A General Approach to Probabilistic Data Mining[1]

Radim Jiroušek, Václav Kratochvíl

Czech Academy of Sciences
Institute of Information Theory and Automation, Praha, Czech Republic,
radim@utia.cas.cz, velorex@utia.cas.cz

Abstract. The paper describes principles enabling us to express the knowledge hidden in a multidimensional probability distribution – a distribution that is assumed to have generated the input data – into the form legible by humans, into the form expressible in a plain language. The generality of this approach arises from the fact that we do not assume any type of probability distribution. The basic idea is that the analysis of such a multidimensional distribution is, because of its computational complexity, intractable, and therefore we construct its approximation in a form of a decomposable model, which provides an easy interpretation. The process should be controlled by an expert in the field of application, and the presented principles give him instruction, how, using the tools from probability and information theories, to get satisfactory results.

Keywords: approximation, probability models, conditional independence, decomposition, information content, ambiguity

1 Introduction

There is abundant literature on data mining and, quite naturally, a great number of different definitions explaining what the authors understand by this notion. All the authors agree that data mining is a process discovering interesting relationships that are to be found in large databases, a process uncovering useful information that can be expressed in the form of knowledge. And this is the point in which the individual data mining processes differ from each other. Some of the authors consider data mining to be a part of machine learning, as e.g., [5], and therefore they represent the discovered knowledge in a form of specific models. Some others look for relations representable in a form of IF-THEN rules (usually loaded with some uncertainty). As we are now going to support with arguments, in the described approach we look for a knowledge representable with the tools of probability theory.

In the beginning, the artificial intelligence refused probability theory for knowledge representation and inference because of several reasons. Among them,

[1] This survey lecture is patterned on the manuscript of the book [6], and on preceding papers of the authors.

Table 1: An example of a *direct proportion.*

Age	Proportion of patients with disease D
less than 40	2.1 %
40 – 49	7.3 %
50 – 59	14.6 %
60 – 69	31.1 %
70+	44.9 %

the rigidity of statistical methodology and the complexity of the respective computational procedures played important roles. It was in the middle of the eighties of the last century when the probability theory started penetrating into the field of artificial intelligence thanks to the papers like [2], and the tools based on probabilistic graphical models [15, 16]. Naturally, we do not claim that the probability theory is an approach capable to represent all forms of knowledge, but it is general enough that it can serve for the purpose of this paper. It can represent a logical implication (IF-THEN rule) by a two-dimensional distribution (four-fold table) with one zero value. If this rule is loaded with uncertainty, then it contains instead of the zero a small probability. In probability logic, the validity of implication is formalized as a conditional probability. Moreover, probability table (distribution), like in Table 1, can represent a type of dependence we express in words "the older, the greater the risk of suffering from disease D." There are many other types of dependence that can be read from a respective low-dimensional probability table. If the reader cannot find a more general description of dependence, it is always possible to express it as a series of conditional probabilities, i.e., a series of IF-THEN rules, the validity of which corresponds to the value of the corresponding conditional probability. But keep in mind that it is reasonable to explain the type of dependence only when the respective probability table is low-dimensional, as a rule, the dimension should usually be lower than 5.

Among all possible types of dependence, the most important is *independence*, or more generally, *conditional independence*, about which we will speak later in more details. Before proceeding to a more formal exposition let us admit that the probability theory has also its limits: it cannot model *ambiguity*. It has been known since the middle of the last century [4] that humans do not like ambiguity. They prefer situations when they know probabilities of the alternatives influencing their decision, and hate situations when the probabilities are fully unknown. This phenomenon is called an Ellsberg paradox, and it is known that classical probability theory cannot treat such situations easily. The only way how to overcome this problem is to employ some generalization designed for treating uncertain probabilities, as, e.g., belief function theory.

