Stochastic and Physical Monitoring Systems
SPMS 2016
June 20 - June 24, 2016
Dobřichovice, Czech Republic

Feasibility Study of an Interactive Medical Diagnostic Wikipedia

Jiří Grim

Department of Pattern Recognition, Institute of Information Theory and Automation of the Czech Academy of Sciences, P.O. BOX 18, CZ-18208 Prague 8, Czech Republic

Email: grim@utia.cas.cz

Abstract. Considering different application possibilities of product distribution mixtures we have proposed three formal tools in the last years, which can be used to accumulate decision-making know-how from particular diagnostic cases. First, we have developed a structural mixture model to estimate multidimensional probability distributions from incomplete and possibly weighted data vectors. Second, we have shown that the estimated product mixture can be used as a knowledge base for the Probabilistic Expert System (PES) to infer conclusions from definite or even uncertain input information. Finally we have shown that, by using product mixtures, we can exactly optimize sequential decision-making by means of the Shannon formula of conditional informativity. We combine the above statistical tools in the framework of an interactive open-access medical diagnostic system with automatic accumulation of decision-making knowledge.

Key words: Multivariate statistics; Medical diagnostics; Product mixtures; Incomplete data; Sequential classification; EM algorithm.

1 Introduction

The great success of Wikipedia is based on the joint effort of great many people bringing together small pieces of knowledge in textual form. Motivated by the surprising extent and quality of this information source we propose a statistical platform to accumulate decision-making know-how from particular diagnostic cases. This project is based on three statistical tools developed in recent years in different applications of product distribution mixtures.

First, in a series of papers we have studied different aspects of the structural mixture model [7], [14], [15], [21] to estimate probability distributions in multidimensional spaces from incomplete [16] and possibly weighted data vectors [20]. By means of a structural "background" substitution technique we can evaluate conditional distributions in terms of subsets of variables while ignoring the remaining variables. Second, we have proposed to use the estimated discrete mixture of product components as a knowledge base for the

Probabilistic Expert System (PES, [8], [9], [10], [16]) with the aim to infer conclusions from either definite or uncertain input information. Given a sub-vector of input values or a probability distribution on the input subspace, we can thus compute the corresponding conditional distributions of arbitrary output variables. Finally, we have shown that, in case of mixtures of product components we have a unique possibility to evaluate the Shannon formula of conditional informativity exactly [23]. By means of this criterion we can choose the most informative questions with respect to any subset of goal variables. In this paper we discuss the possibility of combining the above formal tools in the framework of a statistical open-access interactive diagnostic system with optimally controlled dialog and automatic accumulation of decision-making knowledge. Natural application area of a sequential interactive decision-making system is medical diagnostics [1], [2], [4].

The basic idea of the project is to accumulate large statistical data sets by means of an interactive diagnostic application freely available online. We assume that anonymous users can be motivated by diagnostic information to specify some symptoms in an interactively controlled dialog. The final output protocol including symptoms and diagnoses can be stored as an anonymous by-product in the database, which is the fundamental source for estimating the probabilistic knowledge base.

The diagnostic result has to be formulated as a recommendation to consult a physician and therefore the protocol of the dialog should contain maximum useful information including symptoms and possible diagnoses. The physician should be motivated to join in the interactive dialog and possibly correct or complete the patient's data according to his/her personal opinion. In this way the active cooperation of the physicians can improve the general validity of the data.

There is obviously no guarantee that the statistical database arising in an open-access mode will be reliable and error free but, in this respect, the diagnostic system provides automatic self-correcting possibilities. The estimated product mixture model can be used "backwards" to eliminate or suppress the incorrect or suspicious data by weighting. In view of the self-correcting mechanisms concerning the data, variables and components, the process of designing hypotheses and collecting data may be viewed on as rather robust and open for anybody.

A typical feature of the considered medical decision-making is a large number of discrete diagnostic and symptom variables. It is therefore important that, at both the estimation and application stages, the structural mixture model can be treated as having no fixed dimension. We can arbitrarily add or remove variables or mixture components at any level of the design process and, simultaneously, we can optimize the mixture parameters by using incomplete data.

The goal of this paper is to describe the theoretical background of the proposed diagnostic system in formal statistical terms. In Section 2 we first formalize the problem of medical decision-making. Section 3 describes the role of the Probabilistic Expert System and its application in a controlled dialog scheme. Section 4 describes the "engine" of the diagnostic system concerning the details of estimating the probabilistic knowledge base from data. In Section 5 we discuss the problem of initial parameters, and Section 6 summarizes the concluding remarks. However, the problem of interactive medical diagnostics is extremely complex and there are many related medical aspects and questions going beyond the scope of this paper.

2 Problem of Medical Decision-Making

The goal of medical diagnostics is to derive diagnostic conclusions from a set of symptoms and informative data. Without essential loss of generality, we can describe the input information (symptoms) by a set of discrete variables $x_1, x_2, \ldots, x_K$, and the diagnoses by discrete variables $y_1, y_2, \ldots, y_J$ since possible continuous variables can be discretized. A typical diagnostic variable will be binary (negative, positive) but can be of a general discrete type in the case of several mutually exclusive alternatives. Below, we sometimes include all discrete variables into a single N-dimensional vector for convenience:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}, \quad x_n \in \mathcal{X}_n, \quad \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N, \quad y_j = x_{K+j}, \ j = 1, 2, \ldots, J,$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_J) \in \mathcal{Y} \equiv \mathcal{X}_{K+1} \times \cdots \times \mathcal{X}_{K+J}, \quad N = K + J.$$

A particular diagnostic case can thus be described by a vector $\boldsymbol{x} \in \mathcal{X}$ containing some known symptoms $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ and some related diagnoses $y_{j_1}, y_{j_2}, \ldots, y_{j_l}$. Obviously, such a vector will be always incomplete or even sparse since we have to consider a large number of possible symptoms and diagnostic variables, while only a small part of them is known or can be specified in a particular case.[1] For this reason the possibility of learning from incomplete data is essential.

