 1. Two sources report about A.

Using the Bayes rule and the assumption of conditional independence of tests given the patient's state we can compute the posterior probability as

$$\begin{aligned} & P(A = yes \mid T_1 = yes, T_2 = yes) \\ & = c \cdot P(T_1 = yes, T_2 = yes, A = yes) \\ & = c \cdot P(A = yes) \cdot P(T_1 = yes \mid A = yes) \cdot P(T_2 = yes \mid A = yes) \end{aligned} \tag{7}$$

$$\begin{aligned} & P(A = no \mid T_1 = yes, T_2 = yes) \\ & = c \cdot P(T_1 = yes, T_2 = yes, A = no) \\ & = c \cdot P(A = no) \cdot P(T_1 = yes \mid A = no) \cdot P(T_2 = yes \mid A = no) \ , \end{aligned} \tag{8}$$

where c is again the normalization constant. In our example

$$P(A = yes \mid T_1 = yes, T_2 = yes) = c \cdot 0.7 \cdot 0.6 \cdot 0.2 = c \cdot 0.084 = \frac{7}{9} \tag{9}$$

$$P(A = no \mid T_1 = yes, T_2 = yes) = c \cdot 0.3 \cdot 0.1 \cdot 0.8 = c \cdot 0.024 = \frac{2}{9} \tag{10}$$

If we are to decide then we choose $A = yes$ since $P(A = yes \mid T_1 = yes, T_2 = yes) > P(A = no \mid T_1 = yes, T_2 = yes)$. Note that the uncertainty about the states of A has substantially decreased.

4 Belief revision based on summary statistics

If in the example from the previous section all tests have the same sensitivity $P(T_i = yes \mid A = yes)$ and specificity $P(T_i = no \mid A = no)$ then we can use $n(a), a \in \{yes, no\}$ - the number of tests with result a - as a summary statistics. Let $n = \sum_a n(a)$. The revised beliefs in A are (for $a = yes, no$)

$$P(A = a \mid \boldsymbol{t}) = c \cdot P(A = a) \cdot \left(\begin{array}{c} P(T_i = a \mid A = a)^{n(a)} \\ \cdot (1 - P(T_i = a \mid A = a))^{n - n(a)} \end{array} \right) \quad . \tag{11}$$

Now we will regard several experts $E_i, i = 1, \ldots, \ell$ and assume that each expert E_i performed tests $T_i^j, j = 1, \ldots, n_i$ with observed results $\boldsymbol{t}_i = (t_i^1, \ldots, t_i^{n_i})$ and reports $P(E_i = yes \mid \boldsymbol{t}_i)$ what she believes should be the final belief about A after her report is taken into account. Further assume that each expert E_i used the virtual evidence method (discussed in Section 3) to combine the observed results using the sensitivity and the specificity of each test by computing

$$P(E_i = a \mid \boldsymbol{t}_i) \;\; c_i \cdot P(A = a) \cdot \left(\begin{array}{c} \prod_{j \in J_i(a)} P(T_i^j = a \mid A = a) \\ \cdot \prod_{j \in J_i \setminus J_i(a)} 1 - P(T_i^j = a \mid A = a) \end{array} \right) \quad , \tag{12}$$

where $J_i = \{1, \ldots, n_i\}$, $J_i(a) = \{j \in J_i : t_i^j = a\}$, $P(T_i^j = yes \mid A = yes)$ is the sensitivity of T_i^j, $P(T_i^j = no \mid A = no)$ is the specificity of test T_i^j, and c_i is the normalization constant.

Assume that all tests performed by all experts are independent given the state of A. An example of a Bayesian network model corresponding to tests made by two experts E_1 and E_2 is given in Figure 2.

