

Symptom Proximity in Diagnostic Problem

Otakar Kříž

Prague, Czech Republic

Copyright ©2017 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract An algorithm SP (= Symptom Proximity) is suggested for solving discrete **diagnostic problem**. It is based on probabilistic approach to decision-making under uncertainty, however, it does not use knowledge integration from marginal distributions.

Keywords Probabilistic Decision Making, Diagnostic Problem

1 Informal analysis

The ability of decision-making belongs to the most important mechanism build in complex organisms. Optimal strategies for decision-making determine the success or even survival of different animal species (Darwin theory), trading companies or individual patients.

As mentioned e.g. in [5], decision-making can take place on a purely intuitive basis (i.e. an individual decision is provided by a genetically inborn mechanism or it can be given by experience acquired during the life span of a decision-maker and stored in neural structures of his brain). Or, decision-making can take place on the basis of a more considered strategy and in the framework of a formalized theory.

When collecting a large number of facts relating to a situation, there appears a point, in certain phase, where facts are "abstracted" to a piece of knowledge. (The process is referred to as hegelian dialectic concept of "quantitative change leads to qualitative change".) Knowledge is usually expressed in form of sets of implications. Then, in the context of decision-making, observed facts (called *evidences*) on an individual object are used as antecedents and the required decision is, hopefully, provided via laws of logics from the succedents (of the implications). However, even this approach involves implicitly certain degree of uncertainty (e.g. when to switch from facts to a law/knowledge).

The tolerable precision is domain dependent. In physics, a theory must explain data with precision to six decimal positions. In soft sciences, models are less demanding. Therefore, one can generalize that decision-making cannot be separated from uncertainty. Partially, it is due to the men-

tioned "ever present lack of data" and partially, there exists an "uncertainty" which of formal theories of uncertainty to use as a model. Even in probability, taken usually as a normative theory (for uncertainty), there exist alternatives (see e.g.[4]).

One of the most fruitful lines in probabilistic decision-making (see subsection 2.1) can be divided into four phases:

First, select less-dimensional distributions, considered as *marginals* of a *theoretical joint distribution*, as input knowledge.

Second, one can construct (integrate) an explicit formula for the *joint distribution*. It can be done via strong assumptions like conditional independence between marginals expressed as *graph models* .

Third, given an *evidence* about an object (i.e. measured on the object), the formula is reduced (marginalized) to a simpler form where the *evidence* can be directly applied.

Fourth, the "best" decision is the one that yields largest *conditional probability on diagnosis variable* (i.e. containing decisions) given the observed *evidence*.

The topic of this paper is a method/algorithm SP that does not use the *marginals* (and assumptions about their conditional independence) and finds an *approximation* of the largest conditional probability directly from the data (called *statistical file* in the sequel and defined in subsection 2.3).

Let us suppose that both models (i.e. "marginal" one and SP) are finite and discrete.

(Explicitly, we do not consider e.g. *family of log-linear distributions* and no *estimation* of their *parameters* from the *statistical file* takes place as is usual in *standard statistics*).

There is a discrete number of *symptom variables*, discrete number of *symptoms* in range of each *symptom variable*, discrete number of diagnoses/decisions and finite number of objects in the *statistical file* (representing knowledge).

It should be stressed that if one uses the term "measured" it should be interpreted as "measured and discretized". E.g. biochemical values, though continuous, are reduced to dichotomies (greater or less than a threshold) or trichotomies ("standard values", "too low", "too high").

This finiteness and discretization need not be a limiting factor, as in many decision-making situations, there is small number of alternatives, as well. In most frequent situations, the decision-making takes place on the basis of not too many

symptom variables .

In "marginal approach", the *joint probability* formula is integrated once forever from *marginals* (elicited, in their turn, from a "learning" *statistical file*).

Then, this **unique** formula is used for each *evidence* available for the object/patient/case.

Whereas, "marginal models" construct an **unique** *approximative distribution* of an **unique** *theoretical joint distribution*, the SP-model uses for decision-making an **ad hoc** *approximation of conditional probability* (derived from the **unique** *theoretical joint distribution*). (This link is established by the supposition that the *statistical file* was generated (by nature?) according to a *theoretical joint distribution*.) This **ad hoc** *approximation of conditional probability* is generated freshly from scratch for each set of *symptom variables* (carrying *evidence* measured on the object/patient). (In other words, the term **ad hoc** relates to a disclosed set of *symptom variables*.)

The aim of SP is not to "predict" the shape of *a posteriori probabilities* of all considered diagnoses, but only to find the most probable diagnosis/decision and even the numerical value of its *a posteriori probability* is not important. In other words, what matters is just the first place in this ranking.

What is new ?

In eighties, it was believed that whereas the data (i.e. *statistical file*) is not sufficient for estimating the *joint distribution*, it should do for less-dimensional tables (i.e. supposed *marginals*). Moreover, these *marginals* were considered as given externally and reliable as "incorporated truth". Beside being given from outside, the *marginals* were just few. One might use a *marginal* for integration or leave it out. Not ask for new ones! The argument that *marginals* cannot be obtained otherwise than from the data was not paid sufficient attention to. If it would have been taken into consideration, one could ask for more *marginals*, but, at the same time, it would raise questions like "How many *marginals* should we ask for?" and "What would be the best composition of those *marginals*?".

The central notion was the *approximation* of the *joint distribution* whereas it is the *approximation of conditional probability* that SP takes as its "flag ship". At the same time, the role of disclosed *symptom variables* carrying *evidence* is stressed for SP, as well. Then, the detour "integration to joint distribution" and subsequent marginalization is not needed any more.

A specific methodology based on proximity (of vector of *evidences* and respective vectors constructed from cases in the *statistical file*) in space of *symptom variables* is used in SP algorithm.

To explain the theory of the SP algorithm, simple *metrics* is introduced in section 3.

