KYBERNETIKA — VOLUME 31 (1995), NUMBER 5, PAGES 465 – 480

INTEGRATION OF EXPERT KNOWLEDGE INTO A PROBABILISTIC EXPERT SYSTEM

JANA VEJVALKOVÁ

The paper deals with the problem of integration of additional expert information in the form of univariate marginal distributions into a probabilistic knowledge base defined by a discrete distribution mixture. The suggested solution consists in constructing the I-projection of the original knowledge base on the class of distributions satisfying the additional conditions formulated by experts. The computation of the I-projection is based on the iterative proportional fitting procedure (IPFP) originally designed for contingency tables. The procedure is modified for distribution mixtures with product components and the convergence of the resulting algorithm is proved. Practical application of the method is illustrated by a numerical example.

1. INTRODUCTION

The purpose of expert systems is to enable us the efficient use of knowledge and experience accumulated in different fields of human activities [4]. In practice a great deal of information obtained from experts as well as from the data provided by users is not known with certainty. An important feature of expert systems is therefore the processing of uncertain information which can be well formalized in the framework of probability theory.

Considering the probabilistic approach to expert systems we assume that the input and output information is expressed in terms of discrete random variables

$$v_1, v_2, \ldots, v_N \tag{1}$$

taking values from finite sets $X_1, X_2, \ldots, X_N$, respectively. The uncertainty of the variable v_n, $n \in \{1, \ldots, N\}$ is characterized by a univariate probability distribution on X_n:

$$\text{Prob}\,\{v_n = x\} = p_n(x) \geq 0,\ x \in X_n, \quad \sum_{x \in X_n} p_n(x) = 1. \tag{2}$$

The knowledge base of a common probabilistic expert system is usually closely related to the joint distribution of the involved variables. The knowledge base of the probabilistic expert system PES [3] has the form of a finite distribution mixture

(weighted sum) with M product components:

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} w_m F(\boldsymbol{x}|m), \quad F(\boldsymbol{x}|m) = \prod_{n=1}^{N} p_n(x_n|m),$$

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N), \quad \boldsymbol{x} \in \boldsymbol{X}; \ \boldsymbol{X} = X_1 \times X_2 \times \ldots \times X_N, \tag{3}$$

$$w_m \geq 0, \ \sum_{m=1}^{M} w_m = 1, \ \sum_{x \in X_n} p_n(x|m) = 1, \ n \in \{1, \ldots, N\}, \ m \in \{1, \ldots, M\}.$$

Here $w_m \geq 0$ denotes the apriori weight of the mth component $F(\boldsymbol{x}|m)$ and $p_n(x|m)$ is the discrete distribution of the variable v_n corresponding to the mth component.

An important advantage of mixture (3) is a simple computation of any marginal distribution by omitting superfluous terms in the products $F(\boldsymbol{x}|m)$. In addition, the required form of the mixture is not restrictive since any multivariate discrete distribution can be expressed in this form if the number of components M is sufficiently large.

The output information of the probabilistic expert system PES is expressed either by conditional probability distributions which can be obtained according to Bayes formula in case of definite input or by the formula of complete probability if the input information is uncertain.

One way to obtain the knowledge base is to compute the maximum likelihood estimates of finite mixtures from data using the iterative EM algorithm [2].

Another possibility is to design the knowledge base in cooperation with experts. In practice the components of the mixture defined as products of univariate distributions may correspond to different mutually exclusive situations, diagnoses, hypotheses, etc. and can be directly designed by experts. Unfortunately, the underlying assumption of conditional independence of variables for each diagnosis or situation is rather restrictive.

In the following the problem of the integration of expert information into the probabilistic knowledge base is formulated for a general type of finite distribution mixtures.

2. PROBLEM OF INTEGRATION OF EXPERT INFORMATION INTO THE KNOWLEDGE BASE

Let $\mathcal{P}$ be the set of all discrete probability distributions Q on the product space $\boldsymbol{X}$:

$$Q: \ \boldsymbol{X} \to \langle 0, 1\rangle, \quad \sum_{\boldsymbol{x} \in \boldsymbol{X}} Q(\boldsymbol{x}) = 1; \quad \boldsymbol{X} = X_1 \times X_2 \times \ldots \times X_N, \tag{4}$$

where $X_n, n \in \{1, \ldots, N\}$ are finite sets. On the set $\mathcal{P}$ we introduce the metric:

$$\rho(P, Q) = \|P - Q\| = \sum_{\boldsymbol{x} \in \boldsymbol{X}} |P(\boldsymbol{x}) - Q(\boldsymbol{x})|; \quad P, Q \in \mathcal{P}. \tag{5}$$

Theorem 1. Metric space $(\mathcal{P}, \rho)$ with the metric (5) is compact.

Proof. It is a well known fact. □

We denote as $\mathcal{S}_M \subset \mathcal{P}$ the set of discrete distributions on $\boldsymbol{X}$ having the form of a finite mixture (3) with M product components.

We shall assume that the original knowledge base is defined by a joint probability distribution $R \in \mathcal{S}_M$ – designed by experts or estimated from data – and that the experts cooperating in designing the knowledge base supply some additional information in the form of true univariate distributions for all variables v_n, $n \in \{1, \ldots, N\}$:

$$\text{Prob}\,\{v_n = x\} = q_n(x) \geq 0, \quad x \in X_n, \quad \sum_{x \in X_n} q_n(x) = 1. \tag{6}$$

If some distributions q_n are not given explicitly, we could reduce the problem to the corresponding subspace. Another possibility is to complete the expert information with the respective marginal distributions of the original distribution R. In that case we would prefer to keep the unspecified marginal structure of the original knowledge base unchanged.

In order to integrate the given additional information into the original knowledge base $R \in \mathcal{S}_M$ we seek a new knowledge base $R^\star \in \mathcal{S}_M$ satisfying prescribed marginal constraints (6) and differing from the original mixture R as little as possible.