For the sake of statistical decision-making the relationship between the diagnostic and symptom variables can be described in full generality by a joint probability distribution. Considering mixtures of product components, we approximate unknown discrete probability distributions in the form

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m), \quad F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m), \quad \mathcal{M} = \{1, \ldots, M\} \tag{1}$$

$$w_m \geq 0, \quad \sum_{m \in \mathcal{M}} w_m = 1, \quad \sum_{\xi_n \in \mathcal{X}_n} f_n(\xi_n|m) = 1, \quad \mathcal{N} = \{1, \ldots, N\}$$

where $x_n \in \mathcal{X}_n$ are general discrete variables, $\mathcal{N}$ is the index set of variables, w_m are probabilistic weights, $\mathcal{M}$ is the component index set and $f_n(\cdot|m), n \in \mathcal{N}$ are univariate discrete probability distributions defined by the probabilities $f_n(\xi_n|m)$, $\xi_n \in \mathcal{X}_n$. We recall that discrete product mixtures are not identifiable (cf. [12], Lemma 1) but they are not restrictive as a statistical model since any discrete distribution can be expressed in the form (1), (cf. [12], Remark 1).

From the point of view of medical diagnostics the product mixtures have two important advantages: they can be estimated by means of EM algorithm [26],[27], [3], [6] and, especially, arbitrary marginal distribution is easily evaluated by omitting superfluous terms in the products. In particular, if we denote $P_C(\boldsymbol{x}_C)$ a marginal distribution of the mixture (1) corresponding to variables $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$:

$$\boldsymbol{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C, \quad C = \{i_1, i_2, \ldots, i_k\} \subset \mathcal{N}, \tag{2}$$

then we can write

$$P_C(\boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} w_m F_C(\boldsymbol{x}_C|m) = \sum_{m \in \mathcal{M}} w_m \prod_{s=1}^{k} f_{i_s}(x_{i_s}|m). \tag{3}$$

[1] We can formally describe the missing value as an additional discrete option.

The last Equation (3) is essential since, using the marginal property, we can estimate the mixture parameters from incomplete data (cf. [16]) and, by considering a structural mixture model [7], we can treat the decision problem as having no fixed dimension (cf. Section 4.1).

3 Probabilistic Expert System

The estimated product mixture (1) can be used as a knowledge base for the Probabilistic Expert System (PES), (cf. [8], [9], [10], [16]). In this way we can use the efficient inference mechanism of PES to derive diagnostic conclusions from the input variables (symptoms) in terms of conditional distributions. If we denote $P_{DC}(\boldsymbol{y}_D, \boldsymbol{x}_C)$ the marginal distribution of the mixture (1) corresponding to the symptom variables $\boldsymbol{x}_C$ and diagnostic variables $\boldsymbol{y}_D$

$$\boldsymbol{y}_D = (y_{j_1}, y_{j_2}, \ldots, y_{j_l}) \in \mathcal{Y}_D, \quad \mathcal{D} = \{j_1, j_2, \ldots, j_l\} \subset \mathcal{N}, \tag{4}$$

then we can write

$$P_{DC}(\boldsymbol{y}_D, \boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} w_m F_D(\boldsymbol{y}_D|m) F_C(\boldsymbol{x}_C|m) = \sum_{m \in \mathcal{M}} w_m \prod_{s=1}^{l} f_{j_s}(y_{j_s}|m) \prod_{s=1}^{k} f_{i_s}(x_{i_s}|m).$$

Thus, for any subset of symptoms $\boldsymbol{x}_C \in \mathcal{X}_C$ and diagnostic variables $\boldsymbol{y}_D \in \mathcal{Y}_D$ we can compute the related conditional distributions by the formula

$$P_{D|C}(\boldsymbol{y}_D|\boldsymbol{x}_C) = \frac{P_{DC}(\boldsymbol{y}_D, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m \in \mathcal{M}} W_m(\boldsymbol{x}_C) F_D(\boldsymbol{y}_D|m). \tag{5}$$

Here $W_m(\boldsymbol{x}_C)$ are the component weights corresponding to a given vector of symptoms $\boldsymbol{x}_C \in \mathcal{X}_C$:

$$W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j \in \mathcal{M}} w_j F_C(\boldsymbol{x}_C|j)}, \qquad m \in \mathcal{M}. \tag{6}$$

The conditional distribution (5) represents a general exact response to the definite input information $\boldsymbol{x}_C \in \mathcal{X}_C$. However, for an interested user it could be more intuitive to evaluate the conditional distributions of diagnostic variables separately, e.g.,

$$P_{j|C}(y_j|\boldsymbol{x}_C) = \frac{P_{jC}(y_j, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m \in \mathcal{M}} W_m(\boldsymbol{x}_C) f_j(y_j|m), \quad j \in \mathcal{D}. \tag{7}$$

In case of uncertain input represented in full generality by a given probability distribution $P^\star_C(\boldsymbol{x}_C)$, we have to make the substitution (5) or (7) in the formula of complete probability:

$$P^\star_D(\boldsymbol{y}_D) = \sum_{x_C \in \mathcal{X}_\mathcal{C}} P_{D|C}(\boldsymbol{y}_D|\boldsymbol{x}_C) P^\star_C(\boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} W^\star_m F_D(\boldsymbol{y}_D|m), \tag{8}$$

$$P^\star_j(y_j) = \sum_{x_C \in \mathcal{X}_\mathcal{C}} P_{j|C}(y_j|\boldsymbol{x}_C) P^\star_C(\boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} W^\star_m f_j(y_j|m) \tag{9}$$

where

$$W_m^\star = \sum_{x_C \in \mathcal{X}_\mathcal{C}} W_m(\boldsymbol{x}_C) P_C^\star(\boldsymbol{x}_C). \tag{10}$$

Recall that the uncertain input information in the general form $P_C^*(\boldsymbol{x}_C)$ is rarely known and must be approximated if only a partial knowledge is available, e.g., only the marginal distributions $P_n^*(x_n), n \in \mathcal{C}$, [9].