As details matter and redundancy is preferable to ambiguity, the essence of the algorithm SP is described twice in section 4.2. Namely, via a flowchart (see Figure 1.) and via a code written in a symbolic language. The latter description makes possible to derive computational complexity for SP in section 5.

However, the real strength of SP lies in the fact that this

approach is fast enough to operate on large *statistical files* with large number of *symptom variables*.

As far as decision quality is concerned, the situation is not so transparent. On one hand side, it is possible to construct examples where SP dominates the "marginal approach". It may happen when input *marginals* are "at wrong position" to the disclosed *symptom variables* carrying *evidences*. However, in general, the dominance of SP cannot be directly proved and one can only compare SP and one of many possible "marginal" algorithms, with different sets of *marginals* and for different *evidence*-carrying *symptom variables*. And it has to be done in many combinations. This "branching expansion" is multiplied by two when we consider two testing techniques. First, "testing" and "learning" *statistical files* are identical. Second, the "leave one out" technique, described in section 6, is used.

The final impression (from this multicriterial decision-making) is that SP does very well. It is based on experience from many comparative runs. If it is not dominant, SP belongs to Pareto optimum, at least.

It should be noted that SP does not use the *empirical distribution* derivable from the *statistical file*. (The "mass" of the *empirical distribution* is concentrated only in the *atoms* of the defining *set algebra* that appears in cases from the *statistical file* and it is zero anywhere else).

One may ask why this line of research (i.e. "without *marginals*") was not followed as a main stream before?

In my opinion, during the last decades, external material conditions (computerization and existence of large databases) changed a lot. In the meantime, the original paradigm was refined and improved by many researchers. The model became so rich, successful and invested in that the inertia prevented an investigation if basic suppositions cannot be weakened.

2 Specific statement of the problem

2.1 Historical background

Firat attempts for machine-assisted decision-making under uncertainty are marked by rule-based expert systems Mycin and Prospector in early eighties. Weights in rules were interpreted as conditional probabilities. But the way the rules were combined was not a probabilistic one. The same held for systems with fuzzy number approach. At that time, Albert Perez (in [11]) raised the requirement that partial knowledge should be "integrated" intensionally i.e. using the concept of theoretical joint distribution P . Knowledge was understood as probability or conditional probability elicited either from experts or observed from experiments. The best way to keep it, at least partially, complete and homogenous was to assume that it comes in form of less-dimensional distributions that were supposed to be *marginals* of the theoretical joint P . Thanks to smaller sizes, *marginals* could be estimated from available data. The main effort in the subsequent research was concentrated on the way how to assemble effective approximations of the joint P . The formulation of the task was known as *marginal problem* already in [6] and its

specific solution was suggested even before in [2]. Different models, connected with names like Lauritzen, Spiegelhalter, Dempster, Shafer, Pearl, Dawid, were studied with assumptions about conditional independence of variables appearing in P that helped to integrate the marginals. At present, there exist professional software packages (e.g. Hugin) supporting the decision-making on commercial basis. As, beside different algorithms, even the selection of proper marginals may be a problem of its own (see e.g. [9] and [10]), this paper tries to study an alternative to the marginal approach.

2.2 The layout of the paper

1. Informal (i.e. without notation) analysis of the diagnostic problem in probabilistic setting is given in section 1.
2. There are historical reminiscences explaining the position of the suggested method in a broader context in subsection 2.1.
3. Basic notions are defined including the formulation of the *diagnostic problem* and describing the role of the *statistical file* \mathcal{F} in subsection 2.3.
4. The essential features of the algorithm SP are laid down in section 3.
5. Realisation of SP is described in section 4 via a flowchart (in subsection 4.1) and via symbolic programming language (in subsection 4.2).
6. On the basis of the latter description, computational complexity C_{SP} of SP in terms of "length" $l = |\mathcal{F}|$ of the file \mathcal{F} and of its "width" $n = |\{\xi_1, \xi_2, \dots, \xi_n\}|$ is estimated and verified experimentally for different values of l and w in section 5.
7. "Discernment power" of SP (i.e. absolute values or percentage of wrong classifications) is tested for different "apertures" (sets of symptom variables whose values are disclosed to SP as evidence). Testing is performed both via method "leave one out" as well as on all data in simulation runs. The results are compared with a simple marginal-based algorithm under the same testing conditions in section 6.
8. In section 7, there are features of SP sorted as "pros" and "cons". Then, there are two open questions suggested for further investigation and finally, there is a summarizing conclusion.

2.3 Basic notions and notations

Let (Ω, \mathcal{X}, P) be a probabilistic space, $\eta = \xi_0, \xi_1, \xi_2, \dots, \xi_n$ be finite sets and $\xi_r : (\Omega, \mathcal{X}, P) \rightarrow (\xi_r, 2^{\xi_r})$ for $r = 0, 1, 2, \dots, n$ be measurable functions.

Though the topic is defined in a formal way, the names of objects in the universe of discussion (e.g. diagnosis, symptoms etc.) are taken from the field of medicine to give

them a semantical interpretation and ease up understanding of basic notions and character of their interaction.

The mutual behaviour of all random variables $\eta, \xi_1, \xi_2, \dots, \xi_n$ is described by a theoretical joint probability distribution $P_{\eta \xi_1 \xi_2 \dots \xi_n}$. Decision making under uncertainty with probabilistic background can be interpreted as the diagnostic problem with the following formulation:

Diagnostic problem: Find the diagnosis $d(s_1, s_2 \dots s_n) \in \eta$ that is the most probable (according to the $P_{\eta \xi_1, \xi_2 \dots \xi_n}$) on the set $\{\omega \in \Omega \mid \xi_1(\omega) = s_1 \ \& \ \xi_2(\omega) = s_2 \ \& \ \dots \ \xi_n(\omega) = s_n\}$ for a given (i.e. observed) arbitrary combination $(s_1, s_2 \dots s_n)$ of values of *symptom variables* from the cartesian product $\Xi = \xi_1 \times \xi_2 \times \dots \times \xi_n$. If we wish to predict the values of diagnostic variable η , the conditional probability $P_{\eta \mid \xi_1 \xi_2 \dots \xi_n}$ (derivable from $P_{\eta \xi_1 \xi_2 \dots \xi_n}$) should be used instead of $P_{\eta \xi_1 \xi_2 \dots \xi_n}$.