To simplify notation we denote as $\mathcal{E}_n$, $n \in \{1, \ldots, N\}$ the set of all discrete distributions on $\boldsymbol{X}$ which satisfy the nth marginal constraint from (6):

$$\mathcal{E}_n = \{P \in \mathcal{P} : P_n(x) = q_n(x) \text{ for all } x \in X_n\}, \tag{7}$$

where

$$P_n(x) = \sum_{x_1 \in X_1} \cdots \sum_{x_{n-1} \in X_{n-1}} \sum_{x_{n+1} \in X_{n+1}} \cdots \sum_{x_N \in X_N} P(x_1, \ldots, x_{n-1}, x, x_{n+1}, \ldots, x_N) \tag{8}$$

and as $\mathcal{E}$ the set of discrete distributions on $\boldsymbol{X}$ which satisfy all N marginal constraints:

$$\mathcal{E} = \{P \in \mathcal{P} : P_n(x) = q_n(x) \text{ for all } x \in X_n,\ n = 1, \ldots, N\} = \bigcap_{n=1}^{N} \mathcal{E}_n. \tag{9}$$

Remark 1. The sets $\mathcal{E}, \mathcal{E}_1, \ldots, \mathcal{E}_N \subset \mathcal{P}$ are nonempty, convex and closed in $(\mathcal{P}, \rho)$.

As the measure of difference between two distributions we choose the I-divergence used by Csiszár in [1].

Definition 1. The *I-divergence* $I(P, Q)$ of the distributions P, $Q \in \mathcal{P}$ (also called the discrimination information or the relative entropy of P with respect to Q) is defined by the formula:

$$I(P, Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}. \tag{10}$$

Here and in the following we understand

$$0 \log \frac{0}{a} = 0 \text{ for } a \geq 0, \quad a \log \frac{a}{0} = +\infty \text{ for } a > 0. \tag{11}$$

Remark 2. Let us recall that the I-divergence of two probability distributions is always nonnegative and vanishes iff the distributions are identical. It is not symmetrical and does not satisfy the triangular inequality, therefore it is not a metric. Nevertheless, owing to suitable properties, it is often used as a measure of difference of two distributions.

Definition 2. Let $\mathcal{A} \subset \mathcal{P}$ be a nonempty closed convex set and $Q \in \mathcal{P}$ be such that $I(\tilde{P}, Q) < \infty$ holds for some $\tilde{P} \in \mathcal{A}$. Then the distribution $Q^* \in \mathcal{A}$ satisfying the condition:

$$I(Q^*, Q) = \min_{P \in \mathcal{A}} I(P, Q) \tag{12}$$

is called the *I-projection* of the distribution Q on the set $\mathcal{A}$.

Remark 3. Since X is finite the following equivalence is true:

$$I(P, Q) < \infty \Leftrightarrow (Q(x) = 0 \Rightarrow P(x) = 0 \text{ for all } x \in X),$$

i.e. the finiteness of $I(P, Q)$ is equivalent to the absolute continuity of P with respect to Q, $P \ll Q$.

Remark 4. It can be shown that the I-projection Q^* is determined uniquely since $\mathcal{A}$ is a convex set and $I(P, Q)$ is strictly convex function of the variable P.

A necessary and sufficient condition for the existence of the I-projection is formulated in the following theorem.

Theorem 2. Let $R \in \mathcal{P}$ be a distribution such that $I(\tilde{Q}, R) < \infty$ (i.e. $\tilde{Q} \ll R$) holds for some $\tilde{Q} \in \mathcal{E}$. Then R has a unique I-projection $R^* \in \mathcal{E}$ on the set $\mathcal{E}$.

Proof. The proof is analogous to that of Theorem 2.1 in [1]. The uniqueness of the I-projection follows from Remark 4. □

Now we can briefly summarize the problem under consideration and the idea of its solution:

The original knowledge base $R \in \mathcal{S}_M$ and some additional expert information in the form of a set of univariate distributions q_n, $n \in \{1, \ldots, N\}$ are given. In order to integrate this expert knowledge into our expert system we shall construct a new knowledge base $R^\star \in \mathcal{S}_M$ as the I-projection of the distribution R on the set $\mathcal{E}$. If $R^\star$ exists it evidently satisfies marginal constraints (6); it is determined uniquely (see Theorem 2) and minimizes the distance from R in the sense of Definition 2.

The main idea of the present approach is to compute the I-projection $R^\star$ by means of the so-called iterative proportional fitting procedure (IPFP). This procedure had been originally designed for contingency tables but it can be modified for the considered special class of mixtures.

As it will be shown later the constructed iterative sequence of distributions converges to the desired solution $R^\star \in \mathcal{E}$ in the sense of I-divergence and pointwisely, too.

3. IPFP PROCEDURE

The iterative proportional fitting procedure (IPFP) was originally designed to adjust the relative frequencies in a contingency table to some apriori known marginal probabilities. The procedure is based on cyclic norming of the rows and columns of a contingency table until the convergence of the entries. In this paper the IPFP procedure is applied to distribution mixtures of product components (3).

At first we shall describe the construction of the iterative sequence in general case when the starting term $R \in \mathcal{P}$ has not necessarily the form of distribution mixture (3). The modification for distribution mixtures will be mentioned in Section 6.

Let $R \in \mathcal{P}$. The iterative sequence produced by the IPFP procedure we denote as

$$P^{(k,\ell)}, \quad k = 0, 1, \ldots \; ; \quad \ell = 0, 1, \ldots, N, \tag{13}$$

where we set $P^{(0,0)} = R$ and $P^{(k,N)} \equiv P^{(k+1,0)}$ for all k.

If $P^{(k,\ell-1)}$ is the term in the step $(k, \ell - 1)$ then, analogously to [5], the next iteration will be obtained by the recurrent formula:

$$P^{(k,\ell)}(\boldsymbol{x}) = C^{(k,\ell)}(x_\ell)\; P^{(k,\ell-1)}(\boldsymbol{x}), \quad \boldsymbol{x} \in \boldsymbol{X}, \tag{14}$$

where

$$\begin{aligned} C^{(k,\ell)}(x_\ell) &= \frac{q_\ell(x_\ell)}{P_\ell^{(k,\ell-1)}(x_\ell)} \quad \text{if} \quad P_\ell^{(k,\ell-1)}(x_\ell) \neq 0, \\ C^{(k,\ell)}(x_\ell) &= 1 \qquad\qquad\quad \text{if} \quad P_\ell^{(k,\ell-1)}(x_\ell) = 0. \end{aligned} \tag{15}$$

Here q_ℓ is the given univariate distribution and $P_\ell^{(k,\ell-1)}$ is the ℓth marginal of $P^{(k,\ell-1)}$. The iterative sequence defined by formulae (13), (14), (15) we denote briefly as $\left\{ \left\{ P^{(k,\ell)} \right\}_{\ell=0}^{N} \right\}_{k=0}^{\infty}$ and prove its important properties.