2 Basic Notions and Notation

Let us assume the records from the available data represent observations of random variables, which are, in this paper, denoted by upper-case characters of Latin alphabet (like $X, Y, \ldots$). Finite sets of values of these variables are denoted by $\mathbb{X}_X$, $\mathbb{X}_Y, \ldots$. Thus, for example, if variable Y denote a 'sex' of a respondent than $\mathbb{X}_Y$ contains just two values corresponding to 'female' and 'male'. Most of the time we will deal with sets of variables denoted by bold-face characters $\mathbf{K}$, $\mathbf{L}$, $\mathbf{M}$, $\mathbf{N}$. Thus, $\mathbf{K}$ may be $\{X, Y, W\}$. By a *state* of variables $\mathbf{K}$ we understand any combination of values of the respective variables, i.e., in the considered case $\mathbf{K} = \{X, Y, W\}$, a state is an element of a Cartesian product $\mathbb{X}_X \times \mathbb{X}_Y \times \mathbb{X}_W$. For the sake of simplicity, this Cartesian product is denoted $\mathbb{X}_{\mathbf{K}}$. For a state $y \in \mathbb{X}_{\mathbf{K}}$ and $\mathbf{L} \subset \mathbf{K}$, by $y^{\downarrow \mathbf{L}}$ we denote a *projection* of $y \in \mathbb{X}_{\mathbf{K}}$ into $\mathbb{X}_{\mathbf{L}}$, i.e., $y^{\downarrow \mathbf{L}}$ is the state from $\mathbb{X}_{\mathbf{L}}$ that is got from y by dropping out all the values of variables from $\mathbf{K} \setminus \mathbf{L}$.

Probability tables (distributions) are denoted by characters of Greek alphabet (κ, λ, μ, ν, π). Recall that it means that $\kappa(\mathbf{K}) : \mathbb{X}_{\mathbf{K}} \longrightarrow [0,1]$, for which $\sum_{x \in \mathbb{X}_{\mathbf{K}}} \kappa(x) = 1$.

Having a probability distribution $\kappa(\mathbf{K})$, and a subset of variables $\mathbf{L} \subset \mathbf{K}$, $\kappa^{\downarrow \mathbf{L}}$ denote a *marginal distribution* of κ defined for each $x \in \mathbb{X}_{\mathbf{L}}$ by the formula

$$\kappa^{\downarrow \mathbf{L}}(x) = \sum_{y \in \mathbb{X}_{\mathbf{K}} : y^{\downarrow \mathbf{L}} = x} \kappa(y).$$

For a probability distribution $\kappa(\mathbf{K})$, we introduce a conditional distribution in a standard way. For disjoint $\mathbf{L}, \mathbf{M} \subseteq \mathbf{K}$, by a conditional distribution of variables $\mathbf{L}$ given variables $\mathbf{M}$ we understand any function $\kappa^{\mathbf{L}|\mathbf{M}} : \mathbb{X}_{\mathbf{L} \cup \mathbf{M}} \longrightarrow [0,1]$ meeting the following two conditions:

- $\forall x \in \mathbb{X}_{\mathbf{L} \cup \mathbf{M}} \;\; \kappa^{\downarrow \mathbf{L} \cup \mathbf{M}}(x) = \kappa^{\mathbf{L}|\mathbf{M}}(x) \cdot \kappa^{\downarrow \mathbf{M}}(x^{\downarrow \mathbf{M}})$,
- $\forall$ fixed $x \in \mathbb{X}_{\mathbf{M}}$ function $\kappa^{\mathbf{L}|\mathbf{M}}$ as a function of variables $\mathbf{L}$ is a probability distribution, i.e., $\sum_{y \in \mathbb{X}_{\mathbf{L} \cup \mathbf{M}}; y^{\downarrow \mathbf{M}} = x} \kappa^{\mathbf{L}|\mathbf{M}}(y) = 1$.

Due to the latter condition, the argument y of the function $\kappa^{\mathbf{L}|\mathbf{M}}$ is often split into two complementary pieces $y^{\downarrow \mathbf{L}}$ and $y^{\downarrow \mathbf{M}}$, and its value is depicted as $\kappa^{\mathbf{L}|\mathbf{M}}(y^{\downarrow \mathbf{L}} | y^{\downarrow \mathbf{M}})$.