If we want to combine reports from experts $E_1, \ldots, E_\ell$ then, first, we must discard the prior information[3] $P(A = a)$ included ℓ-times and then we can simply multiply the terms altogether with the prior information:

$$P(A = a \mid \boldsymbol{t}) = c \cdot P(A = a) \cdot \prod_{i=1}^{\ell} \frac{P(E_i = a \mid \boldsymbol{t}_i)}{P(A = a)} \quad . \tag{13}$$

5 Belief revision in a complex probabilistic model

Quite often, we are interested not only in one variable, but in a whole set of variables V that are interdependent. Within the probability framework the dependency is represented by a probabilistic model. If all variables are finite-valued

[3] We assume that all experts use the same prior information.

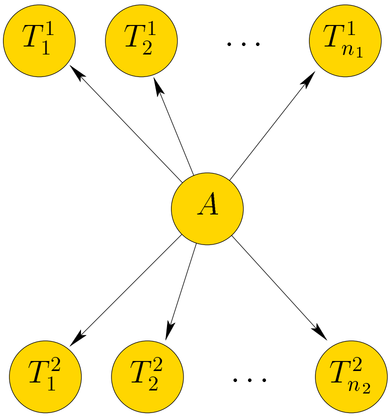

Fig. 2. Two experts report test results.

then the probabilistic model is a discrete probability distribution $P(V)$ that defines probability values for all combinations of states of variables from V.

In such a model a change of beliefs about one variable, say A, has an impact on beliefs about all variables dependent on A. If an expert E_1 reports $P(E_1)$, which she believes should be the probability of A after her report is taken into account, then the formula used for updating the whole model $P(V)$ to $P'(V)$ is called Jeffrey's rule [6]. Let $V' = V \setminus \{A\}$, v' be a combination of values of variables from V', and a a value of A. The Jeffrey's rule is (for all combinations of v', a):

$$P'(V' = v', A = a) = P(V' = v', A = a) \cdot \frac{P(E_1 = a)}{P(A = a)} \quad . \tag{14}$$

The basic principle underlying this rule is the principle of *probability kinematics*, which can be viewed as a principle that aims at minimizing belief change in the model[4]. In [1] it is shown that also the method of virtual evidence commits to this principle and that the difference between these two methods is the way uncertain evidence is specified. In the virtual evidence method we specify sensitivity, specificity, and an observed outcome, while when using the Jeffrey's rule we specify the resulting effect.

We will exploit this correspondence when we have a complicated situation when several experts have different opinion on what the values of $P'(A)$ should be after the revision. We will assume that each expert opinion $P(E_i)$, $i = 1, \dots, \ell$ about A is based on a (possibly only virtual) test and combined with prior probability $P(A)$ as described in Section 4. Then we can follow the approach discussed in Section 4 and proceed using formula 13 in the following way:

[4] Note that it holds that $P'(V' \mid A) = P(V' \mid A)$ while $P'(A) = P(E_1)$.

- discard the influence of the prior probability from the experts opinions, i.e. for $i = 1, \dots, \ell$ and for all states a of A

$$P'(E_i = a) = \frac{P(E_i = a)}{P(A = a)} \quad \text{and} \tag{15}$$

- use the modified distribution to compute for all combinations of v', a

$$P'(V' = v', A = a) \propto P(V' = v', A = a) \cdot \prod_{i=1}^{\ell} P'(E_i = a) \ . \tag{16}$$

This method can be used even if the expert reports are only partially overlapping or not overlapping at all, e.g. when E_1 and E_2 report about two different variables A and B. Now, let $V' = V \setminus \{A, B\}$. Then the formula for the updated model is

$$P'(V' = v', A = a, B = b) \tag{17}$$

$$\propto P(V' = v', A = a, B = b) \cdot \frac{P(E_1 = a) \cdot P(E_2 = b)}{P(A = a) \cdot P(B = b)} \ . \tag{18}$$

If A and B are dependent in the probabilistic model $P(V)$ then it may happen that $P'(A) \neq P(E_1)$ and $P'(B) \neq P(E_2)$. It is all right since, typically, the experts can not take the information from the other experts into account (for example, because the experts report at the same time).