In the following we suppose that the assumptions of Theorem 2 are satisfied so that $R \in \mathcal{P}$ has the unique I-projection $R^\star \in \mathcal{E}$ on the set $\mathcal{E}$.

Lemma 1. Let $Q \in \mathcal{E}$, $I(Q, R) < \infty$. Then for each iteration $P^{(k,\ell)}$ the following implication holds:

$$P^{(k,\ell)}(\boldsymbol{x}) = 0 \;\Rightarrow\; Q(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \boldsymbol{X}. \tag{16}$$

Proof. For the first iteration $R = P^{(0,0)}$ this assertion is true because $Q \ll R$. Let us take $\boldsymbol{x} \in \boldsymbol{X}$. If $P^{(k,\ell)}(\boldsymbol{x}) = 0$ then, according to formula (14), either $P^{(k,\ell-1)}(\boldsymbol{x}) = 0$ or $P^{(k,\ell-1)}(\boldsymbol{x}) \neq 0$ and $C^{(k,\ell)}(x_\ell) = 0$.

In the case of $P^{(k,\ell-1)}(\boldsymbol{x}) = 0$ the induction hypothesis implies that $Q(\boldsymbol{x}) = 0$ immediately. In the opposite case, $P^{(k,\ell-1)}(\boldsymbol{x}) \neq 0$ and $C^{(k,\ell)}(x_\ell) = 0$ imply that $q_\ell(x_\ell) = 0$ (see (15)). Since $Q \in \mathcal{E}$ it holds:

$$0 = q_\ell(x_\ell) = Q_\ell(x_\ell) \geq Q(\boldsymbol{x}) \geq 0. \tag{17}$$

□

Proposition 1. $P^{(k,\ell)} \in \mathcal{P}$ for all k, ℓ.

Proof. Let us suppose by induction that $P^{(k,\ell-1)} \in \mathcal{P}$. Then the next iteration $P^{(k,\ell)}$ (see (14)) is evidently nonnegative and

$$\sum_{\boldsymbol{x} \in \boldsymbol{X}} P^{(k,\ell)}(\boldsymbol{x}) = \sum_{x \in X_\ell} C^{(k,\ell)}(x)\; P_\ell^{(k,\ell-1)}(x) = \sum_{x \in X_\ell} q_\ell(x) = 1 \tag{18}$$

because according to Lemma 1

$$P_\ell^{(k,\ell-1)}(x) = 0 \quad \Rightarrow \quad q_\ell(x) = 0, \quad x \in X_\ell. \tag{19}$$

□

Proposition 2. If $Q \in \mathcal{E}$ then

1. $I(Q, R) < \infty \;\Rightarrow\; I(Q, P^{(k,\ell)}) < \infty$ for all k, ℓ
2. $I(Q, R) = \infty \;\Rightarrow\; I(Q, P^{(k,\ell)}) = \infty$ for all k, ℓ.

Proof. The first assertion follows from Lemma 1 and Proposition 1. To prove the second one we shall show that the following implication holds:

$$I(Q, P^{(k,\ell-1)}) = \infty \quad \Rightarrow \quad I(Q, P^{(k,\ell)}) = \infty. \tag{20}$$

$I(Q, P^{(k,\ell-1)}) = \infty$ implies that $Q(\tilde{\boldsymbol{x}}) > 0$ along with $P^{(k,\ell-1)}(\tilde{\boldsymbol{x}}) = 0$ for some $\tilde{\boldsymbol{x}} \in \boldsymbol{X}$. However, according to recurrent formula (14)

$$P^{(k,\ell)}(\tilde{\boldsymbol{x}}) = C^{(k,\ell)}(\tilde{x}_\ell)\; P^{(k,\ell-1)}(\tilde{\boldsymbol{x}}) = 0 \tag{21}$$

so that $I(Q, P^{(k,\ell)}) = \infty$, too. □

Proposition 3. $P^{(k,\ell)} \in \mathcal{E}_\ell$ for all k and $\ell \in \{1, \ldots, N\}$.

Proof. According to iterative formula (14), (15) and implication (19) the following relation can be proved easily:

$$P_\ell^{(k,\ell)}(x) = C^{(k,\ell)}(x)\ P_\ell^{(k,\ell-1)}(x) = q_\ell(x) \text{ for all } x \in X_\ell. \tag{22}$$

□

Proposition 4. If $Q \in \mathcal{E}$ then

$$I(Q, P^{(k,\ell-1)}) = I(Q, P^{(k,\ell)}) + I(P^{(k,\ell)}, P^{(k,\ell-1)}) \text{ for all } k \in \{0, 1, \ldots\}. \tag{23}$$

Proof. First we consider such distributions $Q \in \mathcal{E}$ for which $I(Q, R) < \infty$, i.e. $Q \ll R$. Then $I(Q, P^{(k,\ell-1)}) < \infty$ and $I(Q, P^{(k,\ell)}) < \infty$ (see Proposition 2) and according to iterative formula (14) we can write

$$I(Q, P^{(k,\ell)}) = I(Q, P^{(k,\ell-1)}) - \sum_{x \in X_\ell} Q_\ell(x)\ \log C^{(k,\ell)}(x). \tag{24}$$

Distribution Q satisfies the ℓth marginal constraint so that using formula (15), Proposition 3 and formerly proved implication (19) it can be easily shown that

$$\sum_{x \in X_\ell} Q_\ell(x)\ \log C^{(k,\ell)}(x) = \sum_{x \in X_\ell} q_\ell(x)\ \log C^{(k,\ell)}(x) = I(P^{(k,\ell)}, P^{(k,\ell-1)}) < \infty. \tag{25}$$

In the case of $Q \in \mathcal{E}$, $I(Q, R) = \infty$, both $I(Q, P^{(k,\ell-1)})$ and $I(Q, P^{(k,\ell)})$ are infinite (see Proposition 2) while $I(P^{(k,\ell)}, P^{(k,\ell-1)})$ is finite. Therefore relation (23) is true, too. □

4. TRANSITIVITY OF THE I-PROJECTION

Now we can prove the transitive property of the I-projection.

