ELSEVIER International Journal of Medical Informatics 45 (1997) 65–74

International Journal of
Medical
Informatics

Information theoretical approach to constitution and reduction of medical data

Jana Zvárová [a],*, Milan Studený [b]

[a] *EuroMISE Center, Charles University and Academy of Sciences of Czech Republic, Pod vodarenskou věží 2,*
182 07 Prague, Czech Republic

[b] *Institute of Information Theory and Automation, Academy of Sciences of Czech Republic, Pod vodarenskou věží 4,*
182 08 Prague, Czech Republic

Abstract

In medical decision problems it is very important to use the most relevant piece of information for decision making. We focus on a special case of diagnostic decision making when we can measure many symptoms and signs and we have to make diagnostic conclusions. We can state the problem as follows. We can measure symptoms and signs of a patient, denoted by $s_1, s_2,..., s_k$, and we have to decide about a possible diagnosis d. We know that the symptoms and signs have different costs $w_1, w_2,...w_k$ when they are examined. Of course, each symptom, sign or their combination has a different predictive value for the diagnosis. Our task is to find out the combination of symptoms from given data with a sufficient informative value for diagnostic decision making. However, simultaneously we look for a combination of symptoms and signs with minimal costs among those carrying sufficient information. For that reason we will describe approaches based on information measures of statistical dependence and to show the idea of the program CORE (constitution and reduction of data) prepared for practical applications in medicine. © 1997 Elsevier Science Ireland Ltd.

Keywords: Constitution of medical data; Reduction of medical data; Measure of statistical dependence; Multiinformation; Conditional independence

1. Introduction

A special problem in medical decision making occurs when a decision-maker has too much unstructured empirical information at his/her disposal. The model situation is a large database concerning previous patients and involving many symptoms and signs, where some of them may have no influence on a concrete diagnostic task concerning a new patient. In fact, it is a special case of a general problem of choice of a relevant piece

* Corresponding author. E-mail: zvarova@uivt.cas.cz

1386-5056/97/$17.00 © 1997 Elsevier Science Ireland Ltd. All rights reserved.
PII S0020-7101(97)00036-6

of information for decision making. This general problem has appeared also in information theory, where some tools used for solving matters were developed. We have in mind various information–theoretical measures of mutual information, statistical dependence or conditional statistical dependence. In this paper we show that these measures of dependence can be applied also in medical decision making.

In the following section we will describe a general diagnostic situation in which our method can be used. We will concentrate on two problems, described in Section 3. The first problem is the *constitution of data*, i.e. the problem which a combination of symptoms from a given database has sufficient information value for diagnostic decision making. The second problem is the *reduction of data*, i.e. the problem of how to remove redundant pieces of information that are sometimes caused by mutual dependence among relevant symptoms. In the fourth section we will recall some concepts from information theory, namely definitions and properties of several concrete measures of dependence and conditional dependence. Algorithms for constitution and reduction of data based on these concepts will then be described on a theoretical level in Section 5. Then we show the idea of the computer program CORE (constitution and reduction of data) prepared for practical application in medicine. In the last section of the paper we will outline another possible use of the above mentioned information–theoretical measures of dependence. They can be utilized for estimating qualitative models of conditional independence for small groups of variables, i.e. for extraction of qualitative information from data.

2. General description of the considered situation

Let us describe the situation we have in mind. We already mentioned that diagnoses will be based on some measured symptoms or signs. Our problem is how to choose relevant symptoms for such decision making.

Let us specify our assumptions more concretely. We will suppose that a big set of *symptom variables* $S = \{s_1,\ldots, s_k\}$, $k \geq 1$ is given. Each symptom variable has finitely many, but at least two possible values. The values can be both quantitative (for example the scale of temperature) but also qualitative (i.e. presence or absence of a certain factor). Moreover, we will suppose that each symptom $t \in S$ is assigned a certain nonnegative weight $w_t \geq 0$, describing the cost of obtaining the value of the variable t. For instance, it can reflect monetary expenses of the corresponding test. Nevertheless, a more general point of view on the cost can also be taken: certain invasive methods for obtaining data can be painful or risky and therefore the obtained symptom variable should be considered as costly.