Optimal decision: The value of diagnosis d from η that should be selected if the values of symptom variables are $(s_1, s_2 \dots s_n)$ to keep the wrong classifications of d as low as possible), called Bayes solution, is for each $(s_1, s_2 \dots s_n) \in \xi_1 \times \xi_2 \times \dots \times \xi_n$

given by the formula

$$d_{opt}(s_1, s_2 \dots s_n) = \underset{d \in \eta}{\operatorname{argmax}} P_{\eta \mid \xi_1 \xi_2 \dots \xi_n}(d \mid s_1, s_2 \dots s_n) \quad (1)$$

So far the theory. Unfortunately, in the "real world", we are never given the theoretical distribution $P_{\eta \xi_1 \xi_2 \dots \xi_n}$ in full and directly. To compensate for this, we expect to have some indirect information about $P_{\eta \xi_1 \xi_2 \dots \xi_n}$ that will be called *knowledge base* and denoted by \mathcal{K} . It is done by postulating a set of conditions that we believe the theoretical $P_{\eta \xi_1 \xi_2 \dots \xi_n}$ fulfills.

Marginal problem: Using the concept of *marginal problem*, see [6], *knowledge base* \mathcal{K} is given as a set of "low-dimensional" distributions (i.e. number of variables in the distribution does not exceed e.g. 10.), postulated as theoretical *marginal distributions* of the $P_{\eta \xi_1, \xi_2 \dots \xi_n}$. Beside the marginals, there are usually made assumptions about conditional independence holding between groups of random variables. It is interesting that the topic was so attractive that it was addressed in several waves, usually after 20 years. Original and interesting ideas were not just the product of the last two decades but go back much deeper. See e.g. [2], [6],[3], [1]. Instead of the unknown $P_{\eta \xi_1 \xi_2 \dots \xi_n}$, we try to construct (from the marginals) its suitable approximation $\hat{P}_{\eta \xi_1 \xi_2 \dots \xi_n}$ that could play its role in the formula (1).

If existence of marginals is postulated, it is natural to ask where do they come from. Therefore, another notion should be specified.

Statistical file F: Let $(\omega_1, \omega_2, \dots, \omega_s)$ be a sequence, where individual $\omega_i \in \Omega$ denote realizations of a random selection

from Ω ,

then the sequence $(\eta(\omega_l), \xi_1(\omega_l), \xi_2(\omega_l) \cdots \xi_n(\omega_l))_{l=1}^s$ of points in cartesian product $\eta \times \xi_1 \times \xi_2 \times \dots \times \xi_n$ is a statistical file F of size s (i.e. $s = |\mathcal{F}|$) and $(\mathcal{F})_r$ is the r -th member of the sequence \mathcal{F} .

There exists a taciturn assumption that decision making about a concrete case (patient) should be very fast (about 1 sec/pers.). On the other hand, longer time (e.g. hours of CPU time) devoted to selecting and populating the marginals (in the learning phase) is tolerable. This may be one of the reasons why the "marginal approach" is the standard way.

However, using marginals for "integrating" $\hat{P}_{\eta\xi_1\xi_2 \dots \xi_n}$ and its subsequent conditioning need not be mandatory for solving the diagnostic problem.

3 Basic idea of SP algorithm

An algorithm, called SP (= Symptom Proximity), tries to construct necessary conditional probabilities directly from available statistical data file \mathcal{F} . Basic idea of SP can be explained by the assumption "Patients with similar symptoms should have a similar diagnosis". Hence, the name of the algorithms SP interpretes the similarity as a proximity in the sense of a very natural metrics.

Proximity metrics ρ :

$$\rho : \Xi \times \Xi \longrightarrow \mathbf{R} \quad (\mathbf{u}, \mathbf{v}) \longmapsto n - \sum_{i=1}^n \delta((\mathbf{u})_i, (\mathbf{v})_i)$$

where $\delta(\cdot, \cdot)$ is the Kronecker function and $(\mathbf{u})_i$ is the i -th component of the sequence \mathbf{u} . The mapping ρ is a metrics (i.e. reflexivity, symmetry, triangular inequality) on Ξ that can be used for defining equivalence classes on Ξ . For each fixed $\mathbf{v} \in \Xi$, there exist $n + 1$ sets $C_0(\mathbf{v}), C_1(\mathbf{v}), \dots, C_n(\mathbf{v})$, where $C_k(\mathbf{v}) = \{\mathbf{u} \in \Xi \mid \rho(\mathbf{u}, \mathbf{v}) = k\}$.

The next step is to estimate $P(C_k(\mathbf{v}))$. It can be done, in a natural way, using available data (i.e. the statistical file \mathcal{F}).

$$P(C_k(\mathbf{v})) = \sum_{j=1}^{|\mathcal{F}|} \delta(\rho(((\mathcal{F})_j)_{\Xi}, \mathbf{v}), k) / |\mathcal{F}|$$

where $(\mathcal{F})_j$ is the j -th vector from file \mathcal{F} i.e. $(\mathcal{F})_j \in \eta \times \Xi$. Similarly, $((\mathcal{F})_j)_{\Xi}$ is that part of the j -th vector $(\mathcal{F})_j$ that corresponds to symptom variables i.e. $((\mathcal{F})_j)_{\Xi} \in \Xi$. We are interested in the set $C_k(\mathbf{v})$ with smallest k but at the same time such that $P(C_k(\mathbf{v})) > 0$. Let us denote this optimal k as k_0

Finally, the conditional probability $P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v})$ of η on $C_{k_0}(\mathbf{v})$ can be defined.

