DINES: A POSSIBILITY OF DIRECT DECISION MAKING WITHIN THE FRAME OF INES

OTAKAR KŘÍŽ

The idea to construct expert systems with the strictly probabilistic background, as suggested by Perez [6] proved to be both promising and fruitful. It found its implementation in the form of INtensional Expert System — INES (Perez, Jiroušek [7]). Certain modification of INES inference mechanism is presented in this paper. The new experimental system DINES is intended as an efficient sparring partner for the original INES to study further ways for enhancement of its discernment power.

1. BASIC FEATURES OF INES SYSTEM

The original idea of the system INES belongs to Perez who, inspired by his previous work [5], suggested in [6] a model of an expert system strictly based on probabilistic principles. We shall not enumerate advantages of intensional systems over rule-based ES but some important features of INES should be briefly mentioned. All uncertainty is completely expressed by an unknown theoretical distribution $P_{\eta\xi_1\xi_2\ldots\xi_n}(d, x_1, x_2, \ldots \ldots, x_n)$. η is a random variable taking values from a set of different diagnoses, ξ_i correspond to symptoms (when medical terminology is used). ES INES tries to find such approximation $\bar{P}$ of this actual distribution that, beside being consistent with all knowledge at our disposal, minimizes decision losses. This approach has not only good theoretical foundations but it enables also to look for the appropriate approximation without knowing the actual distribution. Another interesting feature of INES is that it accepts knowledge in a form of statistical data. It is possible to elicit information from experts in a rule-like way where "weights" have the strict interpretation of conditional probability (see [4]), but the most natural way of feeding knowledge into INES is a set of oligodimensional distributions. Their statistical estimates are easy to obtain from a training data set with a sufficient number of variables and of course with reliable diagnosis. The term "oligodimensionality" reflects the fact that due to a trade-off between discernment power, robustness and computational complexity the "input" distributions – supposed marginals of the actual distribution – describe relations between 2, 3, 4 or exceptionally 5 random variables. Each variable in its turn takes values from a set of several elements only. This explanation should justify the abbreviation "oligodistribution" in the sequel. In INES, the decision itself is based not directly on the actual distribution but on its restrictions. The success of the whole approach was strongly dependent on efficient and fast algorithms to achieve the restriction within seconds. The task was accomplished for different classes of approximating distributions [7] and since then the discernment power of INES has been tested on several occasions [2]. Though the original idea to test continually the theoretically guaranteed merits of INES in a direct confrontation with an extensional system failed, because of some technical difficulties, the necessity to "measure" the power of INES by comparing it with an efficient sparring partner over the same knowledge base remained. The DINES system (D-stands for degenerated or direct) is the answer to the demand. "Direct" intimates that by using conditional oligodistributions right from the start we formally bypass generating the joint distribution and the subsequent restrictions. "Degenerated" reflects the fact, that due to heuristics used, certain assumptions of the original INES are simplified.

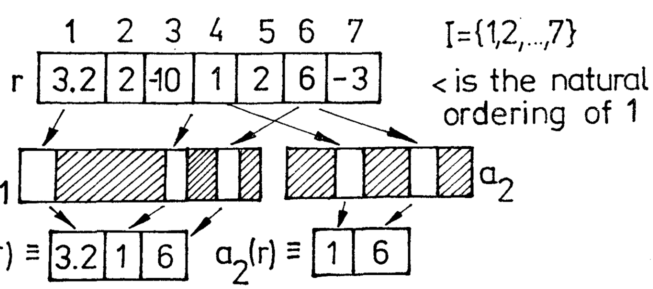
The system INES consists of three distinct parts (program modules): MODIFIER, INTEGRATOR, APPLICATOR. The module MODIFIER accepts knowledge in all admissible forms (see [4]) and creates as its output a set of consistent oligodistributions. This set is passed over to the module INTEGRATOR, that generates structures necessary to construct the approximating joint distribution. These two modules, that are in fact building the problem knowledge base, are supposed to work off-line for hours of CPU time. The situation is different with the module APPLICATOR, that has to respond within seconds and to supply conditional probabilities of diagnoses for given evidence (symptoms).