There is an interesting option to globally evaluate the reliability of the inference mechanism given the symptoms $\boldsymbol{x}_C$. Let us recall that the conditional distribution of diagnostic variables $\boldsymbol{y}_D$ is given as a weighted sum of the component distributions $F_D(\boldsymbol{y}_D|m)$, (cf. (5)) where the conditional weight $W_m(\boldsymbol{x}_C)$ reflects the particular importance of the m-th component, given the the symptoms $\boldsymbol{x}_C$. On the other hand the unconditional component weights w_m reflect the support of the respective components in the database. In this sense the scalar product of both weight vectors

$$\rho(\boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} w_m W_m(\boldsymbol{x}_C), \quad \boldsymbol{x}_C \in \mathcal{X}_C, \quad (0 < \rho(\boldsymbol{x}_C) < 1) \tag{11}$$

can be viewed as a global reliability measure of the conditional distributions $P_{D|C}(\boldsymbol{y}_D|\boldsymbol{x}_C)$ given the symptoms $\boldsymbol{x}_C \in \mathcal{X}_C$.

In recent years the product mixtures (1) have repeatedly been applied to pattern recognition, e.g., to recognition of hand-written numerals [14], [15], [23]. Let us note that classification based on the Bayes formula can be viewed as a special case of the probabilistic inference mechanism (9) if the variable y_j defines the recognized classes.

3.1 Relevant Diagnostic Variables

By means of the probabilistic inference mechanism we can compute the conditional distributions of all diagnostic variables given a set of input values (symptoms). However, the number of possible diagnoses may be very large and for the sake of an efficient interactive dialog the attention should be focused on a reasonably small subset of significant diagnostic variables, e.g., by ordering their importance.

The idea of the first choice is to compare the conditional and unconditional (prior) distributions of diagnostic variables since any substantial change of the distribution (caused by the given conditioning symptoms) suggests a risk which should be evaluated in more detail. We can formally compute the absolute difference between the two distributions $P_{j|C}(y_j|\boldsymbol{x}_C)$, $P_j(y_j)$:

$$\triangle P_j = \sum_{y_j \in \mathcal{Y}_j} |P_j(y_j) - P_{j|C}(y_j|\boldsymbol{x}_C)|, \quad j \in \mathcal{D}$$

or, e.g., the Kullback-Leibler information divergence:

$$I(P_j(\cdot)||P_{j|C}(\cdot|\boldsymbol{x}_C)) = \sum_{y_j \in \mathcal{Y}_j} P_j(y_j) \log \frac{P_j(y_j)}{P_{j|C}(y_j|\boldsymbol{x}_C)},$$

or some other suitable dissimilarity measure.

Unfortunately, formal dissimilarity does not reflect the medical meaning of diagnostic variables. The medical relevance can be introduced by specific weights but the user should have an option to express his personal interest and influence the direction of the dialog.

Let us recall that the probabilistic inference mechanism yields the conditional distributions of diagnostic variables in a multidimensional form (5). The general formula is not very intuitive from the point of view of a user but it can be useful for computation of conditional informativity of questions in the next section. Thus the underlying computational complexity is the main reason to keep the number of relevant diagnostic variables within certain limits. We recall that there is no irreversible information loss in the reduced number of diagnoses since the subsequent application of the inference mechanism takes all diagnostic variables into account again.

3.2 Optimal Choice of Questions

In the case of medical decision-making we always assume that the final decision is done by a physician and therefore the main purpose of the interactive system is to accumulate maximum diagnostically relevant information for the final consultation. The output of the inference mechanism supplies the conditional probabilities of possible diagnoses and may be useful for the physician. Thus the key problem of the interactively controlled dialog is the optimal choice of diagnostically relevant symptom variables (questions) from the available set. At each stage the user (patient) should be offered a set of additional questions ordered according to their informativity. The interactive process can be continued as long as the user is motivated to answer additional questions.

The evaluation of symptoms typically suggests several suspicious diagnoses. The goal of the so-called differential diagnostics in medical decision-making is to eliminate the irrelevant diagnoses by answering additional questions. The fundamental problem of choosing additional informative symptom variables (questions) can be solved optimally in the formal context of the probabilistic knowledge base.