To shorten the shape of the final expression for the $P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v})$, an auxillary variable $D(\mathcal{F}, \mathbf{v}, k_0, d)$ will

be introduced. It denotes the number of cases (patients) from the statistical file \mathcal{F} that "belong" to the equivalence class $C_{k_0}(\mathbf{v})$ and at the same time they have the diagnosis d

$$D(\mathcal{F}, \mathbf{v}, k_0, d) = \sum_{j=1}^{|\mathcal{F}|} \delta(\rho(((\mathcal{F})_j)_{\Xi}, \mathbf{v}), k_0) \delta(((\mathcal{F})_j)_{\eta}, d)$$

(Similarly according to the above used notation, $((\mathcal{F})_j)_{\eta}$ stands for the value of the diagnosis that is in the j -th vector of the statistical file \mathcal{F} i.e. $((\mathcal{F})_j)_{\eta} \in \eta$.)

Then, $P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v})$ is defined by

$$P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v}) = D(\mathcal{F}, \mathbf{v}, k_0, d) / (|\mathcal{F}| \cdot P(C_{k_0}(\mathbf{v})))$$

If $\mathbf{v} = (s_1, s_2, \dots, s_n) \in \Xi$, we may approximate the conditional probability $P_{\eta|\xi_1\xi_2 \dots \xi_n}(d|s_1, s_2, \dots, s_n)$ appearing in formula (1) by the $P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v})$ so that $P_{\eta|\xi_1\xi_2 \dots \xi_n}(d|s_1, s_2, \dots, s_n) = P_{\eta|C_{k_0}(\mathbf{v})}(d|\mathbf{v})$ and formula (1) can be applied as the decision rule in the SP algorithm.

The algorithm SP is presented in a symbolic programming language in section 4.2. The complexity of the algorithm SP will be defined, in section 5, as a function of size $|\mathcal{F}|$ of the data file \mathcal{F} and as a function of number n of symptom variables. The complexity is verified on real data by measuring time required for making decision for one person.

The decision quality (or discernment power) is dealt with in section 6. In principle, it is the number of wrong classifications what is measured. However, it may be defined more formally:

Let $\mathcal{L} \subset \mathcal{F}, \mathbf{v} \in \Xi$. Further, let $SP(\mathcal{L}, \mathbf{v}) \in \eta$ denote decision of SP when evidence (about a patient) is \mathbf{v} and algorithm SP has the "learning" file \mathcal{L} at his disposal. Then, "discernment power" of SP can be measured by percentage of wrong classifications either as

$$100 \left[1 - 1/|\mathcal{F}| \sum_{j=1}^{|\mathcal{F}|} \delta(SP(\mathcal{F}, ((\mathcal{F})_j)_{\Xi}), ((\mathcal{F})_j)_{\eta}) \right]$$

or with the formula

$$100 \left[1 - 1/|\mathcal{F}| \sum_{j=1}^{|\mathcal{F}|} \delta(SP(\mathcal{F} \setminus (\mathcal{F})_j, ((\mathcal{F})_j)_{\Xi}), ((\mathcal{F})_j)_{\eta}) \right]$$

This second approach is referred to as "Leave one out" technique. (There are more details on the technique in section 6.)

The results will be compared with one simple algorithm using the "marginal approach" in section 6.

4 Detailed description of SP

There is a theory of SP in section 3 and there is a realization of SP in this section 4.

As mentioned above in section 2, SP algorithm is described twice. Namely, via a flowchart (see Figure 1.) and via a code written in a symbolic language.

Strictly speaking, realization of an algorithm is an executable file (i.e with extension **.exe**) that is "understood" e.g. by any PC.

Less strictly, realization of an algorithm is syntactically and semantically correct text file "written" by a man-programmer in a general purpose programming language (like Fortran, C, C++, Visual Basic) that is "understood" by a compiler of the respective language. (This text file has extensions like **.for**, **.c**, **.vbs**.)

Next level is symbolic language. It is an abstraction of general purpose languages. It is "written" by man and it can be "understood" by another man to the extent that he/she can rewrite it to correct text in general purpose language.

The last common representation of an algorithm is its flowchart. It is a "text" written by man-programmer in graphical language form. It has to be "understood" by another man. It has conventional structures like action blocks (rectangles), decision blocks (rhombus or trapezoid) and lines where arrows determine the "flow" of the program. Some symbols are overloaded. E.g. symbol "=" is interpreted as an assignment in action blocks whereas it is a relational operator in decision blocks.

To summarize, there is always potential place left for ambiguity and misunderstanding. Therefore, both descriptions of the realization are used in parallel and they are supplemented by informal comments that explain semantics of used structures (in context of theoretical foundations of the algorithm).

There is a slight functional difference between the flowchart and the symbolic language descriptions of the SP algorithm.

The flowchart describes entering **one** statistical file \mathcal{L} (from the problem area) and entering a **sequence** of several new *evidences* \mathbf{v} for which their optimal diagnoses are calculated.

In the symbolic language, there is a body of a *function* **SP** where only **one** given statistical file \mathcal{L} and only **one** evidence \mathbf{v} are entered as *parameters*.

Then, the corresponding diagnosis $d_{opt}(\mathbf{v})$ is calculated and returned when the *function* **SP** ends. The difference is only a formal one and the essence of SP algorithm is the same.