Theorem 3. Let $R^\star \in \mathcal{E}$ be the I-projection of the distribution $R = P^{(0,0)} \in \mathcal{P}$ on the set $\mathcal{E}$.

Then $R^\star$ is the I-projection of each iteration $P^{(k,\ell)}$ on the set $\mathcal{E}$, i.e. for all $k = 0, 1, \ldots$ and $\ell = 0, 1, \ldots, N$

$$\infty > I(R^\star, P^{(k,\ell)}) = \min_{Q \in \mathcal{E}} I(Q, P^{(k,\ell)}). \tag{26}$$

Proof. Let us suppose by induction that

$$\infty > I(R^\star, P^{(k,\ell-1)}) = \min_{Q \in \mathcal{E}} I(Q, P^{(k,\ell-1)}). \tag{27}$$

For each $Q \in \mathcal{E}$ we can write equation (23) (see Proposition 4) and after minimizing:

$$\min_{Q\in\mathcal{E}} I(Q, P^{(k,\ell-1)}) = \min_{Q\in\mathcal{E}} I(Q, P^{(k,\ell)}) + I(P^{(k,\ell)}, P^{(k,\ell-1)}). \tag{28}$$

Using the induction hypothesis we have

$$\infty > I(R^\star, P^{(k,\ell-1)}) = \min_{Q\in\mathcal{E}} I(Q, P^{(k,\ell)}) + I(P^{(k,\ell)}, P^{(k,\ell-1)}) \tag{29}$$

which, according to Proposition 4 again, implies that

$$\infty > I(R^\star, P^{(k,\ell)}) = \min_{Q\in\mathcal{E}} I(Q, P^{(k,\ell)}). \tag{30}$$

□

5. CONVERGENCE PROPERTIES

In this section it will be proved that the iterative sequence $\left\{\left\{P^{(k,\ell)}\right\}_{\ell=0}^{N}\right\}_{k=0}^{\infty}$ converges in the sense of the I-divergence to the I-projection of the original knowledge base $R = P^{(0,0)} \in \mathcal{P}$ on the set $\mathcal{E}$. Since the product space $\boldsymbol{X}$ is finite the convergence is pointwise, too.

Theorem 4. The iterative sequence

$$\left\{\left\{P^{(k,\ell)}\right\}_{\ell=0}^{N}\right\}_{k=0}^{\infty}; \quad (k,\ell) = (0,0),(0,1),\ldots,(0,N) \equiv (1,0),\ldots,(1,N),\ldots \tag{31}$$

converges pointwisely to the I-projection $R^\star$ of the distribution R on the set $\mathcal{E}$.

Proof. According to Proposition 4 we can write for the I-projection $R^\star$ equation (23)

$$I(R^\star, P^{(k,\ell-1)}) = I(R^\star, P^{(k,\ell)}) + I(P^{(k,\ell)}, P^{(k,\ell-1)}). \tag{32}$$

Hence

$$\infty > I(R^\star, P^{(k,\ell-1)}) \geq I(R^\star, P^{(k,\ell)}) \geq 0. \tag{33}$$

Because $\left\{\left\{I(R^\star, P^{(k,\ell)})\right\}_{\ell=0}^{N}\right\}_{k=0}^{\infty}$ is a bounded monotone real sequence, it has a finite limiting value. Approaching the limit in equation (32) we can write

$$I(P^{(k,\ell)}, P^{(k,\ell-1)}) \to 0, \quad k \to \infty \text{ in the sense of (31)}, \tag{34}$$

which implies that

$$I(P^{(k,\ell)}, P^{(k,\ell-1)}) \to 0, \ k \to \infty \text{ for each fixed } \ell \in \{1,\ldots,N\}. \tag{35}$$

According to the inequality $\|P - Q\| \leq \sqrt{2I(P,Q)}$ mentioned in [1] it holds

$$\|P^{(k,\ell)} - P^{(k,\ell-1)}\| \to 0, \ k \to \infty \text{ for each fixed } \ell \in \{1,\ldots,N\}. \tag{36}$$

Let us take an arbitrary subsequence of iterative sequence (31). From this subsequence we can choose a sequence which converges to an element $Q^\star \in \mathcal{P}$ (see Theorem 1). As the number of variables N is finite, this convergent sequence must contain an infinite number of terms with the same (some) index $\tilde{\ell} \in \{1, \ldots, N\}$; thus, it must contain a subsequence $\left\{P^{(k_n,\tilde{\ell})}\right\}_{n=1}^{\infty}$ which converges to $Q^\star$, too.

Without any loss of generality let us consider $\tilde{\ell} = 1$. The limiting element $Q^\star$ of the sequence $\left\{P^{(k_n,1)}\right\}_{n=1}^{\infty}$ lies in the set $\mathcal{E}_1$ since $\mathcal{E}_1$ is closed in $(\mathcal{P}, \rho)$:

$$P^{(k_n,1)} \to Q^\star \in \mathcal{E}_1, \; n \to \infty. \tag{37}$$

The sequence $\left\{P^{(k_n,2)}\right\}_{n=1}^{\infty}$ converges to the element $Q^\star$, too, because of (36) and

$$\|P^{(k_n,2)} - Q^\star\| \leq \|P^{(k_n,2)} - P^{(k_n,1)}\| + \|P^{(k_n,1)} - Q^\star\| \tag{38}$$

and $Q^\star$ lies also in the set $\mathcal{E}_2$ since $\mathcal{E}_2$ is closed in $(\mathcal{P}, \rho)$. It can be analogously shown that $P^{(k_n,3)} \to Q^\star$ and $Q^\star \in \mathcal{E}_3$, ..., $P^{(k_n,N)} \to Q^\star$ and $Q^\star \in \mathcal{E}_N$, thus on the whole

$$\lim_{n\to\infty} P^{(k_n,\ell)} = Q^\star \text{ for each fixed } \ell \in \{1, \ldots, N\}; \quad Q^\star \in \bigcap_{\ell=1}^{N} \mathcal{E}_\ell = \boldsymbol{\mathcal{E}}. \tag{39}$$

We shall show that $Q^\star = R^\star$.