We use the fact that, unlike other methods (cf. [1], [2], [5], [24], [25]), the knowledge base estimated in the form of product mixture provides a unique possibility to evaluate at any decision level the exact conditional informativity (in the Shannon sense) of the remaining symptom variables [23]. In other words, given a subset of symptoms $\boldsymbol{x}_C$ and a set of potentially relevant diagnostic variables $\boldsymbol{y}_D$,

$$\boldsymbol{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C, \quad \boldsymbol{y}_D = (y_{j_1}, y_{j_2}, \ldots, y_{j_l}) \in \mathcal{Y}_D, \tag{12}$$

we can compute we can evaluate the exact Shannon informativity with respect to the chosen diagnoses $\boldsymbol{y}_D$ for all the remaining symptom variables x_n, $n \notin \mathcal{C}$. In particular, we can directly write equations for both the related marginals and the conditional distributions the exact conditional Shannon informativity - with respect to the chosen diagnoses $\boldsymbol{y}_D$. In particular, we can directly write Eqs. for both the related marginals and the conditional distributions

$$P_{nC}(x_n, \boldsymbol{x}_C) = \sum_{m \in \mathcal{M}} w_m F_C(\boldsymbol{x}_C|m) f_n(x_n|m), \tag{13}$$

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{nC}(x_n, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m \in \mathcal{M}} W_m(\boldsymbol{x}_C) f_n(x_n|m), \tag{14}$$

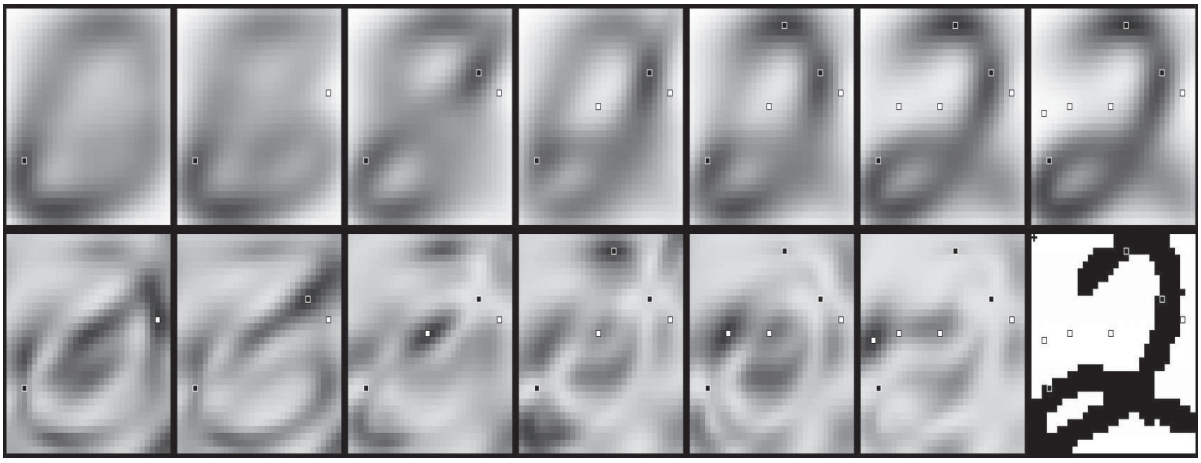

Figure 1: Sequential recognition of numerals. In the first row the images show the changing "expectation" of the classifier according to the currently uncovered raster fields. The images in the second row illustrate the conditional informativity of the hidden raster fields. The last image is the true underlying numeral two.

$$P_{DCn}(\boldsymbol{y}_D, \boldsymbol{x}_C, x_n) = \sum_{m \in \mathcal{M}} w_m F_D(\boldsymbol{y}_D|m) F_C(\boldsymbol{x}_C|m) f_n(x_n|m) \tag{15}$$

$$P_{D|Cn}(\boldsymbol{y}_D|\boldsymbol{x}_C, x_n) = \frac{P_{DCn}(\boldsymbol{y}_D, \boldsymbol{x}_C, x_n)}{P_{Cn}(\boldsymbol{x}_C, x_n)} = \sum_{m \in \mathcal{M}} W_m(\boldsymbol{x}_C, x_n) F_D(\boldsymbol{y}_D|m). \tag{16}$$

Here $W_m(\boldsymbol{x}_C, x_n)$ are the conditional component weights given the input vector $\boldsymbol{x}_C \in \mathcal{X}_C$ and the evaluated symptom variable x_n:

$$W_m(\boldsymbol{x}_C, x_n) = \frac{w_m F_C(\boldsymbol{x}_C|m) f_n(x_n|m)}{\sum_{j \in \mathcal{M}} w_j F_C(\boldsymbol{x}_C|j) f_n(x_n|j)} = \frac{w_m F_C(\boldsymbol{x}_C|m) f_n(x_n|m)}{P_{Cn}(\boldsymbol{x}_C, x_n)}. \tag{17}$$

In view of Equation (16) the conditional Shannon informativity of the symptom variable $x_n, n \notin \mathcal{C}$ with respect to the diagnostic variables $\boldsymbol{y}_D$ can be computed for any given subvector $\boldsymbol{x}_C \in \mathcal{X}_C$ by means of the Shannon formula

$$I_{\boldsymbol{x}_C}(\mathcal{Y}_D, \mathcal{X}_n) = H_{\boldsymbol{x}_C}(\mathcal{Y}_D) - H_{\boldsymbol{x}_C}(\mathcal{Y}_D|\mathcal{X}_n) \tag{18}$$

where $H_{\boldsymbol{x}_C}(\mathcal{Y}_D), H_{\boldsymbol{x}_C}(\mathcal{Y}_D|\mathcal{X}_n)$ are the respective Shannon entropies:

$$H_{\boldsymbol{x}_C}(\mathcal{Y}_D) = \sum_{y_D \in \mathcal{Y}_D} -P_{D|C}(\boldsymbol{y}_D|\boldsymbol{x}_C) \log P_{D|C}(\boldsymbol{y}_D|\boldsymbol{x}_C), \tag{19}$$

$$H_{\boldsymbol{x}_C}(\mathcal{Y}_D|\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} P_{n|C}(x_n|\boldsymbol{x}_C) \sum_{y_D \in \mathcal{Y}_D} -P_{D|Cn}(\boldsymbol{y}_D|\boldsymbol{x}_C, x_n) \log P_{D|Cn}(\boldsymbol{y}_D|\boldsymbol{x}_C, x_n). \tag{20}$$

Recall that high Shannon information $I_{\boldsymbol{x}_C}(\mathcal{Y}_D, \mathcal{X}_n)$ reflects a strong statistical dependence between the variable x_n and the diagnostic variables $\boldsymbol{y}_D$ given the symptoms $\boldsymbol{x}_C \in \mathcal{X}_C$. It is maximum if the variable x_n uniquely determines the value of $\boldsymbol{y}_D$ given

$\boldsymbol{x}_C$. On the other hand the information (18) is zero if the variable x_n and the sub-vector $\boldsymbol{y}_D$ are conditionally independent given $\boldsymbol{x}_C \in \mathcal{X}_C$.