4.1 Flowchart of SP

It represents the activity of the algorithm in a simplified form. It is a link between the theory and the realization. Therefore, the symbols from both "worlds" are used. E.g. the first action block describes the filling of the array L with values from the statistical file denoted as \mathcal{L} to stress it is for "learning". Names of the symptoms, usually in form of text strings, are densely coded into integers. Denotation $L[0-n, 1-|\mathcal{L}|]$ says that two-dimensional array L has $n+1$ positions in the first dimension. (Position "0" is used for storing diagnoses, positions "1" to "n" are for storing values of *symptom variables*). The second dimension is for indexing

individual records from \mathcal{L} . Variable *maxcount* corresponds to k_0 via definition $k_0 = n - \text{maxcount}$. There are three loops in flowchart which are realized as **for**-cycles in the symbolic language. Filling the two-dimensional array LD with zeroes takes place anytime new *evidence* \mathbf{v} is entered. In the end, after running through the whole L matrix, the array element $LD(i,j)$ contains the number $D(\mathcal{L}, \mathbf{v}, n - i, d_j)$. It is easy to see that elements of the matrix LD are integers and their sum is $|\mathcal{L}|$.

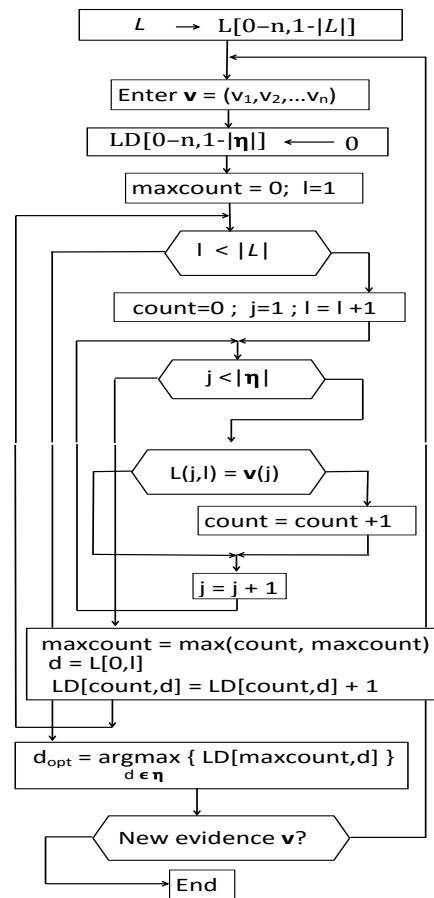


Figure 1. Flowchart of SP.

It would be possible to calculate the approximation of conditional probability $P_{\eta|C_{k_0}}(\mathbf{v})(d|\mathbf{v})$ of η on $C_{k_0}(\mathbf{v})$. (It can be done by normalization of the row $LD(\text{maxcount}, *)$ of the array LD . i.e. by dividing elements in the row $LD(\text{maxcount}, *)$ by their sum.) However, as we are interested only in the ranking, this normalization is not necessary and we look for d_{opt} by looking for maximal value in the row $LD(\text{maxcount}, *)$ of matrix LD , only. That way, we save computational time by skipping unnecessary actions.

Semantics of LD (level distance) can be seen from the fact that the row $LD(i, *)$ contains number of all records/vectors from \mathcal{L} that have just i mutual *symptom variables* with the *evidence* \mathbf{v} . In other words, they belong to the equivalence class $C_{n-i}(\mathbf{v})$.

4.2 Description of SP in a symbolic language

SP algorithm can be used in different roles. It may be a simple "one-shot" decision-making, repeated decision-making for different apertures (sets of symptom variables whose values are disclosed to *SP* as evidence), using *SP* in a general testing scheme or it may be adapted to specific testing via "Leave one out" technique.

Instead of using one highly parametrized form of *SP* algorithm, it seems better, for didactical reasons, to use several stand-alone modifications. However, only the most simple version, under the name *function SP*, will be presented in this paper. Specific modifications built on its basis (and entitled SPL and SPA) will be mentioned in other sections. The following symbolic description is kept as simple as possible.

First, though the variables have their specific denotation reflecting their semantics, they are coded as integers or arrays of integers to make *SP* faster.

Second, tests and resulting exceptions in inconsistent situations such as $|\mathcal{L}| = 0$ or $|\eta| = 0$ are omitted !

Third, this version of *SP* algorithm is defined for all n available symptom variables. However, it can be easily modified if not values of all symptom variables are known, but if only some of them are provided as evidences (i.e. in case of smaller aperture).

Function **SP** returns the value $d_{opt}(\mathbf{v})$ for each given $\mathbf{v} = (v_1, v_2 \dots v_n) \in \xi_1 \times \xi_2 \times \dots \times \xi_n$

```

1  function SP (v)
2  read  $\mathcal{L} \rightarrow L(0 - n, 1 - |\mathcal{L}|)$ 
3  for  $j = 1, |\eta|$ 
4  for  $i = 0, n$ 
5   $LD(i, j) = 0$ 
6  next  $i$ 
7  next  $j$ 
8   $maxcount = 0$ 
9  for  $l = 1, |\mathcal{L}|$ 
10  $count = 0$ 
11 for  $j = 1, n$ 
12 if  $L(j, l) = \mathbf{v}(j)$  then
13  $count = count + 1$ 
14 endif
15 next  $j$ 
16 if  $maxcount < count$  then
17  $maxcount = count$ 
18 endif
19  $d = L(0, l)$ 
20  $LD(count, d) = LD(count, d) + 1$ 
21 next  $l$ 
22  $max = 0; d_{opt} = 0$ 
23 for  $j = 1, |\eta|$ 
24  $val = LD(maxcount, j)$ 
25 if  $max < val$  then
26  $max = val$ 
27  $d_{opt} = j$ 
28 endif
29 next  $j$ 
30  $SP = d_{opt}$ 
31 end
```

Comments to the code of *SP*:

1.1 expresses that *SP* is a function $SP : \Xi \rightarrow \eta$ i.e. accepts as argument the vector \mathbf{v} and returns the optimal diagnosis $d_{opt}(\mathbf{v})$.

1.2 learning file \mathcal{L} is stored in an array L . The value "0" in first dimension is for values of η .

1.3 - 1.7 sets zero values to the array LD (level distance) where metrics will be stored in the sequel.