6 Revision of model parameters

Now, assume the variables of interest are parameters of a probabilistic model - for example, sensitivity $R = P(T = yes \mid A = yes)$ of a test T of an illness A and specificity $S = P(T = no \mid A = no)$ of the same test[5]. Note that these variables are continuous with states $r, s \in \langle 0, 1 \rangle$. Now $P(R = r)$ and $P(S = s)$ are probability density functions, which means that $\int_0^1 P(R = r)\ dr = 1$ and $\int_0^1 P(S = s)\ ds = 1$.

Assume two experts independently evaluated one test T - both of them performed several tests on patients with known diagnosis (on both - sick and non-sick patients) with the observed results being $\boldsymbol{t}_i$ for expert i. They report their estimates of r and s computed as relative frequencies ($i = 1, 2$):

$$\hat{r}_i = \hat{P}(T = yes \mid A = yes, \boldsymbol{t}_i) = \frac{n_i(T = yes, A = yes)}{n_i(A = yes)} \ , \tag{19}$$

$$\hat{s}_i = \hat{P}(T = no \mid A = no, \boldsymbol{t}_i) = \frac{n_i(T = no, T = no)}{n_i(A = no)} \ , \tag{20}$$

where $n_i(t, a)$ denotes the number of occurrence of $T = t, A = a$ observed by expert E_i, $n_i(A = a) = \sum_t n_i(T = t, A = a)$, and $n_i = \sum_a n_i(A = a)$.

[5] We assume that the test properties are stable.

Again assume that results of performed tests are independent given the illness A. Experimental data $\boldsymbol{t} = \{\boldsymbol{t}_1, \boldsymbol{t}_2\}$ satisfying the assumptions given above are called identically independently distributed (i.i.d.) data. The posterior probability

$$P(R = r \mid \boldsymbol{t}) = c_r \cdot P(R = r) \cdot \prod_{i=1}^{2} r^{n_i(T=yes,A=yes)} \cdot (1-r)^{n_i(T=no,A=yes)} \tag{21}$$

$$P(S = s \mid \boldsymbol{t}) = c_s \cdot P(S = s) \cdot \prod_{i=1}^{2} r^{n_i(T=no,A=no)} \cdot (1-r)^{n_i(T=yes,A=no)} \quad . \tag{22}$$

A task is to estimate most probable values $\hat{r}, \hat{s}$ of variables R and S given the data $\boldsymbol{t}$, i.e. to find

$$\hat{r} = \arg\max_r P(R = r \mid \boldsymbol{t}) \tag{23}$$

$$\hat{s} = \arg\max_s P(S = s \mid \boldsymbol{t}) \tag{24}$$

These parameters are then called *maximum likelihood estimates* of r and s.

In our example after little algebra[6] we get that

$$\begin{aligned} \hat{r} &= \frac{n_1(A = yes)}{n_1(A = yes) + n_2(A = yes)} \cdot \hat{r}_1 + \frac{n_2(A = yes)}{n_1(A = yes) + n_2(A = yes)} \cdot \hat{r}_2 \\ &= w_1(yes) \cdot \hat{r}_1 + (1 - w_1(yes)) \cdot \hat{r}_2 \quad . \end{aligned} \tag{25}$$

Thus, we combined expert information using a weighted arithmetic average. Similarly, we can get the formula for the *maximum likelihood estimate* of s.