To be able to formulate our description in a more precise way let us introduce several conventions. The components of an n-dimensional vector r are denoted $(r)_l$ for $l = 1, 2, \ldots\ldots, n = \dim(r)$. Further let a pair $(I, <)$ stand for an ordered set where $I \subset \mathbb{N}$ is a set of positive integers and $<$ is the symbol for its complete ordering (not necessarily the natural one). Let $\pi_i((I, <))$ denote the ith element from I according to ordering $<$. (π_i is thus the usual projector taking the ith component from elements of a linear space). For a fixed n let there exist a nonempty finite subset T of a vector space $\mathbb{R}^n$: $T \neq \emptyset$, $|T| < +\infty$, $T \subset \mathbb{R}^n$. It is possible to define a mapping A

$$A\colon \mathscr{P}((\{1, 2, \ldots, n\}, <)) \times T \to \mathbb{R} \cup \mathbb{R}^2 \cup \mathbb{R}^3 \ldots \cup \mathbb{R}^n$$

$$((I, <), r) \mapsto ((r)_{\pi_1((I,<))}, (r)_{\pi_2((I,<))}, \ldots, (r)_{\pi_{|I|}((I,<))}),$$

where $\mathscr{P}(M)$ is the potential set of a set M. Let us now consider a system of ordered sets $\{(I_j, <)\}_{j=1}^m$. With the aid of the mapping A we can construct a set of the corresponding functions $\{A((I_j, <), \cdot)\}_{j=1}^m = \{a_j(\cdot)\}_{j=1}^m$. Inspired by the following example we shall call the functions a_j: $T \to \mathbb{R} \cup \mathbb{R}^2 \ldots \cup \mathbb{R}^n$ apertures (or windows).



To indicate inversely the ordered set of variables that corresponds to any particular aperture a_j we shall introduce the function "support of aperture" denoted with sign$^{\sim}$

$$\sim : \{A((I_j, <), \cdot)\}_{j=1}^{m} \to \mathscr{P}(\{1, 2, \dots, n\})$$
$$\tilde{a}_j(\cdot) = \tilde{A}((I_j, <), \cdot) \mapsto (I_j, <)$$

For the above example $\tilde{a}_1 = (1, 4, 6)$ $\tilde{a}_2 = (4, 6)$.

The main interpretation of the introduced notions is the following: For each respondent (or patient) r of a training set T aperture a (corresponding to a question or test) is a function that assembles a vector $a(r)$ (answers to the question a about patient r). In the vector $a(r)$ there will appear only values of those variables from the vector $r \in T$ that are described by the support $\tilde{a}$ of the aperture a ordered conform to initial fixed ordering! Now, we may formally describe the function of the INTEGRATOR and APPLICATOR modules. The task of the INTEGRATOR is to create an approximation $\bar{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ of the all-explaining unknown joint distribution $P_{\eta\xi_1\xi_2\cdots\xi_n}$. The approximation $\bar{P}_{\eta\xi_1\xi_2\cdots\xi_n}$ can be thought of as certain function I realised by the algorithm of INTEGRATOR and parametrized by all "input" oligodistributions $o_j, j = 1, \dots, m$

(1) $$\bar{P}_{\eta\xi_1\xi_2\cdots\xi_n}(d, x_1, \dots, x_n) = I(o_1, o_2, \dots, o_m)\,(d, x_1, x_2, \dots, x_n)$$