We know that $R^\star$ is the I-projection of each iteration $P^{(k_n,\ell)}$ on the set $\boldsymbol{\mathcal{E}}$ (see Theorem 3). Therefore, for an arbitrarily chos n but fixed $\tilde{\ell} \in \{1, \ldots, N\}$, we can write

$$I(R^\star, P^{(k_n,\tilde{\ell})}) \leq I(Q^\star, P^{(k_n,\tilde{\ell})}) \tag{40}$$

and letting n approach infinity in the above relation (it is possible since $\boldsymbol{X}$ is finite) we get

$$I(R^\star, Q^\star) \leq I(Q^\star, Q^\star) = 0. \tag{41}$$

Hence $Q^\star = R^\star$.

Since the I-projection $R^\star$ is unique we have proved that each convergent subsequence of $\left\{P^{(k,\ell)}\right\}_{k=0}^{\infty}$ converges to $R^\star$ in $(\mathcal{P}, \rho)$. That is why the limiting element in $(\mathcal{P}, \rho)$ of the whole iterative sequence is the I-projection $R^\star$.

The pointwise convergence of the sequence follows from the convergence in $(\mathcal{P}, \rho)$. □

6. MODIFICATION OF IPFP FOR DISTRIBUTION MIXTURE

Let us suppose now that the first term of the iterative sequence is a distribution mixture $R \in \boldsymbol{\mathcal{S}}_M$. We shall prove that in such case each term of the sequence produced by the IPFP algorithm has the form of finite distribution mixture of product components:

$$P^{(k,\ell)}(\boldsymbol{x}) \;=\; \sum_{m=1}^{M} w_m^{(k,\ell)} \prod_{n=1}^{N} p_n^{(k,\ell)}(x_n|m), \quad \boldsymbol{x} \in \boldsymbol{X},$$

$$w_m^{(k,\ell)} \geq 0, \qquad \sum_{m=1}^{M} w_m^{(k,\ell)} = 1,$$
$$\sum_{x \in X_n} p_n^{(k,\ell)}(x|m) = 1, \; k = 0, 1, \ldots ; \; \ell = 0, 1, \ldots, N \tag{42}$$

where $w_m^{(k,\ell)}$ is the weight of the mth component and $p_n^{(k,\ell)}(x|m)$ is the univariate distribution of the variable v_n corresponding to the mth component in the step (k, ℓ).

Proposition 5. Let $R \in \mathcal{S}_M$. Then $P^{(k,\ell)} \in \mathcal{S}_M$ for all k, ℓ, i.e. there exist weights $w_m^{(k,\ell)}$ and probability distributions $p_n^{(k,\ell)}(x|m)$ so that the iteration $P^{(k,\ell)}$ computed according to formula (14), (15) can be written in the form (42).

Proof. Let us suppose by induction that $P^{(k,\ell-1)} \in \mathcal{S}_M$, i.e.

$$P^{(k,\ell-1)}(\boldsymbol{x}) = \sum_{m=1}^{M} w_m^{(k,\ell-1)} \prod_{n=1}^{N} p_n^{(k,\ell-1)}(x_n|m), \tag{43}$$

where

$$w_m^{(k,\ell-1)} \geq 0, \qquad \sum_{m=1}^{M} w_m^{(k,\ell-1)} = 1,$$
$$p_n^{(k,\ell-1)}(x|m) \geq 0 \text{ for } x \in X_n, \qquad \sum_{x \in X_n} p_n^{(k,\ell-1)}(x|m) = 1. \tag{44}$$

Let us set

$$w_m^{(k,\ell)} = \sigma_m^{(k,\ell)} w_m^{(k,\ell-1)}, \quad \sigma_m^{(k,\ell)} = \sum_{x \in X_\ell} C^{(k,\ell)}(x)\, p_\ell^{(k,\ell-1)}(x|m), \tag{45}$$

$$p_n^{(k,\ell)}(x|m) = p_n^{(k,\ell-1)}(x|m) \text{ for } n \neq \ell, \quad x \in X_n, \tag{46}$$

$$p_\ell^{(k,\ell)}(x|m) = \frac{1}{\sigma_m^{(k,\ell)}}\, C^{(k,\ell)}(x)\, p_\ell^{(k,\ell-1)}(x|m), \quad x \in X_\ell, \tag{47}$$

where $C^{(k,\ell)}$ is defined in (15) and the ℓth marginal of the mixture $P^{(k,\ell-1)}$ has the form:

$$P_\ell^{(k,\ell-1)}(x_\ell) = \sum_{m=1}^{M} w_m^{(k,\ell-1)} p_\ell^{(k,\ell-1)}(x_\ell|m). \tag{48}$$

We shall show that $\sigma_m^{(k,\ell)} > 0$ for all m. Let $m \in \{1, \ldots, M\}$ be fixed. All terms of the sum $\sigma_m^{(k,\ell)}$ are nonnegative. We shall prove by contradiction that at least one of them is positive. Let us suppose that

$$C^{(k,\ell)}(x)\, p_\ell^{(k,\ell-1)}(x|m) = 0 \text{ for all } x \in X_\ell. \tag{49}$$

Let us denote as $\tilde{X}_\ell \subset X_\ell$ the set of points in which the distribution $p_\ell^{(k,\ell-1)}(x|m)$ takes positive values:

$$p_\ell^{(k,\ell-1)}(x|m) > 0 \text{ for } x \in \tilde{X}_\ell \ ; \ p_\ell^{(k,\ell-1)}(x|m) = 0 \text{ for } x \in X_\ell, \ x \notin \tilde{X}_\ell. \tag{50}$$