1.9 - 1.21 For each $l \in \mathcal{L}$, number of symptom variables with coinciding values (symptoms) is calculated in the variable *count*. Increasing $LD(count, \eta(l))$ by one increases chances of diagnosis $\eta(l)$ to become optimal d_{opt} if the decision should take place at the level *count*.

1.16 - 1.18 stores in *maxcount* the up-to-now achieved maximal number of coincidences.

1.23 - 1.30 finds in $LD(maxcount, j)$ such diagnosis d_j that would, on the level *maxcount*, define the winning diagnosis d_{opt} . Naturally, if the number of cases from \mathcal{L} is small (and that would result in objections from statistical point of view), it is possible to perform search for optimal diagnosis on a level *count* smaller than *maxcount* that would have more objects than level *maxcount*. Or even, it is possible to sum $LD(ct, j)$ for $ct = count$ to *maxcount* in an array $D(1-|\eta|)$ and search for d_{opt} in this array. (However, this modification of *SP* is not available in the presented version.)

The link of the code with previous formal description may be made more clear if we realize that the variable *maxcount* is related to k_0 . However, their roles are reversed in the sense that *maxcount* is the number of coincidences (of symptoms from \mathbf{u} and \mathbf{v}) and should be as close to n as possible, whereas k_0 is proximity (in the sense of the metrics ρ) and should tend to zero. Further, $\mathbf{v}(j)$ is $(\mathbf{v})_j$ and the value in the array $LD(maxcount, j)$ stores the $D(\mathcal{L}, \mathbf{v}, k_0, d_j)$.

5 Computational complexity

It should be mentioned that experimenting with algorithms was performed on a statistical file \mathcal{F} , from the field of rheumatology with 1089 patients. Diagnosis variable η contained 4 diagnosis and there were 34 symptom variables whose ranges had cardinalities from 2 to 9. That way, no generation of artificial examples was necessary. Nevertheless, this choice has no influence on the substance of *SP*.

Complexity C_{SP} of *SP* algorithm can be measured with respect to the number of symptom variables n , number $|\mathcal{L}|$ of objects in the learning file \mathcal{L} and with respect to the number $|\eta|$ of diagnoses i.e. $C_{SP} = C(n, |\mathcal{L}|, |\eta|)$. Due to the simple structure of *SP*, C_{SP} can be estimated directly:

1.2	$c_1 * n * \mathcal{L} $
1.3 - 1.7	$c_2 * n * \eta $
1.9 - 1.15	$c_3 * \mathcal{L} * (n + c_4)$
1.23 - 1.29	$c_5 * \eta $

$$C_{SP} = n [(c_1 + c_3)|\mathcal{L}| + c_2|\eta|] + c_3c_4|\mathcal{L}| + c_5|\eta|$$

Table 1. Computational time dependence on length $m = |\mathcal{F}|$ of file \mathcal{F}

m	T_{read}	T_{total}	$T_{decision}$
1000	31 msec	1 sec	10 msec
2000	60 msec	2 sec	20 msec
5000	156 msec	25 sec	50 msec
10000	343 msec	100 sec	100 msec
20000	687 msec	400 sec	200 msec

Table 2. Computational time dependence on width n of file \mathcal{F}

n	T_{read}	T_{total}	$T_{decision}$
35	31 msec	0.546 sec	0.5 msec
70	31 msec	1.046 sec	0.9 msec
105	47 msec	1.516 sec	1.3 msec
140	45 msec	1.968 sec	1.8 msec
200	78 msec	2.78 sec	2.59 msec
300	109 msec	4.07 sec	3.78 msec

The assumption of linearity ($c_1, c_2 \dots$) is a bit simplification and valid only for small ranges of $n, |\mathcal{L}|, |\eta|$. If the ranges are greater, then effects like "paging" of memory, the way files are stored in a concrete file system (e.g. **FAT 32** or **NTFS**) and variables used for storing the "coding" numbers may come in play. E.g. values of η are stored in variables of type **integer*1** and therefore should not exceed 256.

Therefore, instead of looking for explicit values for c_1, \dots, c_5 , direct measurements are documented in Table 1 where length $|\mathcal{L}|$ of \mathcal{L} varies from 1000 to 20000 and in Table 2 where width n of \mathcal{L} varies from 35 to 300. Corresponding files \mathcal{L} (e.g. $\mathcal{L}(70, 1089)$ or $\mathcal{L}(35, 20000)$) were generated from original \mathcal{L} (i.e. $\mathcal{L}(35, 1089)$) by repeating respective rows and columns. In the Tables 1 and 2, column T_{read} contains time necessary to read \mathcal{L} . Column T_{total} increases with square power of $|\mathcal{L}|$ as it is the time necessary for $|\mathcal{L}|$ decisions. When T_{total} is divided by $|\mathcal{L}|$, then the times in column $T_{decision}$ are always below 1 sec and therefore completely acceptable. Based both on analysis and direct measurements, complexity of SP is not a problem. Therefore, the limiting factor for better discernment power of SP is an externality i.e. experts should provide bigger data in form of a file \mathcal{F} (or \mathcal{L}).

6 Simulation results

This section should reflect the decision-making quality of SP via simulation results. Simulation is performed on the data mentioned in the beginning of section 5. Though the following example is very simple, it may reveal interesting facts when comparing SP and the decision-making algorithm $A4$ (from [7] and mentioned in [8]) that can serve as a simple representative of marginal-based algorithms. Let knowledge base \mathcal{KB} consist of 3 marginals i.e. $\mathcal{KB} = \{m_1, m_2, m_3\}$.

The marginals can be defined by their "generating" symptom variables (besides implicitly supposed diagnosis variable η). Underscoring of symbol for a marginal denotes the set of "generating" symptom variables. E.g., if $\underline{m}_1 = \{\xi_{25}, \xi_{33}\}$, then $m_1 = P_{\eta\xi_{25}\xi_{33}}$. Then, \mathcal{KB} used for testing was defined via

$$\underline{m}_1 = \{\xi_{25}, \xi_{33}\}, \underline{m}_2 = \{\xi_{26}, \xi_{32}\}, \underline{m}_3 = \{\xi_{27}, \xi_{33}\}$$

The "testing environment" provides an easy way to manipulate with "inputs" to the decision-making algorithm.