Further our assumption is that a set of *decision variables* $D = \{d_1,\ldots, d_m\}$, $m \geq 1$, disjoint with S, is given. Every decision variable should correspond to a concrete diagnostic hypothesis, that means its values should represent possible outcomes of the decision-making procedure about the diagnostic hypothesis. Thus, in case that the decision variable corresponds to a simple diagnostic hypothesis (for example that the patient has tonsillitis) the variable has only two possible values: YES or NO. However, one can consider also a complex hypothesis and in this case decision variables can have more than two possible values. For example, a patient can have rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), ankylosing spondylitis (AS) or none of these diseases (NO). Then the decision variable has four values: RA, SLE, AS, NO. The set of all considered variables $D \cup S$ will be denoted by V.

The decision making will be based on previous observations. Thus, our starting point will be a large *data matrix* $\mathcal{M}$ with n rows (representing previous observations) and $m + k$; columns. Here, each column corresponds to a variable in V and the element of $\mathcal{M}$ in the ith row $(1 \leq i \leq n)$ and in the column corresponding to a variable $v \in V$, denoted by x^i_v, is the value of v obtained by the ith observation. Our idea is that one observation represents one previous patient, whose diagnoses (i.e. values of decision variables) were indubitable and where values of symptom variables were obtained by measurement. Of course, the number of observations n should be so high that one can expect that relationships among variables are reflected in the data matrix. The question whether a data matrix has sufficient information value is delicate and should be answered by an experienced physician or statistician.

The data matrix can serve as a source for a *frequency estimate* of the 'underlying' probability distribution. Namely, for every vector of possible values of considered variables $[y_v]_{v \in V}$ one can compute the relative frequency of its occurrence as a row of the data matrix $\mathcal{M}$, i.e.

$$[y_v]_{v \in V} \mapsto n^{-1} \cdot \mathrm{card}\{i; \quad 1 \leq i \leq n \quad \text{and} \quad [x^i_v = y_v \text{ for all } v \in V]\}.$$

This defines a probability distribution over V. Note that all algorithms described in this paper are based on this distribution, more exactly on information measures of stochastic dependence and conditional stochastic dependence computed for marginals of that frequential estimate. Perhaps one can use another method of estimation of the 'underlying' probability distribution and apply the same algorithms with the only difference that the information measures will be computed for these other estimates.

3. Problem description

Now, let us assume that a physician has a diagnostic hypothesis or a collection of diagnostic hypotheses concerning a new patient. We would like to verify it and therefore we need to know which symptoms or signs are relevant to such decision making. More formally, a set of decision variables $Y \subset D$, called the set of *dependent variables* is given (perhaps Y has just one decision variable). Our task is to find out whether it is possible to make a justified decision concerning variables in Y on the basis of symptom variables in S. If yes, we should find a relatively small set of symptom variables $X \subset S$, called the set of *independent variables*, such that a strong stochastic dependence between X and Y allows us to estimate with high credibility probabilities of values of variables in Y on the basis of values of variables in X. The choice of X should take into account the cost of obtaining values, i.e. $\Sigma_{s \in X} w_s$ should be as low as possible. A more concrete example of actual variables from medicine will be given in Section 6. Sometimes a physician has already indicated symptoms and signs which he/she thinks are 'relevant' to the considered diagnostic decision task. Our method can measure their relevance 'objectively' on basis of medical data. We have experience that sometimes the selected symptoms and signs are not relevant to the considered decision task (because they do not bring sufficient relevant information). Thus, we face the problem of *constitution of data*. That is, we should answer the question whether symptom variables together have a sufficient information value for Y. Therefore we can search for additional symptoms and signs that will have sufficient information value for the decision task in a statistical sense (for details see [1]).

The second step, called the *reduction of data* starts with the set of symptom variables S. However, owing to possible strong mutual dependencies among variables in S perhaps some symptom variables having information value for Y can be omitted because the other variables in S (namely the variables with strong mutual dependency with the omitted variables) may keep that information value. Thus, the result of the procedure should be a set of independent variables $X \subset S$ with sufficient information value for Y and low cost. The set with minimal cost among sets having sufficient information value would be an ideal solution.