Consider two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$. We say that κ and λ are *consistent* if $\pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \kappa^{\downarrow \mathbf{K} \cap \mathbf{L}}$. For two probability distributions defined for the same group of variables, say $\pi(\mathbf{K})$, $\kappa(\mathbf{K})$, we say that κ *dominates* π (in symbol $\pi \ll \kappa$) if

$$\forall \; x \in \mathbb{X}_{\mathbf{K}} \quad (\kappa(x) = 0 \implies \pi(x) = 0) \, .$$

Consider a probability distribution $\pi(\mathbf{N})$, and three disjoint subsets of variables $\mathbf{K}, \mathbf{L}, \mathbf{M}$ ($\mathbf{K} \cup \mathbf{L} \cup \mathbf{M} \subseteq \mathbf{N}$). Let $\mathbf{K}$ and $\mathbf{L}$ be nonempty. We say that groups of variables $\mathbf{K}$ and $\mathbf{L}$ are *conditionally independent given* $\mathbf{M}$ *for distribution* π if

$$\pi^{\downarrow \mathbf{K} \cup \mathbf{L} \cup \mathbf{M}} \cdot \pi^{\downarrow \mathbf{M}} = \pi^{\downarrow \mathbf{K} \cup \mathbf{M}} \cdot \pi^{\downarrow \mathbf{L} \cup \mathbf{M}}.$$

In symbol, this property is expressed by $\mathbf{K} \perp\!\!\!\perp \mathbf{L} | \mathbf{M}\ [\pi]$. In case of $\mathbf{M} = \emptyset$ we use only $\mathbf{K} \perp\!\!\!\perp \mathbf{L}\ [\pi]$ and speak about an unconditional independence.

We have already mentioned that the conditional independence is considered in this paper as a property expressing knowledge about the reality described by the considered probability distribution. Perhaps, it is not visible just from the definition, but it is an easy exercise to show that if $\mathbf{K} \perp\!\!\!\perp \mathbf{L} | \mathbf{M}\ [\pi]$, then

$$\pi^{\downarrow \mathbf{K}|\mathbf{L} \cup \mathbf{M}} = \pi^{\downarrow \mathbf{K}|\mathbf{M}}.$$

In words: if we know a state from $\mathbb{X}_{\mathbf{M}}$, then learning values of variables from $\mathbf{L}$ does not bring us any new information about variables from $\mathbf{K}$. If we know a state from $\mathbb{X}_{\mathbf{M}}$, then groups of variables $\mathbf{K}$ and $\mathbf{L}$ become independent. For example, the intensity of training and the placing of a sportsman in a race are dependent. But conditionally, given the time in which he accomplished the race, these two events become independent. Namely, when knowing the time he achieved in the race, the probability that he wins the race does not change when learning how much time he spent in training.

For a probability distribution π, by its *independence structure* we understand the list of all conditional independence relations holding for π. It explains us, which of the dependence relations are direct, and which are mediated through other variables.

3 Decomposition of Probability Distributions

As said in the introduction, humans can read knowledge from low-dimensional probability tables. It means that when considering a multidimensional distribution one has to decompose it into low-dimensional ones, and the decomposition should be done in the way that it gives evidence about the data [8]. Moreover, any decomposition is required to meet the following properties

- the result of the decomposition are two objects of the same type as the decomposed object;
- both these sub-objects are simpler (smaller) than the original object;
- not all objects can be decomposed;
- there exists an inverse operation (we call it a composition) yielding the original object from its decomposed parts.

These properties made us to accept the following definition.

Definition 1. *We say that a probability distribution* $\pi(\mathbf{M})$ *is* decomposed *into its marginals* $\pi^{\downarrow \mathbf{K}}$ *and* $\pi^{\downarrow \mathbf{L}}$ *if*

1. $\mathbf{K} \cup \mathbf{L} = \mathbf{M}$*;*
2. $\mathbf{K} \neq \mathbf{M}$, $\mathbf{L} \neq \mathbf{M}$*;*
3. $\pi(\mathbf{M}) \cdot \pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}$.

Notice that the third condition is nothing else than $\mathbf{K} \setminus \mathbf{L} \perp\!\!\!\perp \mathbf{L} \setminus \mathbf{K} | \mathbf{K} \cap \mathbf{L}\ [\pi]$, and that the original distribution $\pi(\mathbf{M})$ can be uniquely reconstructed from the marginals $\pi^{\downarrow \mathbf{K}}$ and $\pi^{\downarrow \mathbf{L}}$.