The assumption (49) implies that $C^{(k,\ell)}(x) = 0$ for all $x \in \tilde{X}_\ell$, which, according to formula (15), is equivalent to $P_\ell^{(k,\ell-1)}(x) \neq 0$ and $q_\ell(x) = 0$ for all $x \in \tilde{X}_\ell$. On the whole we can write:

$$\begin{aligned} &P_\ell^{(k,\ell-1)}(x) \neq 0 \text{ and } q_\ell(x) = 0 \text{ for } x \in \tilde{X}_\ell, \\ &P_\ell^{(k,\ell-1)}(x) = 0 \text{ for } x \in X_\ell, \ x \notin \tilde{X}_\ell. \end{aligned} \tag{51}$$

Because $P_\ell^{(k,\ell)} \ll P_\ell^{(k,\ell-1)}$ (see (25)) and $P_\ell^{(k,\ell)} \in \mathcal{E}_\ell$ (see Proposition 3) the relations (51) imply that

$$P_\ell^{(k,\ell)}(x) = 0 \text{ for all } x \in X_\ell \tag{52}$$

which is impossible. We have proved that the dividing by $\sigma_m^{(k,\ell)}$ in expression (47) is correct.

The next iteration, according to formulae (14), (15) and (45)–(47), can be then written as:

$$\begin{aligned} &P^{(k,\ell)}(\boldsymbol{x}) = C^{(k,\ell)}(x_\ell)\ P^{(k,\ell-1)}(\boldsymbol{x}) \\ &= \sum_{m=1}^{M} w_m^{(k,\ell-1)} \sigma_m^{(k,\ell)} \frac{1}{\sigma_m^{(k,\ell)}}\ C^{(k,\ell)}(x_\ell)\ p_\ell^{(k,\ell-1)}(x_\ell|m) \prod_{n=1, n\neq\ell}^{N} p_n^{(k,\ell-1)}(x_n|m) \\ &= \sum_{m=1}^{M} w_m^{(k,\ell)} \prod_{n=1}^{N} p_n^{(k,\ell)}(x_n|m). \end{aligned} \tag{53}$$

As it follows from (14), (15), (45)–(47) $p_n^{(k,\ell)}(x|m)$ are probability distributions and $w_m^{(k,\ell)} \geq 0$. It remains to be proved that $w_m^{(k,\ell)}$ sum to one:

$$\begin{aligned} 1 &= \sum_{\boldsymbol{x}\in\boldsymbol{X}} P^{(k,\ell)}(\boldsymbol{x}) = \sum_{m=1}^{M} w_m^{(k,\ell)} \sum_{\boldsymbol{x}\in\boldsymbol{X}} \prod_{n=1}^{N} p_n^{(k,\ell)}(x_n|m) \\ &= \sum_{m=1}^{M} w_m^{(k,\ell)} \sum_{x_1\in X_1} \cdots \sum_{x_{N-1}\in X_{N-1}} \prod_{n=1}^{N-1} p_n^{(k,\ell)}(x_n|m) \sum_{x_N\in X_N} p_N^{(k,\ell)}(x_N|m) \\ &= \sum_{m=1}^{M} w_m^{(k,\ell)} \sum_{x_1\in X_1} \cdots \sum_{x_{N-2}\in X_{N-2}} \prod_{n=1}^{N-2} p_n^{(k,\ell)}(x_n|m) \sum_{x_{N-1}\in X_{N-1}} p_{N-1}^{(k,\ell)}(x_{N-1}|m) \\ &= \sum_{m=1}^{M} w_m^{(k,\ell)}. \end{aligned} \tag{54}$$

□

Now we summarize the properties of IPFP algorithm modified for distribution mixtures.

Theorem 5. Let $R \in \mathcal{S}_M$ be a distribution mixture such that $I(\tilde{Q}, R) < \infty$ holds for some $\tilde{Q} \in \mathcal{E}$. Then the following three assertions are true:

1. R has a unique I-projection $R^\star \in \mathcal{E}$ on the set $\mathcal{E}$.
2. The iterative sequence $\left\{ \left\{ P^{(k,\ell)} \right\}_{\ell=0}^{N} \right\}_{k=0}^{\infty}$ defined by formulae (42), (45) – (47) converges pointwisely to $R^\star$.
3. $R^\star \in \mathcal{S}_M$.

Proof. Assertions 1 and 2 follow immediately from Theorem 2, Theorem 4 and Proposition 5. It remains to prove that the I-projection $R^\star$ has the form of finite distribution mixture with M product components (3). The sequences

$$\left\{ \left\{ w_m^{(k,\ell)} \right\}_{\ell=0}^{N} \right\}_{k=0}^{\infty}, \quad \left\{ \left\{ p_n^{(k,\ell)}(x|m) \right\}_{\ell=0}^{N} \right\}_{k=0}^{\infty} \tag{55}$$

are bounded for each $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots, N\}$, $x \in X_n$; the numbers M, N and the set $\boldsymbol{X}$ are finite.

Now we utilize repeatedly the fact that from every bounded sequence a convergent subsequence can be chosen. Thus there exists some subsequence of indices $\{(k_i, \ell_i)\}_{i=1}^{\infty}$ and exist values $w_m^\star$, $p_n^\star(x|m)$ such that it holds:

$$\begin{aligned} \lim_{i\to\infty} w_m^{(k_i,\ell_i)} &= w_m^\star \geq 0, \qquad m = 1, \ldots, M, \\ \lim_{i\to\infty} p_n^{(k_i,\ell_i)}(x|m) &= p_n^\star(x|m) \geq 0, \quad m = 1, \ldots, M,\ n = 1, \ldots, N,\ x \in X_n, \end{aligned} \tag{56}$$

$$\sum_{m=1}^{M} w_m^\star = 1, \quad \sum_{x\in X_n} p_n^\star(x|m) = 1. \tag{57}$$