Proper decision-making procedures should be based on symptom variables in X. We have the following idea how to perform it. First, patient values of symptom variables in X will be obtained by corresponding medical examinations. They form a vector of values $[z_v]_{v\in X}$ of independent variables. Second, for each vector of possible values $[y_v]_{v\in Y}$ of dependent variables one computes the conditional probability of $[y_v]_{v\in Y}$ given $[z_v]_{v\in X}$. More exactly, one computes an estimate of that conditional probability, since the basis of the computation is the frequency estimate of the 'underlying' probability distribution, mentioned in the previous section. Thus, the number

$$\hat{p}_{Y|X}([y_v]_{v\in Y}|[z_v]_{v\in X}) = \frac{\text{card}\,\{i;\quad x_v^i = z_v \text{ for all } v\in X \quad \text{and} \quad x_v^i = y_v \text{ for all } v\in Y\}}{\text{card}\,\{i;\quad x_v^i = z_v \text{ for all } v\in X\}}$$

is our estimate of the probability that $[y_v]_{v\in Y}$ is the vector of patient's values of dependent variables. Acceptance or rejection of the original physician's hypothesis could be based on these estimates.

4. Information–theoretical measures of statistical dependence

Our approach is based on information measures of stochastic dependence and of conditional stochastic dependence. Roughly speaking, information measures are nonnegative numerical characteristics of the strength of stochastic dependence between two variables (respectively the strength of conditional dependence between two variables given values of a third variable). A basic requirement the information measure should fulfil is that it is zero if and only if the corresponding random variables are independent (or conditionally independent). These measures have been developed and studied in information theory as tools to estimate the Bayes risk.

Important properties of adequate measures of dependence have been pointed out already in the 60s by A. Perez [2,3], one of the founders of the Czech school of information theory. Mainly, the measure of dependence based on the classic Shannon's information was studied, but also other measures, based on the general concept of f-information were proposed by I. Vajda [8,9]. However, in this paper we will not deal with this general concept of f-information but with its special case, i.e. the classic Shannon's information. Behaviour and suitability of different measures of stochastic dependence were later studied by J. Zvárová [10,11]. Namely, for a large class of information measures it was shown that they attain their maximal values if and only if so-called c-dependency occurs (for details see [11]), which is often equivalent to strict functional (= deterministic) dependency of variables. The concept of multi-information, introduced as a measure of simultaneous dependence, was studied by M. Studený [5,6]. It was shown that the multi information function is closely connected to Shannon's conditional mutual information, which serves as a measure of conditional stochastic dependence. Application of information measures mainly in connection with decision support in medicine was given for example in papers [12,13].

Let us recall definitions of several information measures for the discrete case that we utilize in this paper. Supposing V is a nonempty finite set of variables, for every variable $v \in V$ the symbol $\mathbf{R}_v$ denotes a finite nonempty set of its possible values. For every $\emptyset \neq A \subset V$ let us denote by $\mathbf{R}_A$ the cartesian product $\Pi_{v \in V}$ $\mathbf{R}_v$. For example, $\mathbf{R}_V$ denotes $\Pi_{v \in V}$ $\mathbf{R}_v$ and $\mathbf{R}_{V \setminus A} = \Pi_{v \in V \setminus \mathrm{A}}$ $\mathbf{R}_v$ where $V \setminus A$ is the set of variables belonging to V but not to A.

Let P be a probability distribution on $\mathbf{R}_V$. Having $\emptyset \neq A \subset V$, the marginal distribution of P on $\mathbf{R}_A$ denoted by P^A is defined by the formula

$$P^A(y) = \sum \{P(y, z); \quad z \in \mathbf{R}_{V \setminus A}\}$$

for every $y \in \mathbf{R}_A$.

Note that P^V is simply P and P^i is a shortened form of $P^{\{i\}}$ where $i \in V$. As $\Sigma\{P(z); z \in \mathbf{R}_{V \setminus \emptyset}\} = \Sigma\{P(z); z \in \mathbf{R}_V\}$ is 1 by the definition of probability distribution we accept a natural convention that $P^{\emptyset}(-) = 1$ in the forthcoming formulas.