It is important that probability distributions can be hierarchically decomposed into a system of low-dimensional distributions that cannot be further decomposed. An example of such a hierarchical process represented by a corresponding tree structure can be seen in Figure 1, where distribution $\pi(X, Y, Z, V, W)$ is decomposed into a system of its marginal distributions: $\pi^{\downarrow X}$, $\pi^{\downarrow Y}$, $\pi^{\downarrow \{Y,Z,V\}}$, $\pi^{\downarrow \{Z,V\}}$, $\pi^{\downarrow \{V,W\}}$. Each decomposition was made possible by the fact that the respective conditional independence relation holds for distribution π. It means that the decomposition process from Figure 1 was made possible by the assumption that the following system of conditional independence relations holds for distribution π:

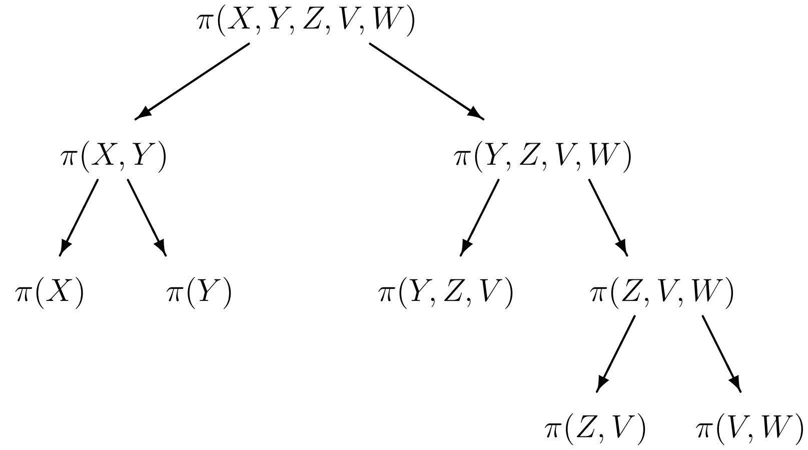

Fig. 1: Hierarchical decomposition of $\pi(X, Y, Z, V, W)$.

- $X \perp\!\!\!\perp \{Z, V, W\} | Y\ [\pi]$;
- $X \perp\!\!\!\perp Y\ [\pi]$;
- $Y \perp\!\!\!\perp W | \{Z, V\}\ [\pi]$;
- $Z \perp\!\!\!\perp W | V\ [\pi]$.

Therefore, having a multidimensional probability distribution (a generator of the considered data) decomposed into its “primes” (low-dimensional distributions that cannot be further decomposed), we are able to express all the knowledge contained in the distribution by

- the list of conditional independence relations enabling the decomposition of the considered multidimensional distribution;
- all the knowledge that can be read from the low-dimensional “prime” distributions.

The only problem is that the process of decomposition of a multidimensional distribution (represented by the considered data file) is, as a rule, computationally intractable. Therefore we have to find an indirect process that yields a decomposition of such a multidimensional distribution, or, which is more realistic, to find an approximation of such a hierarchical decomposition process. And the description of such a process forms content of the rest of the paper.

4 Compositional models

The basic idea is as follows: If we cannot decompose a multidimensional distribution because of the great computational complexity of the respective process, we find the approximation of the considered distribution, which is in a form of a compositional model. It means that its decomposition is "visible" from the structure of the model. Compositional models are multidimensional distributions *composed* from a system of low-dimensional distributions by an operator of a composition realizing an inverse process to the decomposition defined in section.

Recall that $\pi(\mathbf{N})$ can be decomposed into its marginals $\pi(\mathbf{K})$ and $\pi(\mathbf{L})$ if $\mathbf{K} \cup \mathbf{L} = \mathbf{N}$ and $\pi(\mathbf{N}) \cdot \pi^{\downarrow \mathbf{K} \cap \mathbf{L}} = \pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}$. From this, one immediately gets that an inverse operation, the operation of composition is

$$\pi(\mathbf{N}) = \frac{\pi^{\downarrow \mathbf{K}} \cdot \pi^{\downarrow \mathbf{L}}}{\pi^{\downarrow \mathbf{K} \cap \mathbf{L}}}.$$

This is trivial in case that we compose distributions $\pi(\mathbf{K})$ and $\pi(\mathbf{L})$ that are consistent. The question is whether one can compose also inconsistent distributions, i.e., distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$, for which $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} \neq \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$. We need it because the estimates got from a data file with missing values are rarely consistent. Therefore, we advocate the following definition that was first introduced in [7].