First, it makes possible to remove marginals from \mathcal{KB} and second, not all symptom variables, from n possible ones, need to be revealed as "evidences" to decision making-algorithm SP or $A4$. We tested SP or $A4$ alongside for 8 different situations s_1, s_2, \dots, s_8 described in Table 3. E.g., the expression $\{m_1, m_3\} \cap \{\xi_1 - \xi_{33}\}$ (in column "marginals \cap variables" of Table 3) stands for situation s_2 where \mathcal{KB} consists only of marginals $\{m_1, m_3\}$ and values of all 33 symptom variables $\{\xi_1 - \xi_{33}\}$ are submitted as "evidences" to the SP and $A4$. (Naturally, $\{m_1, m_3\}$ has impact only on $A4$, whereas $\{\xi_1 - \xi_{33}\}$ influences both SP and $A4$). Column "active variables" in Table 3 contains symptom variables whose values have influence on $A4$ as result of both conditions. Column "active space" is product of their ranges. As all symptom variables here are dichotomical ones, the values are like 4, 16, 32.

Even the above mentioned denotation for individual marginals is a little simplified. E.g. $m_1 = P_{\eta\xi_{25}\xi_{32}}$ is not enough as it should be also mentioned what data was used for populating the marginal m_1 . This can be expressed by adding the source. E.g. $m_1 = P_{\eta\xi_{25}\xi_{32}}(\mathcal{L})$ stands for the marginal filled from the data set \mathcal{L} . This denotation would do for the column $A4A$, but not for calculating the values for column $A4L$. Then, in fact, there are 1089 different marginals $m_1(\mathcal{L} \setminus t) = P_{\eta\xi_{25}\xi_{32}}(\mathcal{L} \setminus t)$. Those marginals will be populated from 1089 different data files $\mathcal{L} \setminus t$ that have to be created just for the purpose.

The "A" (in column $A4A$ containing number of wrong classifications) stresses that *all* data was used both for learning and testing i.e. $\mathcal{L} = \mathcal{T} = \mathcal{F}$. The "L" (in column $A4L$) has the meaning that the method "Leave one out" was used for the calculation of "discernment power" of algorithm $A4$.

The essence of "Leave one out" technique consists in following steps:

1. The j -th vector from file \mathcal{F} (denoted as $(\mathcal{F})_j$) is removed from \mathcal{F} .
2. The resulting learning file $\mathcal{L}_j = \mathcal{F} \setminus \{(\mathcal{F})_j\}$ is used for populating all necessary marginals in \mathcal{KB} .
3. The "symptom part" $((\mathcal{F})_j)_{\Xi}$ of vector $(\mathcal{F})_j$ is submitted as input to the decision-making algorithm (e.g. SP or $A4$).
4. The resulting decision (e.g. $A4((\mathcal{F})_j)_{\Xi}$) is compared with the "diagnosis part" $((\mathcal{F})_j)_{\eta}$ of vector $(\mathcal{F})_j$ and coincidence of those two diagnoses is "marked" as 1.
5. The same procedure is repeated $|\mathcal{F}|$ times (for $j = 1, 2, \dots, |\mathcal{F}|$) and the normalized sum of coincidences is an indicator of decision quality of the algorithm. Instead of coincidences, we prefer to use percentage of misclassifications as can be seen in second formula in Section 2

It can be observed in Table 4 that L-values are higher than corresponding A-values. In general, SP is slightly better than $A4$, but not always

e.g. $A4L(s8) = 423 < SPL(s8) = 431$. On the basis of other similar experiments, it looks like that advantages of SP may be more prominent but only for A-testing. Especially for \mathcal{KB} with more marginals and when values of all symptom variables are known. As far as "Leave one out" method is concerned and **not** with full-sized evidence, no decisive conclusions can be drawn, so far. However, it seems that SP does quite well and could be used along with other recommended methods.

7 Discussion and conclusions

Among positive features of marginal-less SP algorithm, the following ones can be mentioned :

1. The presented algorithm SP is sufficiently fast i.e. decisions are made within seconds.
2. SP has good discernment power when the tested case t was included in the learning file \mathcal{L} and values of all symptom variables (from \mathcal{L}) are given as input evidence.
3. It is easy to add new cases (or remove old ones if considered as obsolete) to the learning file \mathcal{L} . In marginal-based approach, it is necessary to recalculate the marginals.

4. Problems associated with selection of marginals are avoided (by definition !) and only symptom variables are necessary. In general, values of all symptom variables (present in the learning file \mathcal{L}) should be provided as evidences, if available.
5. Testing via "Leave one out" technique is extremely easy with a small modification in the presented code of SP . It takes approximately the same time as testing on the all data (i.e. when $\mathcal{L} = \mathcal{T}$). Marginal-based algorithm require for "Leave one out" a lot of time for splitting the data file ($|\mathcal{F}|$ times !) and filling the marginals for each split.
6. If an evidence $\mathbf{v} \in \Xi$ turns up, as input for SP , that is **not** present in the learning file \mathcal{L} i.e. $\mathbf{v} \notin ((\mathcal{F})_j)_{\Xi}$ for $j = 1, 2, \dots, |\mathcal{F}|$ then SP recommends the diagnosis with the highest apriori probability i.e.

$$SP(\mathbf{v}) = \operatorname{argmax}_{d_j \in \eta} P_{\eta}(d_j)$$

what sounds as one would expect.

7. If there is a valid piece of knowledge of logical nature (e.g. implications), it is possible to force it out just by repeating artificial records representing it in \mathcal{L} several times.