Having $\emptyset \neq A \subset V$, the *entropy* $H(A)$ is defined by the formula

$$H(A) = \sum \{P^A(x) \cdot \ln \frac{1}{P^A(x)}; \quad x \in R_A \quad \text{and} \quad P^A(x) > 0\},$$

with the convention $\mathrm{H}(\emptyset) = 0$ and the *multi-information* $M(A)$ by the formula

$$M(A) = \sum P^A([x_v]_{v \in A}) \cdot \ln \frac{P^A([x_v]_{v \in A})}{\Pi_{i \in A} P^i(x_i)};$$

where $x_v \in \mathbf{R}_v$ for every $v \in A$ and $P^A([x_v]_{v \in A}) > 0$, with a similar convention $M(\emptyset) = 0$. Both these functions on the power set of V are nonnegative.

For every couple of disjoint sets A, $B \subset V$ *Shannon's mutual information* is the relative entropy of $P^{\mathrm{A} \cup \mathrm{B}}$ with respect to the product of P^A and P^B:

$$I(A; B) = \sum \{P^{A \cup B}(a, b) \cdot \ln \frac{P^{A \cup B}(a, b)}{P^A(a) \cdot P^B(b)}; \quad a \in R_A \quad \text{and} \quad P^{A \cup B}(a, b) > 0\}.$$

Note that (see [11]) $0 \leq I(A; B) \leq \min\{H(A), H(B)\}$ and therefore *Shannon's information measure* of A on B, defined by

$$\delta(A/B) = I(A; B)/H(A)$$

is always a real number between 0 and 1. Note that in our algorithms we use this measure of information as the criterion of whether data has sufficient information value for considered decision variables.

For every triplet of disjoint sets A, B, $C \subset V$ *Shannon's conditional mutual information* is defined by

$$I_c(A; B/C) = \sum \{P^{A \cup B \cup C}(a, b, c) \cdot \ln \frac{P^{A \cup B \cup C}(a, b, c) \cdot P^C(c)}{P^{A \cup C}(a, c) \cdot P^{B \cup C}(b, c)}; \quad a \in \mathbf{R}_A \quad \text{and} \quad b \in \mathbf{R}_B \quad \text{and} \quad c \in \mathbf{R}_C \quad \text{and} \quad P^{A \cup B \cup C}(a, b, c) > 0\}.$$

Note that $I_c(A; B \setminus \emptyset) = I(A; B)$ and moreover it holds (see [6]).

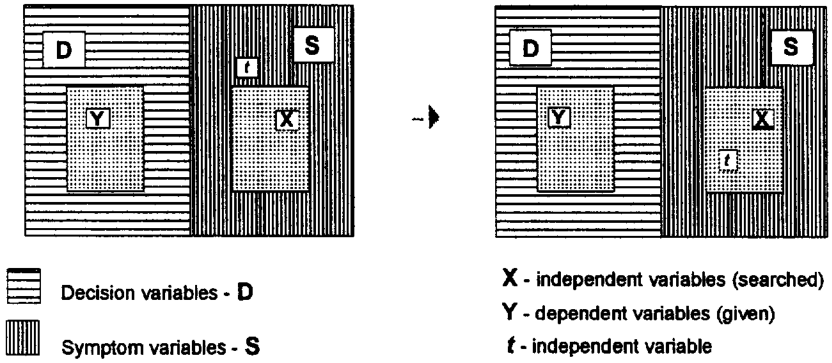

Fig. 1. Forward algorithm.

$$I_c(A; B/C) = M(A \cup B \cup C) + M(C) - M(A \cup C) - M(B \cup C)$$

5. Algorithms

Our algorithms for finding the set of independent variables $X \subset S$ for a given set of dependent variables $Y \subset D$ are not algorithms for finding the solution of a formally precise minimization problem. They have rather a heuristic nature, the choice of variables in X is based on conditional measures of influence between Y and symptom variables which have some interpretation. All our algorithms have in common the stop-criterion which is based on Shannon's information measure of stochastic dependence $\delta(Y|X)$.