To illustrate the power of sequential decision-making we show in Fig.1 an example of recognition of numerals from the NIST Special Database 19 (cf. [23]). The digits are normalized to a 32×32 binary raster. At the beginning the numeral is not visible and the raster fields are uncovered sequentially according to maximum conditional informativity. In the first row Fig. 1 shows the changing "expectation" of the classifier according to the currently uncovered raster fields. The second row shows the corresponding conditional informativity of hidden raster fields. As we can see, merely seven visible raster fields are sufficient to correctly recognize the underlying numeral two.

3.3 Controlled Dialog Scheme

The interactive controlled dialog is the most important open-access user interface. It is the main source of data and therefore it should motivate potential users to accept the form of controlled dialog with the resulting diagnostic information and comments.

Similarly to a natural dialog between a physician and a patient, the user has to specify some symptoms and the Probabilistic Expert System playing the role of a physician should evaluate possible relevant diagnoses (in terms of conditional distributions of diagnostic variables). As mentioned earlier, the list of relevant diagnoses should be restricted to enable efficient evaluation of additional informative questions. A suitable choice can be suggested by means of a formal ordering as discussed in Section 3.1, but the final decision should be made by the user. We recall that the subsequent application of the inference mechanism including the next symptom takes all diagnostic variables into account again.

As shown in Section 3.2 the evaluation of conditional informativity provides an ordered list of the remaining questions which is recomputed after any new specific symptom is revealed. In this way the ordered list of relevant diagnoses $\boldsymbol{y}_D$ always reflects the complete input information $\boldsymbol{x}_C$.

There are many unresolved aspects of the interactive dialog to be considered specifically. Let us recall that the choice of subsequent informative questions based on the conditional Shannon informativity is formally applicable to very complex problems of differential diagnostics, but the computation could become time consuming and the selected questions are optimal only in a formal sense. Obviously, the Shannon formula does not fully reflect the medical importance of variables.

The interactive dialog can be stopped by the user at any time with the possibility to read the final protocol, i.e., the resulting diagnostic conclusions at a suitable level of details and in a desirable form including input symptoms, possible relevant diagnoses, suggested medication and comments. The final output for a user must always contain a recommendation to consult a physician. The protocol of the interactive dialog should be useful for the physician since his cooperation in storing reliable data is of high value. The physician should be motivated to enter the patient's dialog to check the communicated symptoms and verify the resulting diagnoses.

The symptoms and diagnoses in the final form can be stored as an incomplete data vector in the database. For this purpose any uncertain diagnostic output can be stored as a set of data vectors weighted by the corresponding probabilities.

4 Estimation of the Probabilistic Knowledge Base

The maximum-likelihood estimates of mixture parameters can be computed by means of the EM algorithm [3],[26],[27],[6]. Formally, given a finite set $\mathcal{S}$ of independent observations of the underlying N-dimensional random vector

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots\}, \quad \boldsymbol{x} = (x_1, x_2, \cdots, x_N) \in \mathcal{X} \tag{21}$$

we maximize the corresponding log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m) \right] = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m) \right] \tag{22}$$

by means of the following EM iteration equations ($m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S}$):

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \tag{23}$$

$$f_n^{'}(\xi_n|m) = \frac{1}{\sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}} \delta(\xi_n, x_n) q(m|\boldsymbol{x}), \quad \xi_n \in \mathcal{X}_n. \tag{24}$$

Here the apostrophe denotes the new parameter values in each iteration. We recall that the EM iteration equations have to be carefully implemented because there is a risk of multiple latent underflow in Equation (23). For this reason the conditional weights $q(m|\boldsymbol{x})$ have to be evaluated in a logarithmic form and suitably scaled (cf. [21]).

The log-likelihood criterion nearly always has local maxima and therefore the iterative computation depends on the starting point. Nevertheless, in cases of large data sets and a large number of components, possible local maxima do not differ very much and the corresponding approximation quality of the estimated mixture is usually comparable.

4.1 Structural Mixture Model

The structural (subspace) approach to product mixtures makes use of an idea originally proposed within the framework of statistical pattern recognition (cf. [7]). Introducing binary structural parameters $\phi_{mn} \in \{0, 1\}, n \in \mathcal{N}, m \in \mathcal{M}$ we define the mixture components in the form

$$F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}, \tag{25}$$

where $f_n(x_n|0), n \in \mathcal{N}$ are some fixed univariate background distributions. A convenient option is to set them equal to global marginal distributions: $f_n(x_n|0) = P_n^*(x_n), n \in \mathcal{N}$. If we set $\phi_{mn} = 0$ then we can replace any component-specific distribution $f_n(x_n|m)$ with the respective background distribution $f_n(x_n|0)$. We can equivalently write

$$F(\boldsymbol{x}|m) = F(\boldsymbol{x}|0) G(\boldsymbol{x}|m, \boldsymbol{\phi}_m), \quad m \in \mathcal{M}, \quad F(\boldsymbol{x}|0) = \prod_{n \in \mathcal{N}} f_n(x_n|0), \tag{26}$$

where $F(\boldsymbol{x}|0)$ is a nonzero "background" probability distribution and the component functions $G(\boldsymbol{x}|m,\boldsymbol{\phi}_m)$ include the structural parameters $\phi_{mn} \in \{0,1\}$:

$$G(\boldsymbol{x}|m,\boldsymbol{\phi}_m) = \prod_{n\in\mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}}, \quad \boldsymbol{\phi}_m = (\phi_{m1},\cdots,\phi_{mN}) \in \{0,1\}^N. \tag{27}$$