SP has several drawbacks as well:

1. SP can be applied only to nominal variables (i.e. not continuous, not cardinal and even not to ordinal) due to properties of the proximity metrics ρ .
2. As the only testing criterion is number of wrong classifications, SP , in its presented version, is not a proper choice for risk analysis.
3. With decreasing number of symptoms (evidences), discernment power of SP drops as well. (It is similar to marginal-based algorithms, as well.)
4. It is not possible to add additional knowledge about structure of P (e.g. in form of graph models expressing conditional independence among marginals.) All is based on input data represented by the statistical file \mathcal{F} (or \mathcal{L}) only.

There are some open questions to study.

One would expect that decision quality is decreasing with increasing value k of proximity metrics ρ . It can be observed on testing runs. However, one would expect certain invariance or at least monotonicity in recommended d_{opt} . In general, it does not always happen and recommended d_{opt} "oscillates" when moving in array LD to levels with smaller symptom incidence i (or otherwise for ρ higher k value.). Would it be possible to change this behaviour via introducing a sort of "weighting" on *symptom variables*? Up to now,

Table 3. Different testing situations

situations	marginals \cap variables	active variables	active space
s_1	$\{m_1, m_2, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{25}, \xi_{26}, \xi_{27}, \xi_{32}, \xi_{33}$	32
s_2	$\{m_1, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{25}, \xi_{27}, \xi_{33}$	8
s_3	$\{m_2\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{26}, ξ_{33}	4
s_4	$\{m_1\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{25}, ξ_{33}	4
s_5	$\{m_3\} \cap \{\xi_1 - \xi_{33}\}$	ξ_{27}, ξ_{33}	4
s_6	$\{m_2, m_3\} \cap \{\xi_1 - \xi_{33}\}$	$\xi_{26}, \xi_{27}, \xi_{32}, \xi_{33}$	16
s_7	$\{m_1, m_2, m_3\} \cap \{\xi_1 - \xi_{32}\}$	$\xi_{25}, \xi_{26}, \xi_{27}, \xi_{32}$	16
s_8	$\{m_1, m_2, m_3\} \cap \{\xi_{33}\}$	ξ_{33}	2

Table 4. Comparing wrong classifications for SP and A4

situations	SPL	SPA	A4L	A4A
s_1	421	415	427	421
s_2	462	460	463	462
s_3	538	538	538	538
s_4	464	464	464	464
s_5	463	461	463	461
s_6	431	420	423	422
s_7	537	530	539	539
s_8	464	464	464	464

all *symptom variables* were supposed as equally informative. And if it would be effective, how to implement it in the existing realization of the SP? Another topic that would deserve a deeper investigation into is behaviour of SP for small sets of disclosed symptom variables (so called *aperture*) used for defining *evidence*. Though the improvement in decision quality (with respect to alternative approaches) may be only several percentage points, it would be interesting to find if, on average, SP is a reliable tool for decision-making even in this situation.

To summarize, with respect to above mentioned arguments, *SP* can be recommended for decision-making on nominal symptom variables and when a sufficient learning data file is available. It may serve as an alternative to well established marginal-based algorithms for decision-making under uncertainty. It has sound basic philosophy ("Patients with similar symptoms should have a similar diagnosis") and it is theoretically well founded (conditioning on equivalence classes induced by proximity metrics ρ and direct link to the optimal solution represented by formula 1). According to

simulations (described in section 6) on a realistic study case (mentioned in section 5), it does quite well and it does not require sophisticated suppositions (like maximal entropy principle or Markovian blanket) about structure of approximating joint distributions. The problem with marginal selection (i.e. "how many" and "what composition") is circumvented in SP. **Remark:** The original version (entitled "Diagnostic problem without marginals") was published at Wupes15 Workshop at Moninec. This is the extended version of that.

Acknowledgements

I am very grateful to the anonymous referee for valuable comments and suggestions to improve legibility of the paper. My thanks belong also to R.Jiroušek who drew my attention to the fact that suggested metrics can be seen as an instance of Hamming distance known from coding theory.

REFERENCES

- [1] P.Cheeseman: A method of computing generalized Bayesian probability values of expert systems with probabilistic background, in: Proc. 6-th Joint Conf. on AI(IJCAI-83), Karlsruhe
- [2] W.E. Deming, F.F Stephan: On a least square adjustment of sampled frequency table when expected marginal totals are known, Ann.Math.Stat. 11(1940), pp. 427 - 444
- [3] E.T. Jaynes: On the rationale of maximum-entropy methods, Proc. of the IEEE 70 (1980), pp. 939 - 952.
- [4] Terrence L.Fine: Theories of Probability: An examination of foundations, Academic Press,1973)
- [5] David Kahnemann: Thinking: Fast and Slow, (2012)
- [6] H.G. Kellerer: Verteilungsfunktionen mit gegebenen Marginalverteilungen. Zeitschrift für Wahrscheinlichkeitstheorie, 3(1964), pp. 247 -270.

- [7] O. Kříž: A new algorithm for decision making with probabilistic background, in: Transactions of the Eleventh Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, August 27-31, 1990, Vol. B, (Academia, Prague, 1992) pp 135-143
- [8] O. Kříž: Comparing algorithms based on marginal problem. *Kybernetika*, Vol 43, (2007), No.5, 633–647
- [9] O. Kříž: Selecting marginals for decision making based on marginal problem. *WUPES'09*, Vol 43, (2009), No.5, 633–647
- [10] O. Kříž: Mixing marginals for decision making based on marginal problem. In: *WUPES 2012, Proceedings of the 9-th Workshop on Uncertainty Processing*, (T. Kroupa, J. Vejnarová ed.), Mariánské Lázně 2012, pp. 114–125.
- [11] A. Perez: A probabilistic approach to the integration of partial knowledge for medical decision-making (in Czech). In: *Proc. of the 1-st Czechoslovak Congress of Biomedical engineering (BMI83)* (J. Zvárová ed.), Mariánské Lázně 1983, pp. 221–226.