In principle, there is no need to distinguish between algorithms for the constitution of data and algorithms for reduction of data: they both solve the same mathematical problem of choosing a subset of a given set of symptom variables. The only cosmetic difference is that the result of the constitution of data may be the conclusion that symptom variables have not sufficient information value for Y. This should not happen in case of the reduction of data since the reduction procedure should start with the result of constitution procedure, which has a sufficient information value. However, algorithms are based on different heuristics and therefore, we consider some of them more suitable for constitution and some of them for reduction. Each algorithm has a specific *score function* which assigns to a set of decision variables $Y \subset D$, to a set of symptom variables $X \subset S$ and to a symptom variable $t \in S$ a nonnegative number $\kappa(Y, X, t)$ generally interpreted as measure of influence between Y and the single variable t under knowledge of X. Score functions we use are defined by means of information-theoretical characteristics mentioned in the previous section and each of them has special interpretation.

We can classify algorithms as *forward algorithms* where one starts with the empty set of symptom variables and adds variables, *backward algorithms* where one starts with the whole original set of symptom variables S and removes variables, and *combined algorithms* where after application of the forward procedure the backward procedure is used and conversely (see Figs. 1 and 2 for illustration).

The algorithms can also be classified differently. If we take influence among variables as the primary criterion for choice of relevant

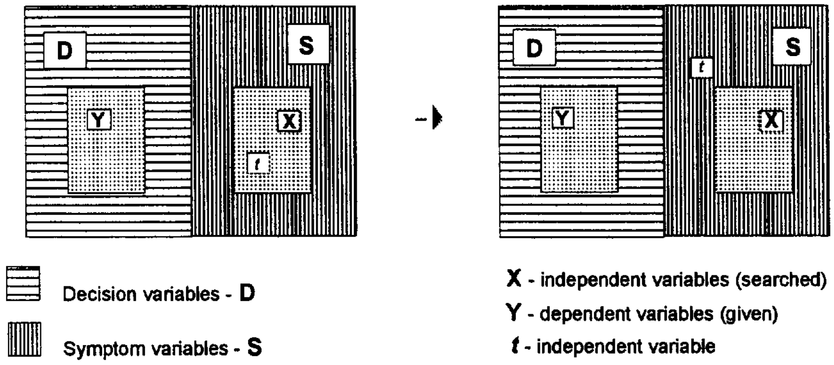

Fig. 2. Backward algorithm

symptom variables and the cost (that is the weight of symptom variables as mentioned in Section 2) as the secondary criterion, then we speak about an *influence-preferring algorithm*. However, if we take the weight w_t of a chosen symptom variable t as the primary criterion and influence among variables as the secondary criterion, then we speak about a *weight-preferring algorithm*.

Our algorithms can be classified according to the primary criterion, that is to influence-preferring and weight-preferring algorithms, and according to the way of forming of X, that is to forward, backward and combined algorithms.

Concerning the *influence-preferring forward* algorithm we offer three possible score functions, applied in case that t is outside of X:

- $\kappa(Y, X, t) = I_c(Y; \{t\}|X)$ is the measure of conditional dependence between Y and t under condition that X is known,
- $\kappa(Y, X, t) = I(Y; X \cup \{t\})$is the measure of dependence between Y and $X \cup \{t\}$,
- $\kappa(Y, X, t) = I(Y; \{t\})$ is the measure of dependence between Y and t.

Moreover, the algorithm uses a parameter $\delta_0 \in (0, 1)$ which should be close to 1. The procedure starts by putting $X_0 = \emptyset$, $Z_0 = S$. One step of the procedure can be described as follows: supposing that X_i, $Z_i \subset S$ are disjoint sets with $X_i \cup Z_i = S$ one finds $t \in Z_i$ such that

- t maximizes $\kappa(Y, X_i, t)$ within Z_i, and
- w_t is minimal within the set of variables from Z_i maximizing $\kappa(Y, X_i, t)$.