Definition 2. *For arbitrary two distributions $\kappa(\mathbf{K})$ and $\lambda(\mathbf{L})$, for which $\lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$ dominates $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}}$, their* composition *is for each $x \in \mathbb{X}_{\mathbf{K} \cup \mathbf{L}}$ given by the following formula*[2]

$$(\kappa \triangleright \lambda)(x) = \frac{\kappa(x^{\downarrow \mathbf{K}})\lambda(x^{\downarrow \mathbf{L}})}{\lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}(x^{\downarrow \mathbf{K} \cap \mathbf{L}})}.$$

In case that $\kappa^{\downarrow \mathbf{K} \cap \mathbf{L}} \not\ll \lambda^{\downarrow \mathbf{K} \cap \mathbf{L}}$ the composition remains undefined.

By a *multidimensional compositional model* we understand a multidimensional probability distribution assembled from a sequences of low-dimensional distributions with the help of the introduced operator of composition, i.e.,

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n.$$

[2] Define $\frac{0 \cdot 0}{0} = 0$.

Unfortunately, the operator of composition is not associative, and therefore the above expression is ambiguous. Therefore, let us make a convention that we will omit the parentheses if the operators are to be performed from left to right:

$$\kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n = (\ldots((\kappa_1 \triangleright \kappa_2) \triangleright \kappa_3) \triangleright \ldots \triangleright \kappa_{n-1}) \triangleright \kappa_n.$$

On the other side, it is important that for these models, similarly to Bayesian networks, efficient computational algorithms were designed that make the application of these models for the inference possible. There are algorithms for the marginalization of compositional models and for computation of conditional distributions. In this paper, we are not interested in this type of applications, and therefore we do not need to present all the properties of the operator of composition, which make the theoretical basis for the design of these computational procedures. At this place, let us highlight just that this operator is generally neither commutative nor associative. Nevertheless, the associativity of the operator of composition would be desirable not only to meet the requirements of mathematical beauty, and to make the design of computational algorithms easier, but also to support some steps of the data-based model construction. Its lack is, in a way, compensated by the existence of a generalized operator of composition, which is called an anticipating operator. It is introduced in the following definition.

Definition 3. *Consider an arbitrary set of variables* $\mathbf{M}$ *and two distributions* $\kappa(\mathbf{K})$*,* $\lambda(\mathbf{L})$*. Their* anticipating composition *is given by the formula*

$$\kappa \textcircled{$\triangleright$}_{\mathbf{M}} \lambda = (\lambda^{\downarrow(\mathbf{M}\setminus\mathbf{K})\cap\mathbf{L}} \cdot \kappa) \triangleright \lambda.$$

Notice, it is a generalization of the operator introduced in Definition 2 in the sense that

$$\kappa \textcircled{$\triangleright$}_{\emptyset} \lambda = \kappa \triangleright \lambda.$$

Therefore, it is clear that it may happen that the result of composition remains undefined. However, it follows immediately from the respective definitions that if $\kappa \triangleright \lambda$ is defined than also $\kappa \textcircled{$\triangleright$}_{\mathbf{M}} \lambda$ is defined. Both $\kappa \triangleright \lambda$ and $\kappa \textcircled{$\triangleright$}_{\mathbf{M}} \lambda$ are distributions defined for the same set of variables.

The main significance of the operator $\triangleright$ is in the following: If $\kappa(\mathbf{K})$, $\lambda(\mathbf{L})$ and $\mu(\mathbf{M})$ are such that $\mu \triangleright (\kappa \textcircled{$\triangleright$}_{\mathbf{M}} \lambda)$ is defined, then

$$(\mu \triangleright \kappa) \triangleright \lambda = \mu \triangleright (\kappa \textcircled{$\triangleright$}_{\mathbf{M}} \lambda).$$

The reader interested in other theoretical issues concerning the operator of composition is referred to [11] and the papers cited there.