Second order probabilities

Observe that there would be no difference if both experts performed ten or ten thousand experiments and got the same values of $\hat{r}_1$ and $\hat{r}_2$. Intuitively, results based on more experiments should be more reliable. In order to be able to represent uncertainty about the values of R we would need (instead of computing only the most likely value $\hat{r}$) to update a prior distribution $P(R)$ using the Bayes rule. This is the basic idea of Bayesian statistics [5], where probability distributions of model parameters are used instead of their single values. It is convenient to assume that the prior distribution has the form

$$P(R = r) = c_0 \cdot r^{n_0(T=yes,A=yes)} \cdot (1-r)^{n_0(T=no,A=yes)} \quad . \tag{26}$$

Observe that if $n_0(T = yes, A = yes) = n_0(T = no, A = yes) = 0$ we have a uniform distribution. In Figure 3 the posterior distributions for $n_1(A = yes) = n_2(A = yes) = 2$ and for $n_1(A = yes) = n_2(A = yes) = 10$ are displayed. In both cases $\hat{r}_1 = 0.7$ and $\hat{r}_2 = 0.8$ and the prior distribution is uniform. Observe that in both cases $\hat{r} = 0.75$ but the probability mass is distributed differently. We can see that the probability mass is concentrated more around the value 0.75 in the latter case.

[6] For simplicity we assume a uniform priors $P(R = r)$ and $P(S = s)$.

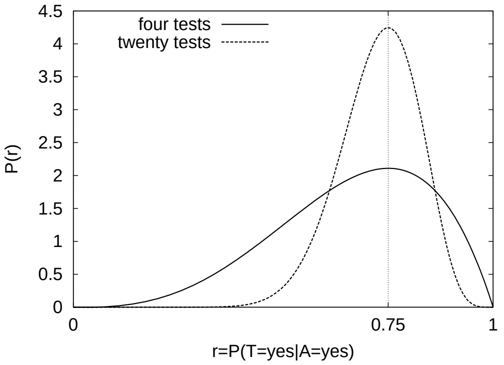

Fig. 3. Posterior probabilities for the sensitivity R of a test T.

Comparisons with belief revision

Assume again a simple example where two experts performed several times a test T of an illness A (both variables have only two states *yes* and *no*).

Next we will show that it is important to distinguish whether we use expert reports to update our belief about a variable (Section 4) or whether we want to revise a parameter of our model (this section).

In the former case each expert E_i perform tests on the same individual and report the number of positive results $n_i(T = yes)$ and the number of negative results $n_i(T = no)$. We define $n_i = n_i(T = yes) + n_i(T = no)$. We use the known sensitivity and specificity $P(T = a \mid A = a), a \in \{yes, no\}$ to combine information from two experts using formula 13. In case of uniform $P(A)$ the formula reduces (for $a \in \{yes, no\}$) to

$$\begin{aligned}P(A = a \mid \boldsymbol{t}) \propto\ & P(T = a \mid A = a)^{n_1(T=a)} \cdot (1 - P(T = a \mid A = a))^{(n_1 - n_1(T=a))} \\ & \cdot P(T = a \mid A = a)^{n_2(T=a)} \cdot (1 - P(T = a \mid A = a))^{(n_2 - n_2(T=a))} \\ \propto\ & P(E_1 = a \mid \boldsymbol{t_1}) \cdot P(E_2 = a \mid \boldsymbol{t_2}) \ ,\end{aligned}$$

which means we combine the information using multiplication. Note that in this case $P(A = a \mid \boldsymbol{t})$ provides information about the tested individual.

In the latter case each expert E_i perform test T on several different individuals. Each expert counts the number $n_i(T = yes, A = yes)$ of positive results for the sick individuals, the number $n_i(T = no, A = no)$ of negative results for the non-sick individuals, and total number of tested sick $n_i(A = yes)$ and non-sick $n_i(A = no)$ individuals. This information is used to estimate $\hat{P}(T = a \mid A = a, \boldsymbol{t}_i), i \in \{1, 2\}, a \in \{yes, no\}$ using formulas 19 and 20. If the priors $P(T \mid A)$ are uniform then the maximum likelihood estimates of sensitivity (formula 25) and specificity are for $a \in \{yes, no\}$

$$\hat{P}(T = a \mid A = a, \boldsymbol{t}) \quad \propto \quad \begin{aligned} & w_1(a) \cdot \hat{P}(T = a \mid A = a, \boldsymbol{t}_1) \\ & + w_2(a) \cdot \hat{P}(T = a \mid A = a, \boldsymbol{t}_2) \ , \end{aligned} \qquad (27)$$

which means we combine the information using addition. Note that in this case $\hat{P}(T = a \mid A = a, \boldsymbol{t})$ is the parameter estimate for the whole population of tested individuals.