No further condition on t is required. Whenever two or more variables comply with both conditions, we arbitrarily select one variable. Then we put $X_{i+1} = X_i \cup \{t\}$, $Z_{i+1} = Z_i \backslash \{t\}$.

If $\delta(Y/X_{i+1}) > \delta_0$, then the procedure stops and the set $X = X_{i+1}$ is the result of the algorithm. If $\delta(Y/X_{i+1}) \leq \delta_0$ and $i + 1 <$ card S, then we repeat the step with X_{i+1} and Z_{i+1}. If $\delta(Y|X_{i+1}) \leq \delta_0$ and $i + 1 =$ card S, then the procedure stops with the conclusion $\delta(Y|S) \leq \delta_0$, i.e. S has not sufficient information value for Y.

Score functions used in the *influence-preferring backward* algorithm are applied in case that t belongs to X:

- $\kappa(Y, X, t) = I_c(Y; \{t\}|X \backslash \{t\})$ is the measure of conditional dependence between Y and t under condition that knowledge about $X \backslash \{t\}$ remain known,
- $\kappa(Y, X, t) = I(Y; X) - I(Y; X \backslash \{t\})$ is the decrease of measure of dependence between Y and X, when t is removed from X.

Except the score function the algorithm has a parameter $\delta_0 \in (0, 1)$ close to 1. The procedure starts by putting $X_0 = S$, $Z_0 = \emptyset$. Supposing X_i, Z_i is a decomposition of S we compute the value of $\delta(Y/X_i)$. If $\delta(Y/X_0) \leq \delta_0$, then the procedure stops with the conclusion that S has not sufficient information value for Y. If $i \geq 1$ and $\delta(Y/X_i) \leq \delta_0$, then the procedure stops and the set $X = X_{i\text{-}1}$ is the result of the algorithm. If $i \geq 1$ and $\delta(Y/X_i) > \delta_0$, then we find arbitrary $t \in X_i$ such that

- t minimizes $\kappa(Y, X_i, t)$ within X_i, and
- w_t is maximal within the set of variables from X_i minimizing $\kappa(Y, X_i, t)$.

and put $X_{i+1} = X_i \setminus \{t\}$, $Z_{i+1} = Z_i \cup \{t\}$ and repeat the step with X_{i+1} and Z_{i+1}.

Influence-preferring combined algorithms can be obtained when the procedure starts as forward algorithm and on its result the backward algorithm is applied (or conversely).

The *weight-preferring forward* algorithm has the same score functions as the influence-preferring forward algorithm, but also a parameter $\delta_0 \in (0, 1)$ close to 1. Moreover, it has a floating parameter $E \geq 0$. The parameter E has the role of internal threshold (changed during the performance of the algorithm) used to 'determine' whether we will consider a symptom variable t sufficiently influential with respect to Y under knowledge of X (unlike the parameter δ_0 which is used to determine whether the overall information in X is sufficient for Y). That is whenever we reset the value of E we perform the following proper procedure, which gives a set of symptom variables X as result.

Proper procedure: We order all variables of S into a sequence $t_1, \ldots, t_n$ such that weights of variables increase in this sequence. The ordering will remain fixed for future possible use of this proper procedure. The procedure starts by putting $X_0 = \emptyset$. The step of the procedure is simple: suppose X_i is determined and $\kappa(Y, X_i, t_{i+1}) \geq E$ we put $X_{i+1} = X_i \cup \{t_{i+1}\}$, otherwise $X_{i+1} = X_i$. After all n steps we put $X = X_n$.