5 Heuristics for Model Construction

As said at the beginning of Section 4, having data, we want to construct a compositional model approximating the distribution that generated the data.

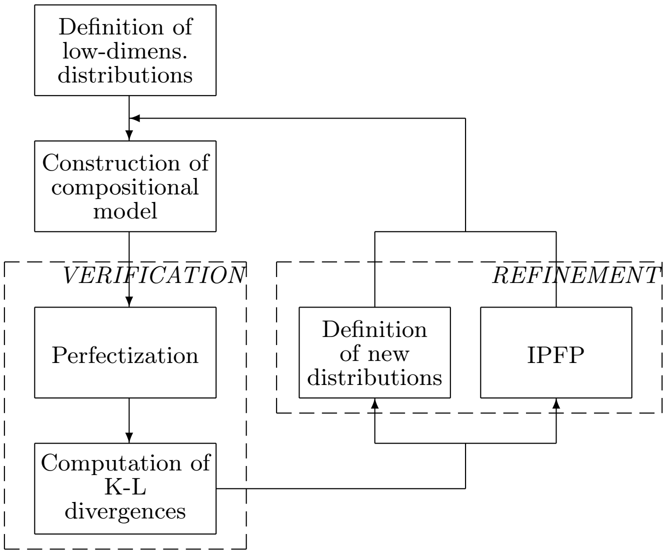
Definition of low-dimens. distributions
Construction of compositional model
VERIFICATION
Perfectization
Computation of K-L divergences
REFINEMENT
Definition of new distributions
IPFP

Fig. 2: Process of model construction

This model is a bearer of the knowledge from the data. Recall the two types of knowledge mentioned at the beginning of Section 3 that can be directly read from a compositional model: First, it is the list of conditional independence relations holding for the distribution represented by the model[3], second, each of the low-dimensional distributions, from which the model is composed, can be interpreted in the way mentioned in the introductory section.

Similarly to, for example, Bayesian network construction, there is no generally accepted "best" approach to data-based compositional model construction. For the purpose of data mining, one possibility is to use a heuristic procedure[4] schematically depicted in Figure 2. Notice that the described process is fully controlled by an expert, who has a possibility to influence the constructed model, and thus consequently also the type of the received knowledge.

As can be seen from the diagram in Figure 2, the process is initiated with the definition of a system of low-dimensional distributions. Regarding the fact that the process cyclically employs steps of *verification* and *refinement*, during which this initial system is gradually changed, the result is fairly independent of the initial selection. For example, starting with all two-dimensional distributions may be quite reasonable (for application to small data files with a limited number of variables one can consider a possibility to start with three-dimensional marginal distributions). In other situations, an expert can select the initial marginal dis-

[3] Instructions for reading all the conditional independence relation from the structure of the model can be found in [12].

[4] For a more detailed description of this process, as well as for the survey of the necessary theoretical background, the reader is referred to [10].

tributions from which the model should be constructed. Generally, we propose to select distributions carrying a greater amount of information. This idea is supported by the assertion, proved in [9] (Corollary 1.). It claims that the higher information content of a compositional model, the better approximation of the unknown distribution. This means that it is necessary to compute the information content (a generalization of the famous *Shannon mutual information*) of a multidimensional distribution, which is for a distribution $\pi(\mathbf{K})$ defined[5]

$$IC(\pi) = Div(\pi \| \prod_{X\in\mathbf{K}} \pi^{\downarrow\{X\}}) = \sum_{x\in\mathbb{X}_{\mathbf{K}}} \pi(x) \log \frac{\pi(x)}{\prod\limits_{X\in\mathbf{K}} \pi^{\downarrow\{X\}}(x^{\downarrow\{X\}})}.$$

It is important to highlight here, that computation of this value is computationally cheap for, so called, perfect models. Therefore, the process of "perfectization" of the model is included into the process of model construction. For perfect models, the information content of the whole multidimensional distribution can be computed from the information content of the individual low-dimensional distributions. This is why the computation of this value for the model is cheap, and also why we want to get low-dimensional distributions with high information content.