In this way, the component functions $G(\boldsymbol{x}|m,\boldsymbol{\phi}_m)$ may be defined on different subspaces. By using substitution (26) we can write the structural mixture in the form

$$P(\boldsymbol{x}) = F(\boldsymbol{x}|0) \sum_{m\in\mathcal{M}} w_m G(\boldsymbol{x}|m,\boldsymbol{\phi}_m), \tag{28}$$

The structural mixture model can be optimized by means of the EM algorithm in full generality (cf. [7], [15], [21]) by maximizing the corresponding log-likelihood function:

$$L = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \log \left[\sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|0) G(\boldsymbol{x}|m,\boldsymbol{\phi}_m)\right].$$

In the following iteration equations, the apostrophe denotes the new parameter values ($m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S}$):

$$q(m|\boldsymbol{x}) = \frac{w_m G(\boldsymbol{x}|m,\boldsymbol{\phi}_m)}{\sum_{j\in\mathcal{M}} w_j G(\boldsymbol{x}|j,\boldsymbol{\phi}_j)}, \quad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} q(m|\boldsymbol{x}), \tag{29}$$

$$f_n^{'}(\xi_n|m) = \frac{1}{\sum_{x\in\mathcal{S}} q(m|\boldsymbol{x})} \sum_{x\in\mathcal{S}} \delta(\xi_n, x_n) q(m|\boldsymbol{x}), \tag{30}$$

$$\gamma_{mn}^{'} = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} q(m|\boldsymbol{x}) \log \frac{f_n^{'}(x_n|m)}{f_n(x_n|0)}. \tag{31}$$

The optimal subset of nonzero parameters $\phi_{mn}^{'}$ is defined by the highest values $\gamma_{mn}^{'}$ and can be chosen by simple thresholding:

$$\phi_{mn}^{'} = \begin{cases} 1, & \gamma_{mn}^{'} > \tau \\ 0, & \gamma_{mn}^{'} \le \tau \end{cases}, \quad \tau = \frac{\alpha}{MN} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \gamma_{mn}^{'}, \quad (0<\alpha<1), \quad m \in \mathcal{M}. \tag{32}$$

In some cases the structural parameters defined by Equation (32) may tend toward overfitting and the following component-specific thresholds $\tau_m, m \in \mathcal{M}$ may be more robust :

$$\phi_{mn}^{'} = \begin{cases} 1, & \gamma_{mn}^{'} > \tau_m \\ 0, & \gamma_{mn}^{'} \le \tau_m \end{cases}, \quad \tau_m = \frac{\alpha}{N} \sum_{n\in\mathcal{N}} \gamma_{mn}^{'}, \quad (0<\alpha<1), \quad m \in \mathcal{M}. \tag{33}$$

Let us recall that the background probability distribution $F(\boldsymbol{x}|0)$ is reduced in Equation (29) as well as in the conditional weights $W_m(\boldsymbol{x}_C)$. In this way Equation (17) includes only the relevant variables:

$$W_m(\boldsymbol{x}_C) = \frac{w_m G(\boldsymbol{x}_C|m,\boldsymbol{\phi}_m)}{\sum_{j\in\mathcal{M}} w_j G(\boldsymbol{x}_C|j,\boldsymbol{\phi}_j)}, \; m \in \mathcal{M}. \tag{34}$$

The structural EM algorithm is directly applicable to data of extreme dimensionality (cf. [13], [18]) but the underlying data set should also be large. A specific feature of medical decision-making are "spars" data vectors which imply less complex statistical relationship between variables. For this reason the estimation of the related structural mixture model could be less demanding concerning the size of the training set - despite the large dimension of the diagnostic problem.

4.2 EM Algorithm for Missing Data

The problem of missing data is a traditional area of mathematical statistics. The data can be made complete by simply omitting the incomplete vectors or variables but, in this way, we lose a large part of the original information. Another possibility is to replace the missing values with some estimates (see e.g., [3]). However, the substituted values are always typical in some sense and therefore the natural variability of data would decrease. Alternatively, the product mixture model enables a simple possibility of directly processing incomplete data since the product components can be reduced to an arbitrary subspace currently specified by the vector $\boldsymbol{x} \in \mathcal{X}$. In other words, we estimate the mixture parameters using only the available data (cf. [16]).

In order to modify the EM algorithm for incomplete data we denote by $\mathcal{N}(\boldsymbol{x}) \subset \mathcal{N}$ the subset of indices of the defined variables in a given vector $\boldsymbol{x}$ and $\mathcal{S}_n \subset \mathcal{S}$ the subset of vectors with the defined value x_n:

$$\mathcal{N}(\boldsymbol{x}) = \{n \in \mathcal{N} : x_n \text{ is defined in } \boldsymbol{x}\}, \quad \mathcal{S}_n = \{\boldsymbol{x} \in \mathcal{S} : n \in \mathcal{N}(\boldsymbol{x})\}.$$