Note that for $E = 0$ the procedure above gives $\tilde{X} = S$, a higher value of E produces less $\tilde{X}$ and for sufficiently high E is $\tilde{X} = 0$. According to the value of $\delta(Y|\tilde{X})$ we can decide whether we will change the value of the parameter E and repeat the proper procedure or whether we stop the algorithm. Namely, if we start with $E = 0$ and if then $\delta(Y|\tilde{X}) \leq \delta_0$, then we stop the algorithm with the conclusion that S has not sufficient information value for Y. If $\delta(Y|\tilde{X}) > \delta_0$, then we choose a higher value of E and repeat the proper procedure with the new parameter. By gradually raising the value of E we reach the situation when $\delta(Y|\tilde{X}) \leq \delta_0$. Then we again decrease the value of E but not below the previous value of E. Thus, we alternatively decrease and increase the value of E with the aim to find minimal $\tilde{X}$ for which $\delta(Y|\tilde{X}) > \delta_0$ (it must exist since only a finite number of sets $\tilde{X}$ may be generated). Then we restart the proper procedure with modification that after each of its step $\delta(Y|X_{i+1})$ is computed. In the case $\delta(Y|X_{i+1}) > \delta_0$ or $i + 1 = \text{card } S$ we stop the procedure and put $X = X_{i+1}$.

The *weight-preferring backward* algorithm has the same score function as the influence-preferring backward algorithm, parameter $\delta_0 \in (0, 1)$ close to 1 and also uses a floating parameter $E \geq 0$. Like in case of the weight-preferring forward algorithm the proper procedure below assigns to E a set $\tilde{X} \subset S$, where $\tilde{X} = S$ for $E = 0$, higher value of E gives less $\tilde{X}$, and $\tilde{X} = \emptyset$ for sufficiently high E.

Proper procedure: We order variables of S in a sequence $t_1, \ldots, t_n$ with decreasing weights and fix the sequence. The procedure starts by putting $X_0 = S$. Its steps are simple: if $\kappa(Y, X_i, t_{i+1}) < E$, then put $X_{i+1} = X_i \setminus \{t_{i+1}\}$, otherwise $X_{i+1} = X_i$. Finally, we put $\tilde{X} = X_n$.

Thus, we start with $E = 0$ and in case $\delta(Y|\tilde{X}) \leq \delta_0$ we conclude that S has not required information value for Y. Otherwise we let E float, like in the previous algorithm, until we reach minimal $\tilde{X}$ with $\delta(Y|\tilde{X}) > \delta_0$ and then we perform a modified proper procedure, where $\delta(Y|X_{i+1})$ is computed after each step. In case $\delta(Y|X_{i+1}) > \delta_0$ and $i+1 <$ card S we continue with a further step of the procedure. In case $\delta(Y|X_{i+1}) \leq \delta_0$ and $i+1 <$ card S we stop the procedure and put $X = X_i$. In case $i+1 =$ card S we stop the procedure and put $X = X_{i+1}$.

Remark: It is not wise to use the same algorithm for reduction of data as for constitution of data, as it should give the same result (provided we did not change the parameter δ_0). We recommend for constitution the inference-preferring backward algorithm and for reduction the weight-preferring forward algorithm.

6. Program CORE

Some of the above mentioned algorithms have been already implemented in a demo-version of the program CORE which is intended for practical application in medicine. The program is written in MS ACCESS and has data matrix $\mathcal{M}$, described in Section 2 as its input (MS ACCESS is a database system from Microsoft). For every A, $B \subset V$ the program computes entropies $H(A)$, $H(B)$, Shannon's mutual information $I(A; B)$, Höffding's coefficients of statistical dependence and minimum probability of error. Moreover, Shannon's information measure of dependence $\delta(A|B)$, its variance, standard deviation and 95% confidence interval can be calculated. A special subprogram computes values of multiinformation function for subsets of V of small cardinality and then Shannon's conditional mutual information $I_c(A; B|C)$ (for every triplet of disjoint sets A, B, $C \subset V$ with card $A \cup B \cup C \leq 4$) which can be used for testing conditional independence statements. We plan to utilize it for estimating conditional independence models for small groups of variables (see the next section).

Moreover, all influence-preferring algorithms described in the previous section are implemented (for some score functions). The other algorithms will be included in a later version of the program CORE.