The difference in handling the information provided by experts is theoretically well founded. In the area of artificial intelligence different methods for fusing expert information are studied, in some of the proposals only intuitive reasons for the use of a kind of average are given.

7 Model revision

We will use a simple example to illustrate the method that can be used to revise original model $P_0(V)$, where $V = \{A, B, C\}$ and A, B, and C are three variables whose relations we would like to model in a model $P(V)$. For example A is *age*, B *religion*, and C is *political orientation*. We conduct a small survey and get data about n_0 individuals. We use the data to create the original model

$$P_0(a, b, c) = \frac{n_0(a, b, c)}{n_0} \quad .$$

Since the population that participated in our survey was only a small subsample of the whole population we would like to revise our model so that it fits the whole population. The information available about the whole population is provided by a census bureau that have made a census or a opinion poll, each of them containing only a two-dimensional subset of $\{A, B, C\}$. For example, a census provided information about relation between *age* and *religion*, an opinion poll provided information about *age* and *political orientation*, and another opinion poll about *religion* and *political orientation*. Only the small survey we conducted provided information about all three variables at the same time. The data patterns are illustrated in Figure 4.

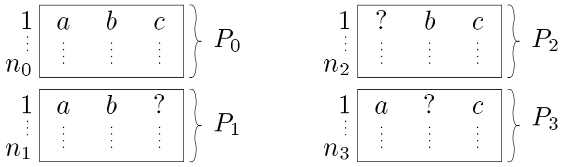

Fig. 4. Data patterns

We can represent the data from the census bureau by probability distributions $P_1(A, B)$, $P_2(B, C)$, and $P_3(A, C)$ computed as:

$$P_1(A, B) = \frac{n_1(a, b)}{n_1}, \quad P_2(B, C) = \frac{n_2(b, c)}{n_2}, \quad P_3(A, C) = \frac{n_3(a, c)}{n_3} \quad .$$

To each probability distribution $P_i, i = 0, 1, 2, 3$ we assign a weight w_i, defined as $w_i = \frac{n_i}{n}$, where $n = n_0 + n_1 + n_2 + n_3$.

Now the task is to find a probability distribution P that is the maximum likelihood estimate given all available data. Under the multinomial model the likelihood of a probability distribution P given data D is

$$\begin{aligned} L(P \mid D) &= \\ \prod_{a,b,c} &\left(P(a,b,c)^{n_0(a,b,c)} \cdot P(a,b)^{n_1(a,b)} \cdot P(b,c)^{n_2(b,c)} \cdot P(a,c)^{n_3(a,c)} \right) \quad , \end{aligned} \tag{28}$$

where $P(a,b)$, $P(b,c)$, and $P(a,c)$ are marginal distributions of $P(a,b,c)$.

It is a consequence of a result proven by Sundberg [13] that for a distribution P that maximizes the likelihood it holds that

$$P(a,b,c) = \begin{array}{l} w_0 \cdot P_0(a,b,c) \cdot \frac{P(a,b,c)}{P(a,b,c)} + w_1 \cdot P_1(a,b) \cdot \frac{P(a,b,c)}{P(a,b)} \\ +w_2 \cdot P_2(b,c) \cdot \frac{P(a,b,c)}{P(b,c)} + w_3 \cdot P_3(a,c) \cdot \frac{P(a,b,c)}{P(a,c)} \end{array} \tag{29}$$

$$= \begin{array}{l} w_0 \cdot P_0(a,b,c) + w_1 \cdot P_1(a,b) \cdot \frac{P(a,b,c)}{P(a,b)} \\ +w_2 \cdot P_2(b,c) \cdot \frac{P(a,b,c)}{P(b,c)} + w_3 \cdot P_3(a,c) \cdot \frac{P(a,b,c)}{P(a,c)} \end{array} \quad . \tag{30}$$