The log-likelihood function for incomplete data is given by Equation

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}(x)} f_n(x_n|m) \right] \tag{35}$$

and the corresponding modified EM iteration equations can be expressed in the form ($m \in \mathcal{M},\ n \in \mathcal{N}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{S}$):

$$q(m|\boldsymbol{x}) = \frac{w_m \prod_{n \in \mathcal{N}(x)} f_n(x_n|m)}{\sum_{j=1}^{M} w_j \prod_{n \in \mathcal{N}(x)} f_n(x_n|j)}, \quad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \tag{36}$$

$$f_n^{'}(\xi|m) = \frac{1}{\sum_{x \in \mathcal{S}_n} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}_n} \delta(\xi, x_n) q(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n. \tag{37}$$

Roughly speaking, we calculate the values $q(m|\boldsymbol{x})$ in Equation (36) only for the variables x_n currently available in $\boldsymbol{x}$ and the new parameters $f_n^{'}(\cdot|m)$ in Equation (37) are estimated only for data vectors $\boldsymbol{x} \in \mathcal{S}_n$ with the defined variable x_n. Obviously, there is a standard trade-off between the percentage of missing values and the estimation accuracy.

4.3 EM Algorithm for Weighted Data

The EM algorithm can be applied to arbitrarily weighted data for example to utilize some external knowledge about the meaning or reliability of data vectors [20]. Denoting $\gamma(\boldsymbol{x})$

the relative frequency of $\boldsymbol{x}$ in $\mathcal{S}$ we can write

$$\boldsymbol{x} \notin \mathcal{S} \Rightarrow \gamma(\boldsymbol{x}) = 0, \quad \Rightarrow \quad \sum_{x \in \mathcal{X}} \gamma(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} 1 = 1 \tag{38}$$

and we can express the log-likelihood function in the following equivalent form

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m)] \; = \sum_{x \in \mathcal{X}} \gamma(\boldsymbol{x}) \log[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m)] \tag{39}$$

We can use the equivalent notation in the EM iteration equations as well:

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad w_m^{'} = \sum_{x \in \mathcal{X}} \gamma(\boldsymbol{x}) q(m|\boldsymbol{x}), \tag{40}$$

$$f_n^{'}(\xi|m) = \sum_{x \in \mathcal{X}} \delta(\xi, x_n) \gamma(\boldsymbol{x}) \frac{q(m|\boldsymbol{x})}{w_m^{'}}, \quad \xi \in \mathcal{X}_n, \quad m \in \mathcal{M} \tag{41}$$

In words, the weighted modification of the log-likelihood function is maximized by weighted EM iteration equations (40), (41).

4.4 Self-Correcting Mechanisms of PES

The probabilistic knowledge base estimated from a large data set in the form of a product mixture can be used "backwards" to evaluate the reliability of the original data. We can identify erroneous or suspicious data vectors $\boldsymbol{x} \in \mathcal{S}$ by computing the log-likelihood $\log P(\boldsymbol{x})$. In recent years this approach has been successfully verified in evaluation of suspicious (malign) locations in screening mammograms (cf. [19], [20], [21]) and identification of defects or abnormalities in textures [17] or impaired pixels in images [22].

In large data sets, the unreliable, suspect or incorrect data can be automatically removed or suppressed by weighting, e.g.,

$$\gamma(\boldsymbol{x}) = \log P(\boldsymbol{x}) / \bar{L}, \qquad \bar{L} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log P(\boldsymbol{x}). \tag{42}$$

In general, by weighting the data in the EM algorithm we influence the form of the estimated distribution with unfavorable consequences for the final accuracy of the estimates [11]. However, unlike standard estimation problems, in the case of medical diagnostics we would increase the reliability of the resulting inference mechanism and suppress the influence of noisy, incorrect or atypical data.

In a similar way we can easily identify "useless" components having very small weights. In view of the related EM equation (40), the component weight w_m reflects how frequently and strongly the component $F(\boldsymbol{x}|m)$ fits to the available data. Except for rare diagnoses, the components with very low weight values w_m can be omitted without substantial effect on the accuracy of the inference mechanism. For the same reason we can remove sparsely used variables x_n for which $\sum_{m \in \mathcal{M}} \phi_{mn}$ is very small or zero.

5 Initial Design of the Knowledge Base

The structural mixture model is a weighted sum of product components. Each component defined as a product of univariate discrete distributions (i.e., histograms) can be viewed as an elementary diagnostic hypothesis. In this sense the initial components can be designed intuitively by specifying the typical co-occurrence of a diagnosis and related symptoms. A natural assumption is cooperation with medical experts; however, given a large training database, the initial intuitively designed components can be reliably optimized by means of the EM algorithm. The EM algorithm would automatically modify the suggested univariate distributions in components or suppress the weights of unsuitably or incorrectly specified components.

An important feature of the EM algorithm is its monotonic property. The mixture parameters at any phase of computation can be viewed as initial estimates and therefore new components can be added sequentially, in the course of iterations [12]. For the same reason a new variable x_n can be included at any level of the design process simply by specifying the corresponding background distribution $f_n(x_n|0)$.

6 Concluding Remarks

The design of an interactive medical diagnostic system is a difficult task, in which a lot of software development is necessary. This paper deals mainly with formal prerequisites for the interactive diagnostic system, but there are many open problems and questions which can hardly be assumed and solved in advance.

At the beginning, the design process should relate to a limited diagnostic area and refer to reliable support by medical experts. Both the symptoms and the diagnoses should be included sequentially in a reasonable hierarchical structuring. In the initial stages the system should refer to a sufficiently large and reliable database, preferably created in cooperation with a well-motivated community, e.g., students or physicians.

Acknowledgements

This project is supported by the Grants No. 14-02652S and 14-10911S of the Czech Science Foundation.

References

[1] M. Ben-Bassat. "Myopic policies in sequential classification". *IEEE Trans. Computers*, **C-27**, 170-174, 1978.