7. Estimating of conditional independence models for small groups of variables

In this section we describe another possible application of information–theoretical measures of dependence. We already mentioned in Section 4 that one can express Shannon's conditional mutual information $I_c(A; B|C)$ for every triplet A, B, C of disjoint variable sets by means of the multiinformation function. The number $I_c(A; B|C)$ is a measure of conditional dependence in the sense that it is always nonnegative and is zero if and only if variables in A are conditionally independent of variables in B given C. This relationship has very clear interpretation, often used in probabilistic expert systems [4]: provided we know values of variables in C, the value of variables in B is not relevant to values of variables in A. Therefore, when one is interested in A and knows already C, it is needless to investigate B (this intuition is used in our algorithms with Shannon's conditional mutual information as score function).

The program CORE offers the possibility to estimate this measure of conditional dependence and therefore opens the perspective of testing of conditional independence statements on the basis of data. There are several reasons why we should limit ourselves to small groups of variables—that is up to five variables. First, the frequency estimates of marginal distributions for small number of variables should be more precise than for more variables, which demand more data for the same accuracy level. Second, the number of models of conditional independence structure superexponentially increases with number of variables. The case of 4 variables is on limit of effective handling (in the case of 4 variables the number of possible structures of conditional independence is around 18 000—see [7]). However, for its representation in a computer one needs in the case of 3 variables 6 bits, in the case of 4 variables 24 bits and in the case of 5 variables 80 bits. Third, it is known that the human brain is able to combine only a few facts simultaneously. Since we are interested in estimation of overall model of conditional independence structure, models involving a lot of variables loose justification.

The above mentioned overall models of conditional independence structure have often a very concrete interpretation. In the area of probabilistic expert systems mainly graphs are used to describe the structure and experts are asked to draw graphs when one needs to elicit structural information from experts. The purpose of estimating conditional independence from data is that one can check whether expert's statements about a structure are consonant with empirical data. Or conversely, one can elicit from data information about structure which can be later either confirmed or refused by experts. Anyway, measures of conditional dependence provide a method how to obtain from unstructured quantitative information (data) qualitative information (structure).

Acknowledgements

The first author was supported from the projects COPERNICUS JRP-10053 and GA ČR 313/93/0616, the second author from the projects CEC CIPA3511CT930053, GA ČR 201/94/0471, and MŠMT VS96008.

References

[1] A. Perez, Information, ε-sufficiency and data reduction problems, Kybernetika 1 (1965) 297–323.

[2] A. Perez, Information-theoretical risk estimates in statistical decision, Kybernetika 3 (1967) 11–21.

[3] A. Perez, ε-admissible simplifications of the dependence structure of a set of random variables, Kybernetika 13 (1977) 439–449.

[4] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan-Kaufmann, San Mateo CA, 1988.

[5] M. Studený, Asymptotic behaviour of empirical multiinformation, Kybernetika 23 (1987) 124–135.

[6] M. Studený, Multiinformation and the problem of characterization of conditional independence relations, Probl. Control Info. Theory 18 (1989) 3–16.

[7] M. Studený, P. Boček, CI-models arising among 4 random variables, in: Proc. 3rd Workshop Uncertainty Processing in Expert Systems, Trest, September 11–15, 1994, pp. 268–282.

[8] I. Vajda, On the f-divergence and singularity of probability measures, Periodica Math Hung 2 (1972) 223–234.

[9] I. Vajda, Theory of Statistical Inference and Information, Kluwer, Dordrecht, 1989.

[10] J. Zvárová, Informačni míry statistické závislosti a výběrové vlastnosti zobecněné entropie řádu α (in Czech), thesis, Institute of Information Theory and Automation and Charles University, Prague, 1973.

[11] J. Zvárová, On measures of statistical dependence, Časopis pro pěstování matematiky 99 (1974) 15–29.

[12] J. Zvárová, A. Perez, J. Nikl, R. Jiroušek, Data reduction in computer-aided medical decision-making, in: J.H. van Bemmel, M. Ball, O. Wigertz (Eds.), Proc. MEDINF083, North-Holland, Amsterdam, 1983, pp. 450–453.

[13] J. Zvárová, B. Srb, Z. Malý, A. Martan, Computer supported information analysis of cardiotocographic parameters, in: J.L. Willems, J.H. van Bemmel, J. Michel (Eds.), Progress in Computer-assisted Function Analysis, North-Holland, Amsterdam, 1988, pp. 339–344.