The corresponding subsequence $\left\{ P^{(k_i,\ell_i)} \right\}_{i=1}^{\infty}$ of the iterative sequence $\left\{ \left\{ P^{(k,\ell)} \right\}_{\ell=0}^{N} \right\}_{k=0}^{\infty}$ converges, of course, to $R^\star$ and in addition it holds:

$$\begin{aligned} \lim_{i\to\infty} P^{(k_i,\ell_i)}(\boldsymbol{x}) &= \lim_{i\to\infty} \sum_{m=1}^{M} w_m^{(k_i,\ell_i)} \prod_{n=1}^{N} p_n^{(k_i,\ell_i)}(x_n|m) \\ &= \sum_{m=1}^{M} w_m^\star \prod_{n=1}^{N} p_n^\star(x_n|m) \quad \text{for all } \boldsymbol{x} = (x_1, \ldots, x_N) \in \boldsymbol{X}. \end{aligned} \tag{58}$$

Uniqueness of the limit implies that

$$R^\star(\boldsymbol{x}) = \sum_{m=1}^{M} w_m^\star \prod_{n=1}^{N} p_n^\star(x_n|m) \quad \text{for all } \boldsymbol{x} = (x_1, \ldots, x_N) \in \boldsymbol{X} \tag{59}$$

so that $R^\star \in \mathcal{S}_M$. □

7. EXAMPLE: PROBABILISTIC SOLUTION OF A LOGICAL PUZZLE

To demonstrate the interesting properties of the probabilistic approach the IPFP algorithm will be aplied to a purely logical problem.

Formulation of the problem

Eight students from different classes of a school – Anthony, Eve, Francis, Charles, John, Mary, Tanya and Peter – represented their classes – I. A, I. B, II. A, II. B, II. C, III. A, III. B and III. C – during a chess championship. We have to find out the classes represented by individual students. All information needed for the correct solution is contained in the following sentences:

1. In the first round Charles played with the student from II. C.
2. The student from I. B came after the first round.
3. In the second round the student from I. A played with Mary.
4. In the second round John played with Eve.
5. After the second round Anthony did not continue.
6. Because of Anthony's absence Francis did not play in the third round.
7. Because of Anthony's absence the student from II. A had no adversary in the fourth round.
8. Because of Anthony's absence John did not play in the fifth round.
9. In the third round Tanya won against the student from I. A.
10. In the third round Charles drew the game with the student from II. B.
11. In the fourth round the student from III. B played with Tanya.
12. In the fourth round Eve played with Charles.
13. After the sixth round the interrupted encounter of students from II. C and III. A was continued.

Let us recall that, as usual, each couple contested during the championship at most once and that each student played one game in one round at most.

Solution

To simplify notation we denote the eight names (Anthony, Eve, Francis, Charles, John, Mary, Tanya, Peter) by the symbols $n_1, n_2, \ldots, n_8$, respectively and the eight classes (I. A, I. B, II. A, II. B, II. C, III. A, III. B, III. C.) by $c_1, c_2, \ldots, c_8$. We introduce two discrete random variables – v_1 (student's name) and v_2 (class) taking the values:

$$\begin{aligned} v_1 \quad &: \quad x_1 \in X_1 = \{n_1, n_2, \ldots, n_8\}, \\ v_2 \quad &: \quad x_2 \in X_2 = \{c_1, c_2, \ldots, c_8\}. \end{aligned} \tag{60}$$

The joint distribution of the variables v_1, v_2 is assumed to be in the form of the finite mixture

$$P(x_1, x_2) = \sum_{m=1}^{M} w_m \, p_1(x_1|m) \, p_2(x_2|m), \quad (x_1, x_2) \in X_1 \times X_2. \tag{61}$$

The mixture components correspond to the mutually exclusive hypotheses of the type $m = (n_i, c_j)$ (the student n_i represents the class c_j). If we introduce two indices then the knowledge base can be rewritten as

$$P(x_1, x_2) = \sum_{i=1}^{8} \sum_{j=1}^{8} w(i,j)\; p_1(x_1|n_i)\; p_2(x_2|c_j) \tag{62}$$

where

$$\begin{aligned} p_1(x_1|n_i) &= \delta(x_1, n_i) &&= 1,\ x_1 = n_i, \\ &&&= 0,\ x_1 \neq n_i, \\ p_2(x_2|c_j) &= \delta(x_2, c_j) &&= 1,\ x_2 = c_j, \\ &&&= 0,\ x_2 \neq c_j, \\ w(i,j) &= P(n_i, c_j) &&= P\{v_1 = n_i, v_2 = c_j\}. \end{aligned} \tag{63}$$

In this way each component corresponds to one possible combination of name and class and the mixture is actually defined by the component weights $w(i,j)$.

First we choose the number of components M as large as possible ($M = 64$). Some of the corresponding hypotheses may be excluded using the information contained in the sentences $1 - 13$ so that the number of components may be reduced.

Table 1. Zero weights.

sentences	⇒ conclusion	zero	sentences	⇒ conclusion	zero
1	Charles ∉ II. C	$w(4,5)$	1, 2	Charles ∉ I. B	$w(4,2)$
1, 12	Eve ∉ II. C	$w(2,5)$	1, 13	Charles ∉ III. A	$w(4,6)$
3	Mary ∉ I. A	$w(6,1)$	3, 4	Eve ∉ I. A	$w(2,1)$
3, 4	John ∉ I. A	$w(5,1)$	5, 7	Anthony ∉ II. A	$w(1,3)$
5, 9	Anthony ∉ I. A	$w(1,1)$	5, 10	Anthony ∉ II. B	$w(1,4)$
5, 11	Anthony ∉ III. B	$w(1,7)$	5, 13	Anthony ∉ III. A	$w(1,6)$
5, 13	Anthony ∉ II. C	$w(1,5)$	6, 7	Francis ∉ II. A	$w(3,3)$
6, 9	Francis ∉ I. A	$w(3,1)$	6, 10	Francis ∉ II. B	$w(3,4)$
7, 8	John ∉ II. A	$w(5,3)$	7, 11	Tanya ∉ II. A	$w(7,3)$
7, 12	Eve ∉ II. A	$w(2,3)$	7, 12	Charles ∉ II. A	$w(4,3)$
9	Tanya ∉ I. A	$w(7,1)$	9, 10	Tanya ∉ II. B	$w(7,4)$
9, 10	Charles ∉ I. A	$w(4,1)$	10	Charles ∉ II. B	$w(4,4)$
10, 12	Eve ∉ II. B	$w(2,4)$	11	Tanya ∉ III. B	$w(7,7)$
11, 12	Eve ∉ III. B	$w(2,7)$	11, 12	Charles ∉ III. B	$w(4,7)$

As it follows from the above thirteen sentences, 28 weights are zero (see Tab. 1). The remaining 36 nonzero weights are equally set: $w(i,j) = 1/36$. At this point the problem represents the classical logical puzzle of "zebra" type: we know that the solution is unique and we have to determine the corresponding eight nonzero weights.