Realization of the box "Computation of Kullback Leibler divergence" in Figure 2 means computing the divergence between the distribution given by data and the distribution represented by a model. It can easily be done for perfect models. It helps the user to decide whether the model reasonably approximates the data distribution. Naturally, one cannot expect that the first choice of the low-dimensional distributions would yield a satisfactory model. This is true even more in situations when starting just with two-dimensional distributions. Comparing the data distribution and the model locally (i.e., computing the Kullback Leibler divergence for marginals corresponding to the low-dimensional distributions, from which the model is set up), the user can see, which parts of the model do not reflect the data properly. It usually means that it is necessary to increase the dimension of some low-dimensional distributions. It can be done, as one can see in Figure 2, in two ways. Having enough data, one can get just new estimates from the data. However, quite often it may be better to get a maximum entropy estimate from several low-dimensional distributions of the original model by the *Iterative Proportional Fitting Procedure* [3].

Let us stress once more that the process in Figure 2 is fully controlled by the expert. The more cycles of the process are performed, the higher dimensions of the input distributions are considered. If the expert had continued ad absurdum, the process would have, in fact, finished with an application of IPFP to all of the initial low-dimensional distributions (which is, as a rule, computationally

[5] $Div(\pi\|\mu)$ denotes the famous Kullback-Leibler divergence defined (for $\pi(\mathbf{K})$ and $\mu(\mathbf{K})$)

$$Div(\pi\|\mu) = \sum_{x\in\mathbb{X}_{\mathbf{K}}} \pi(x) \log \frac{\pi(x)}{\mu(x)}.$$

intractable in practical situations). Therefore, it is obvious that one has to avoid the overfitting of the model.

A *model overfitting* is a well known phenomenon both in statistics [17] and machine learning (artificial intelligence) [1]. It is used to describe a situation when a constructed model reflects the noninformative properties of the source data file (like noise and other misleading properties that each randomly generated data file possesses). Let us illustrate it on two stochastically dependent variables, the dependence of which is known to be linear. Because the dependence is stochastic, if randomly generated data are plotted in a graph, the respective dots are concentrated along a straight line describing the dependence. Naturally, only a part of dots lies directly on the line. If one tries to find a curve that connects all the dots in the plot (see Fig. 3), the model becomes for knowledge discovery useless. Realize that such a complex curve is described (defined) by a much larger number of parameters than the straight line, which can be determined just by two points.

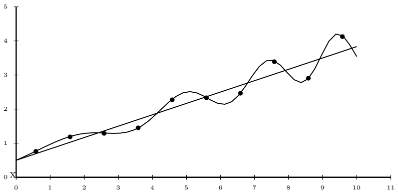

Fig. 3: Overfitted linear dependence

In agreement with other parts of this section, we propose a solution to this problem from the perspective of information theory. Namely, the process of model construction can be viewed as a transformation process; a process transforming the information contained in data into the information represented by a model. Thus, using one of the basic laws of information theory saying that any transformation cannot increase the amount of information, we get the basic restriction laid on models constructed from data: A model is *acceptable* if it does not contain more information than the input data file.

However, the application of this idea hits the problem, how to measure information in a data file, and how to measure information contained in a model. For this, we go back more than a half a century to seminal papers by von Mises [19] and Kolmogorov [14], who explored relations interconnecting randomness, complexity and information. They came with the idea that the amount of information in a sequence of 0's and 1's is increasing with the complexity of the sequence, and that the complexity of such a sequence can be measured by the length of the shortest program[6] generating the sequence. We accept here this

[6] An abstract program for a universal Turing machine.

idea but instead of the length of an (abstract) program we consider the length of a lossless encoding (one can always generate the sequence from its lossless encoding).

Therefore, in agreement with results of Kolmogorov and von Mises, before accepting a final compositional model, we look for the shortest possible encoding of both, data file and the resulting model. In case we get a model, the encoding of which requires more bits than the encoding of data, we are sure, that some undesirable information has been added into the model. In addition to this, we know that regardless of the way the data were collected, they always contain some specific part of the information, employment of which results in the overfitting of the model. It should not be included in the model. Thus we enforce a principle under which we accept only models, the encoding of which is substantially shorter than the encoding of the data file. The meaning of the word substantially is usually left to the user's discretion.