The APPLICATOR module performs the restruction $\bar{P}_{\eta\xi_1\xi_2\cdots\xi_n} \to \bar{P}^{\tilde{a}}_{\eta\xi_1\xi_2\dots\xi_n}$ (the upper indexing stands for the set of variables for which the restriction takes place). The restricted distribution is then evaluated for the values that the aperture a passes to the APPLICATOR for a fixed respondent $r \in T$. We shall assume that final decision for all types of probabilistic expert systems is performed on the basis of a vector $(E_a(P_{d_1|r}), E_a(P_{d_2|r}), \dots, E_a(P_{d_n|r}))$, where $E_a(P_{d_i|r})$ is an estimation of conditional probability $P_{d_i|r}$ of a diagnosis d_i for a respondent r if only values of variables from $\tilde{a}$ are known. $E_a(P_{d_i|r})$ is defined for INES

$$E_a(P_{d_i|r}) = \frac{\bar{P}_{\eta\tilde{a}(d_i, a(r))}}{\bar{P}_{\tilde{a}(a(r))}} \quad i = 1, 2, \dots |\eta|$$

where $\bar{P}_{\eta\tilde{a}}$ resp. $\bar{P}_{\tilde{a}}$ is a shorthand for $\bar{P}^{\eta\tilde{a}}_{\eta\xi_1\xi_2\dots\xi_n}(d_i, a(r))$ resp. $\bar{P}^{\tilde{a}}_{\eta\xi_1\xi_2\dots\xi_n}(a(r))$. For the denominator $\bar{P}_{\tilde{a}}(a(r)) = 0$ we set the result identically to zero. Having in mind (1) we may now transcribe $E_a(P_{d_i|r})$ for INES as

$$E_a(P_{d_i|r}) = \frac{I_{\eta\tilde{a}}(o_1, \dots, o_m)\,(d_i, a(r))}{I_{\tilde{a}}(o_1, \dots, o_m)\,(a(r))}$$

2. MAIN IDEA OF THE DINES SYSTEM

The main object of this paper is to suggest a new inference machine DINES where $P_{d_i|r}$ is estimated in a slightly different way

$$E_a(P_{d_i|r}) = D(E_{a,o_1}(P_{d_i|r}), E_{a,o_2}(P_{d_i|r}), \dots, E_{a,o_m}(P_{d_i|r}))$$

where D is a certain global function (implemented by DINES) $D: \mathbb{R}^m \to \langle 0, 1 \rangle$ and the symbol $E_{a,o_j}(P_{d_i|r})$ stands for estimation of probability $P_{d_i|r}$ that a person $r \in T \subset \mathbb{R}^n$ suffers from a disease d_i provided only values of variables from $\tilde{a}$ are known and supposing that all our "knowledge base" consists of oligodistribution o_j only. Though retaining the symbol o_j, the original notion oligodistribution will be slightly changed in context of DINES. To o_j (originally l-dimensional matrix for $l = |\tilde{o}_j|$) k other objects of the same form and size, denoted by $o_{j|d_i}$, are added to create more complex structure

$$\mathbf{o}_j = (o_j, o_{j|d_1}, o_{j|d_2}, \ldots, o_{j|d_k}) \quad k = |\eta| .$$

Attention should be paid to the fact that at variance with usual conventions each symbol $o_{j|d_i}(\bar{x})$ denotes $P_{\eta|\tilde{o}_j}(d_i, \bar{x})$ i.e. conditional probability of diagnosis variable η to take value d_i if the symptom variables from $\tilde{o}_j$ take the value $\bar{x}$. The whole structure $\mathbf{o}_j$ has the same support $\tilde{o}_j$. Another difference is that $\tilde{o}_j$ at DINES does not contain the "diagnosis" variable η even if it was included at the original INES input. The link to diagnosis is achieved here by conditioning! Roughly speaking D computes a certain "average" of assesments $E_{a,o_j}(P_{d_i|r})$ of probability $P_{d_i|r}$ by individual conditional oligodistributions $o_{j|d_i}$.

One of the differences between the two approaches might be seen in the fact that INES performs at first the integration of knowledge from input oligodistributions and then restriction and evaluation, DINES restricts and evaluates (conditional distributions $o_{j|d_i}$) and then "integrates" the resulting numbers via the function D.