Distribution (62) with the weights defined as above represents the original multidimensional distribution P (the first iteration of the described algorithm). The role

of the additional expert knowledge (univariate marginal constraints) is played now by the natural assumption that the marginal probability distributions of names and classes are uniform:

$$q_1(n_i) = \sum_{j=1}^{8} P(n_i, c_j) = \frac{1}{8}, \quad i = 1, \ldots, 8,$$
$$q_2(c_j) = \sum_{i=1}^{8} P(n_i, c_j) = \frac{1}{8}, \quad j = 1, \ldots, 8. \tag{64}$$

During the computation eight weights appeared to be significantly nonzero. Their convergence is shown in Fig. 1. After 1000 iterations their values were following:

$$w(1,2) = 0.1243 \quad w(2,6) = 0.1241 \quad w(3,7) = 0.1241 \quad w(4,8) = 0.1246$$
$$w(5,4) = 0.1243 \quad w(6,3) = 0.1246 \quad w(7,5) = 0.1240 \quad w(8,1) = 0.125 \tag{65}$$

These nonzero weights correspond to the following correct logical solution of the puzzle:

Anthony ∈ I. B	Eve ∈ III. A	Francis ∈ III. B	Charles ∈ III. C
John ∈ II. B	Mary ∈ II. A	Tanya ∈ II. C	Peter ∈ I. A.

This was the computation in case of complete input information when the solution of the problem was unique in the sense that the eight corresponding weights converged to the value 1/8 (see Fig. 1) whereas the remaining weights approached 0.

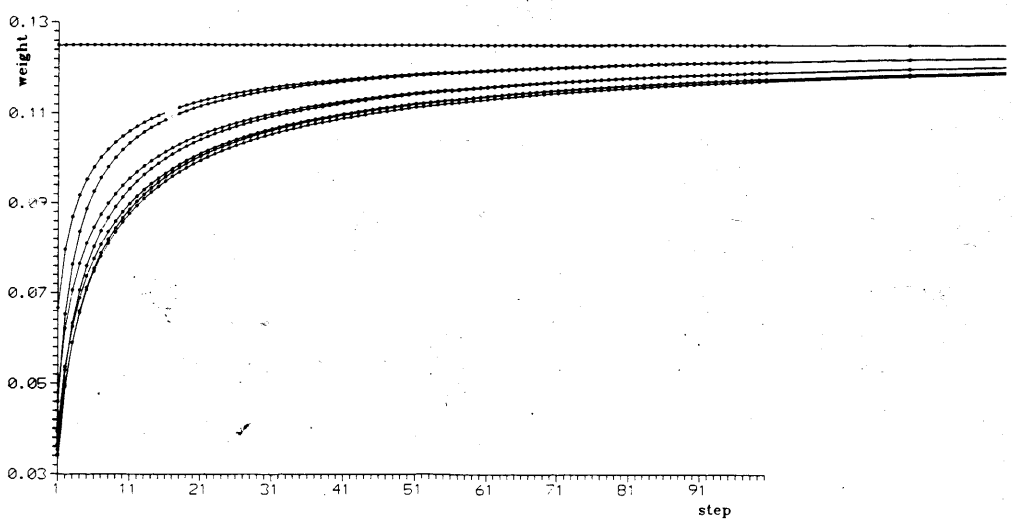

Fig. 1. The convergence of the 8 largest weights.

The properties of the algorithm were tested also for missing input knowledge. If some information of the type $w(i,j) = 0$ is supressed then the solution of the puzzle is not more unique. The algorithm determines "exactly" only some pairs

(i, j) (at best 6 pairs of 8 pairs wanted) in the sense that their weights converge to 1/8. The weights corresponding to the admissible alternatives converge to some nonzero values and the remaining weights converge to zero. Usually the result has a reasonable interpretation: if some couples (i, j) are equally probable then the corresponding weights are equal, too.

8. CONCLUSION

The iterative proportional fitting procedure was originally designed to adjust relative frequencies in contingency tables to some known marginal probabilities. In the present paper it is modified for a special class of distribution mixtures and univariate marginal constraints.

It should be emphasized that the assumed product form of mixture components is essential to enable us the separate norming of any marginal. A generalization for multi-dimensional marginal constraints seems to be a difficult task.

A logical puzzle to illustrate the method was chosen intentionally because the convergence in the case of typically probabilistic problems follows from the proved theorems. The numerical example (see Section 7) shows that even a rather complicated logical puzzle can be solved in the framework of the probabilistic approach.

(Received November 4, 1993.)

REFERENCES

[1] I. Csiszár: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. *3* (1975), 1, 146–158.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. *B 39* (1977), 1–38.

[3] J. Grim: Knowledge representation and uncertainty processing in the probabilistic expert system PES. Internat. J. Gen. Syst. *22* (1994), 2, 103–111.

[4] J. Grim: Probabilistic expert systems and distributions mixtures. Computers and Artificial Intelligence *9* (1990), 3, 241–256.

[5] C. T. Ireland and S. Kullback: Contingency tables with given marginals. Biometrika *55* (1968), 1, 179.

Ing. Jana Vejvalková, katedra matematiky fakulty jaderné a fyzikálně inženýrské ČVUT (Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering – Czech Technical University), Trojanova 13, 120 00 Praha 2. Czech Republic.