6 Conclusions

In this paper we have described main ideas on which probabilistic data-mining methods are based. The term "general" in the title refers to the fact that the approach can be applied to any data that may be assumed to be generated by a random generator – multidimensional probability distribution. This is also why the paper is focused on knowledge describing behavior of such generators. This knowledge may be either qualitative or quantitative. The former can be characterized by the independence structure of a multidimensional probability distribution, by a list of the conditional independence relations holding for the probability distribution. It can express which of the linkages between (among) the considered variables (features, characteristics) are direct, and which are mediated by other variables.

In this contribution, we have presented principles that can be employed in the process of knowledge discovery from data. Naturally, there are many other approaches for this purpose. Our approach is based on the assumption, that we are looking for the knowledge that can be read from a multidimensional probability distribution (data generator) by decomposing the distribution into its "prime" marginals. It takes advantage of the fact that the notion of conditional independence (sometimes also called conditional irrelevance) is a notion used when explaining a knowledge in plain language. Similarly, the knowledge encoded in low-dimensional probability tables can always be expressed with the help of conditional probabilities, i.e., probabilistic (indeterministic) implications.

For this, we had to introduce the concept of compositional models. Naturally, in many places and in particular, in connection with the process of data based model construction, we could only present the main ideas. The interested reader is referred either to original journal papers or to the book [6], which is to be published in 2019 as the first summarizing text on probabilistic compositional models. Let us mention at this place that a number of papers were written also on the compositional models in other uncertainty theories, like possibility theory, belief function theory, and even on compositional models in Shenoy's valuation based systems [18, 13].

Acknowledgement

The research was financially supported by grants GAČR no. 19-06569S (first author) and AV ČR no. MOST-04-18 (second author).

References

1. Petr Berka, *Dobývání znalostí z databází*, Academia, 2003.
2. Peter C Cheeseman, *In defense of probability.*, IJCAI, vol. 2, Citeseer, 1985, pp. 1002–1009.
3. W Edwards Deming and Frederick F Stephan, *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, The Annals of Mathematical Statistics **11** (1940), no. 4, 427–444.
4. Daniel Ellsberg, *Risk, ambiguity, and the savage axioms*, The quarterly journal of economics (1961), 643–669.
5. David J Hand, Heikki Mannila, and Padhraic Smyth, *Principles of data mining (adaptive computation and machine learning)*, MIT Press Cambridge, MA, 2001.
6. Kratochvíl Václav et al. Jiroušek, Radim, *Discrete probabilistic models for data mining*, MatfyzPress, to be published in 2019.
7. Radim Jiroušek, *Composition of probability measures on finite spaces*, Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1997, pp. 274–281.
8. ______, *Decomposition of multidimensional distributions represented by perfect sequences*, Annals of Mathematics and Artificial Intelligence **35** (2002), no. 1-4, 215–226.
9. ______, *On approximating multidimensional probability distributions by compositional models.*, ISIPTA, 2003, pp. 305–320.
10. ______, *Data-based construction of multidimensional probabilistic models with mudim*, Logic Journal of the IGPL **14** (2006), no. 3, 501–520.
11. ______, *Foundations of compositional model theory*, International Journal of General Systems **40** (2011), no. 6, 623–678.
12. Radim Jiroušek and Václav Kratochvíl, *Foundations of compositional models: structural properties*, International Journal of General Systems **44** (2015), no. 1, 2–25.
13. Radim Jiroušek and Prakash P Shenoy, *Compositional models in valuation-based systems*, Belief Functions: Theory and Applications, Springer, 2012, pp. 221–228.
14. A N Kolmogorov, *Tri podchoda k kvantitativnomu opredeleniju informacii*, Problemy peredachi informacii (1965), no. 1, 4–7.
15. Steffen L Lauritzen and David J Spiegelhalter, *Local computations with probabilities on graphical structures and their application to expert systems*, Journal of the Royal Statistical Society. Series B (Methodological) (1988), 157–224.
16. Judea Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, (1988).
17. Thomas P Ryan, *Modern regression methods*, vol. 655, John Wiley & Sons, 2008.
18. Prakash P Shenoy, *A valuation-based language for expert systems*, International Journal of Approximate Reasoning **3** (1989), no. 5, 383–411.

19. Richard Von Mises, *Probability, statistics, and truth*, Courier Corporation, 1981 [Originaly published in German by Springer, 1928].