This formula can be used to define an iterative procedure that converges to a distribution satisfying the necessary condition for the probability distribution maximizing the likelihood. The iterative procedure starts with the original distribution P_0 computed from a complete sample, if it is available, otherwise, the uniform probability distribution is often used as the starting point. In every step i of the procedure we use the probability distribution $P^{(i-1)}$ from the previous step $i-1$ to compute new probability distribution $P^{(i)}$:

$$P^{(i)}(a,b,c) = \begin{array}{l} w_0 \cdot P_0(a,b,c) + w_1 \cdot P_1(a,b) \cdot \frac{P^{(i-1)}(a,b,c)}{P^{(i-1)}(a,b)} \\ +w_2 \cdot P_2(b,c) \cdot \frac{P^{(i-1)}(a,b,c)}{P^{(i-1)}(b,c)} + w_3 \cdot P_3(a,c) \cdot \frac{P^{(i-1)}(a,b,c)}{P^{(i-1)}(a,c)} \end{array} \quad . \tag{31}$$

This procedure is a special case of the *EM-algorithm* [4].

Similarly as in Section 6, instead of computing most likely values of model parameters we could use a posterior probability distribution over the space of all possible model parameters. We will not go into details here. For more information see [2, Chapter 9].

8 Conclusions

In this paper we discussed belief and model revision with uncertain evidence. We conclude by summarizing the lessons we have learned.

First, when dealing with an uncertain evidence we should clarify what the information sources actually report about: is it their reliability and an observed value or is it what they believe should be the final value of a variable of interest. In the first case we should use the *virtual evidence method*, while in the second case we should revise the beliefs using the *Jeffrey's rule.*

Second, we must make a clear distinction whether the information sources provide their beliefs about an individual or about a parameter of a general model, which typically is the value of an entry in a conditional probability distribution. The first case corresponds to *belief revision* while the second to *model revision.*

Third, it is important to realize whether we are interested in the most likely value of a variable or a model parameter or whether we prefer to know the probability distribution over their values. This difference is reflected by the methods we use: *maximum likelihood estimation* or *Bayesian statistics.*

Finally, in Section 7 we gave an example of how we should revise a model when new information is provided in the form of probability distributions defined on different subsets of variables from the model.

Acknowledgments

I am grateful to Marco Valtorta and Finn V. Jensen for interesting discussions on several issues of this paper. This paper also benefited from suggestions for improvements proposed by anonymous reviewers of conference Znalosti 2004.

References

1. H. Chan and A. Darwiche. Revisiting the problem of belief revision with uncertain evidence. In *AAAI Spring Symposium*, 2003.
2. R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer Verlag, New York, 1999.
3. A. P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30:205–247, 1968.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–38, 1977.
5. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition.* Chapman and Hall, 2003. Second Edition.
6. R. C. Jeffrey. *The Logic of Decision.* McGraw-Hill, New York, 1965.
7. F. V. Jensen. *Bayesian Networks and Decision Graphs.* Springer Verlag, New York, 2001.
8. K. G. Olesen, U. Kjærulff, F. Jensen, F. V. Jensen, B. Falck, S. Andreassen, and S. K. Andersen. A MUNIN network for the median nerve — a case study on loops. *Applied Artificial Intelligence*, 3:384–403, 1989. Special issue: Towards Causal AI Models in Practice.
9. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman, San Mateo, CA, 1988.
10. J. Pearl. On two pseudo-paradoxes in Bayesian analysis. *Annals of Mathematics and Artificial Intelligence*, 32:171–177, 2001.
11. G. Shafer. *A Mathematical Theory of Evidence.* Princenton University Press, Princeton, New Jersey, 1976.
12. M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

13. R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, 1:49–58, 1974.
14. P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.
15. L. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
16. L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.