3. CONTRIBUTIONS TO DECISION BY INDIVIDUAL OLIGODISTRIBUTIONS

Let us now define the way the numbers $E_{a,o_j}(P_{d_i|r})$, that appear as arguments in the function D, are constructed. It is in principle always based on the relation between the supports of the aperture a and the oligodistribution o_j.

1. The simplest case arises when for a given aperture a there exists an oligodistribution structure

$$\mathbf{o}_j = (o_j, o_{j|d_1}, o_{j|d_2}, \ldots, o_{j|d_n})$$

such that the support of o_j equals the support of the aperture a:

$$\tilde{o}_j = \tilde{a} \tag{2}$$

$E_{a,o_j}(P_{d_i|r})$ is then just $o_{j|d_i}(a(r))$ meaning a simple looking-up in the corresponding $o_{j|d_i}$ of the structure $\mathbf{o}_j$ with "address" $a(r)$. (Components of patient vector r are supposed to be discretized and coded into integers so that a really maps to vectors of integers.) If (2) holds for more oligodistributions (acquired from different informational sources) we may comprime them as a convex combination and interprete the "weights" as apriori distribution over sources or as our subjective assesment of their credibility.

2. The second case corresponds to the situation that for a given aperture a we find such oligodistribution o_j that for their supports there holds following set inclusion

$$\tilde{a} \subset \tilde{o}_j$$

then the natural choice for $E_{a,o_j}(P_{d_i|r})$ is $o^{\tilde{a}}_{j|d_i}(a(r))/o^{\tilde{a}}_j(a(r))$ if defined or zero otherwise where

$$o^{\tilde{a}}_{j|d_i}(a(r)) = \sum_{\tilde{o}_j/\tilde{a}} o_{j|d_i}(\cdot, a(r)) \cdot o_j(\cdot, a(r))$$

and

$$o^{\tilde{a}}_j(a(r)) = \sum_{\tilde{o}_j/\tilde{a}} o_j(\cdot, a(r))$$

This symbolic notation describes averaging conditional probabilities from $o_{j|d_i}$ – i.e. the probabilities of d_i conditioned by $\tilde{o}_j$. Dots inside brackets stand for values of those variables over which the summation takes place. In this particular case we sum over all variables ξ from $\tilde{o}_j$ that are not in the support of a.

3. The most general and most often encountered situation is when an aperture a and oligodistribution o_j have a nonempty intersection of their supports and non empty differences

$$\tilde{a} \cap \tilde{o}_j \neq 0\,, \quad \tilde{a}/\tilde{o}_j \neq 0\,, \quad \tilde{o}_j/\tilde{a} \neq 0\,,$$

$E_{a,o_j}(P_{d_i|r})$ is then defined as

$$o^{\tilde{a}\cap\tilde{o}_j}_{j|d_i}(a^{\tilde{a}\cap\tilde{o}_j}(r))/o^{\tilde{a}\cap\tilde{o}_j}_j(a^{\tilde{a}\cap\tilde{o}_j}(r))$$

for non-zero denominator or zero otherwise. Where

$$o^{\tilde{a}\cap\tilde{o}_j}_{j|d_i}(a^{\tilde{a}\cap\tilde{o}_j}(r)) = \sum_{\tilde{o}_j/\tilde{a}} o_{j|d_i}(\cdot, a^{\tilde{a}\cap\tilde{o}_j}(r)) \cdot o_j(\cdot, a^{\tilde{a}\cap\tilde{o}_j}(r)) \tag{3}$$

and

$$o^{\tilde{a}\cap\tilde{o}_j}_j(a^{\tilde{a}\cap\tilde{o}_j}(r)) = \sum_{\tilde{o}_j/\tilde{a}} o_j(\cdot, a^{\tilde{a}\cap\tilde{o}_j}(r))$$

4. If $\tilde{a} \cap \tilde{o}_j = 0$ then applying the fact to formula (3) we sum and normalize over the whole $o_{j|d_i}$ getting thus an averaged conditional oligodistribution irrespective of measured symptoms and this corresponds to aprioristic occurrence of different diagnoses in the training data set from which the knowledge base was generated.