[2] M. Ben-Bassat. "Pattern-Based Interactive Diagnosis of Multiple Disorders: The MEDAS System". *IEEE Trans. Pattern. Anal. Mach. Int.*, **PAMI-2**, 148-160, 1980.

[3] 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**, l-38, 1977.

[4] B.J. Flehinger and R.L. Engle. "HEME: A Self-Improving Computer Program for Diagnosis-Oriented Analysis of Hematologic Diseases". *IBM Journal of Research and Development*, **19/6**, 557-564, 1975.

[5] K.S. Fu. *Sequential Methods in Pattern Recognition and Machine Learning*, New York: Academic, 1968.

[6] J. Grim. "On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions". *Kybernetika*, **l8/3**, 173-190, 1982. *http://dml.cz/dmlcz/124132*

[7] J. Grim. "Multivariate statistical pattern recognition with nonreduced dimensionality". *Kybernetika*, **22/2**, 142-157, 1986. *http://dml.cz/dmlcz/125022*

[8] J. Grim. "A dialog presentation of census results by means of the Probabilistic Expert System PES". In: *Proceedings of the Eleventh European Meeting on Cybernetics and Systems Research, Vienna 21-24 April 1992*, (Ed. R.Trappl), World Scientific, Singapore, 997-1005, 1992.

[9] J. Grim. "Knowledge representation and uncertainty processing in the probabilistic expert system PES". *International Journal of General Systems*, **22/2**, 103 - 111, 1994.

[10] J. Grim and P. Boček. "Statistical Model of Prague Households for Interactive Presentation of Census Data". In: *SoftStat'95. Advances in Statistical Software 5*, 271 - 278, Lucius & Lucius: Stuttgart, 1996.

[11] J. Grim, P. Pudil and P. Somol. "Boosting in probabilistic neural networks". In: *Proceedings of the 16th International Conference on Pattern Recognition*, (Kasturi R., Laurendeau D., Suen C. eds.). IEEE Computer Society, Los Alamitos, 136–139, 2002.

[12] J. Grim. "EM cluster analysis for categorical data". *Structural, Syntactic and Statistical Pattern Recognition.* (Eds. Yeung D. Y., Kwok J. T., Fred A.), Springer: Berlin, LNCS **4109**, 640-648, 2006.

[13] J. Grim, M. Haindl, P. Somol, and P. Pudil. "A subspace approach to texture modelling by using Gaussian mixtures". In: *Proc. of the 18th Int. Conf. ICPR 2006*, Eds. B. Haralick, T.K. Ho), 235–238, 2006.

[14] J. Grim and J. Hora. "Iterative principles of recognition in probabilistic neural networks". *Neural Networks*, **21/6**, 838-846, 2008.

[15] J. Grim and J. Hora, "Computational Properties of Probabilistic Neural Networks". *Artificial Neural Networks - ICANN 2010 Part II*, Springer: Berlin, LNCS **5164**, 52-61, 2010.

[16] J. Grim, J. Hora, P. Boček, P. Somol and P. Pudil. "Statistical Model of the 2001 Czech Census for Interactive Presentation". *Journal of Official Statistics*. **26/4**, 673-694, 2010.

[17] J. Grim, P. Somol, M. Haindl and P. Pudil. "A statistical approach to local evaluation of a single texture image". In: *Proc. of the 16-th Annual Symposium PRASA 2005.* (Nicolls F. ed.). University of Cape Town, 171-176, 2005.

[18] J. Grim, J. Novovičová and P. Somol. "Structural Poisson mixtures for classification of documents". In: *Proc. of the 19th International Conference on Pattern Recognition ICPR 2008*, 1-4, 2008. *http://dx.doi.org/10.1109/ICPR.2008.4761669*

[19] J. Grim, P. Somol, M. Haindl and J. Danes. "Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models". *IEEE Transactions on Image Processing*, **18/4**, 765-773, 2009.

[20] J. Grim. "Preprocessing of Screening Mammograms Based on Local Statistical Models". In: *Proc. of the 4th Int. Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2011*, Barcelona, ACM, 1-5, 2011.

[21] J. Grim and G.L. Lee. "Evaluation of Screening Mammograms by Local Structural Mixture Models". In: *Stochastic and Physical Monitoring Systems SPSM 2012*, Prague: Czech Technical University, 51-61, 2012.

[22] J. Grim, P. Somol, P. Pudil, I. Mkov and M.Malec. "Texture Oriented Image Inpainting based on Local Statistical Model". In: *Proc. 10th IASTED Conf. on Signal & Image Processing, SIP 2008*, Calgary : ACTA Press, 2008 - (Cristea, P.), 15-20, 2008.

[23] J. Grim. "Sequential pattern recognition by maximum conditional informativity". *Pattern Recognition Letters*, **45C**, 39-45, 2014.
http:// dx.doi.org/10.1016/j.patrec.2014.02.024

[24] M. Kurzynski, A. Zolnierek. "Sequential pattern recognition: naive Bayes versus fuzzy relation method". In: *Proc. Int. Conf. on Computational Intelligence for Modelling, Control and Automation*, IEEE, **1**, 1165-1170, 2005.

[25] J. Šochman, J. Matas. "WaldBoost - learning for time constrained sequential detection". In: *Proc. Computer Vision and Pattern Recognition, (CVPR 2005)*, IEEE Computer Society Conference on CVPR, **2**, 20-25, 2005.

[26] M.I. Schlesinger. "Relation between learning and self learning in pattern recognition". (in Russian), *Kibernetika*, (Kiev), **/2**, 81-88, 1968.

[27] M.I. Schlesinger and V. Hlaváč. *Ten lectures on statistical and syntactic pattern recognition.* Kluwer Academic Publishers, 2002.