4. GLOBAL COMBINING FUNCTION D

After introducing the construction of the numbers $E_{a,o_j}(P_{d_i|r})$ for an arbitrary pair (a, o_j), we may return to the question of "integrating" the $\{E_{a,o_j}(P_{d_i|r})\}_{j=1}^{m}$ to a single $E_a(P_{d_i|r})$ via the global function D.

Before giving a final answer let us briefly mention some possibilities:

a) to apply certain statistical estimator (after having made some supposition about the family of distributions where we believe to find the actual joint distribution;

b) barycenter approach as proposed by Perez [8];

c) to understand the construction of D as a problem of multicriterial decision with certain Pareto optimal set;
d) it is possible at least formally (if we can justify it) to apply as D some combining functions used in rule-based systems;
e) to apply estimators of the averaging type (arithmetical, geometrical, harmonical etc.);
f) to consider chances of applying gnostical theory (cf. [3]).

The question seems to be still open and will be subject to further experiments. Let us only stress the following idea: It is not the actual value of $P_{d_i|r}$ probability for different diagnoses but their ratio only (with certain threshold strategy) that matters. This ratio should be the invariant of our "knowledge geometry" and this fact may bring about as its consequence more simple and robust forms for the function D. As the first prototype of the global function D for DINES systems we have adopted the following heuristical approach:

First some more conventions: $R(I)$ denotes an arbitrary partition of a set I, $|R(I)|$ is the number of its components and for each component $S_i \in R(I)$ $|S_i|$ is the number of elements S_i consists of.

Let us have an aperture a and let us consider different partition $R(\tilde{a})$ of the set $\tilde{a}$ fulfilling

a) $$\forall_{S_i \in R(\tilde{a})} \; \exists_{o_j \in O} \; S_i = \tilde{a} \cap \tilde{o}_j$$

Components S_i of the partition are thus assembled only from such oligodistributions o_j that "coincide" with the aperture a

b) $$\frac{1}{|R(\tilde{a})|} \sum_{i=1}^{|R(\tilde{a})|} |\tilde{a} \cap \tilde{o}_{S_i}| \to \max$$

where $o_{S_i} = o_j^{S_i}$ for o_j generating S_i: $S_i = \tilde{a} \cap \tilde{o}_j$. The condition b) favours the partitions $R(\tilde{a})$ with a small number of "large" components. In the sequel, the symbol $\mathscr{R}(\tilde{a})$ will be used for all partitions $R(\tilde{a})$ of $\tilde{a}$ with the mentioned properties a) and b). Now we can define

(4) $$E_a(P_{d_i|r}) = \max_{R(\tilde{a}) \in \mathscr{R}(\tilde{a})} \{ (\prod_{j=1}^{|R(\tilde{a})|} \max \{ E_{a,o_{S_j}}(P_{d_i|r}), \varepsilon f(S_j) \})^{1/|R(\tilde{a})|} \}$$

where $f(S_j)$ may be chosen e.g. as

(5) $$|S_j| - 1$$

or

(6) $$\prod_{\xi_k \in S_j} |\xi_k| / |T|$$

The formula (4) for the DINES global function D needs some comment. In fact, we recommend the geometrical mean of estimations $E_{a,o_{S_j}}(P_{d_i|r})$ where the components S_j of the partition $R(\tilde{a})$ are in a sense the largest possible ones. The internal maximization along with f and ε guarantees non zero values of $E_a(P_{d_i|r})$ even for degenerated o_{S_j}. The variant (5) requires exact zeroes only if one-dimensional margi-