Probabilistic Aspects of Score Systems

Manfred Schramm
Institut für Informatik
Technische Universität München
schramm@pit-systems.de

Bertram Fronhöfer
Institut für Informatik
Technische Universität Dresden
fronhoefer@pit-systems.de

Keywords: Score Systems, Probabilistic Reasoning, Maximum Entropy, Automated Diagnosis, Independence

Abstract

Due to their simple applicability score systems are in widespread use as a tool for decision taking.

Unfortunately, as we all feel, they are somehow not apt to take into account interdependencies among the variables (symptoms/attributes) which are input to them when trying to decide an actual case; a drawback which is overcome by probabilistic systems. In order to analyze which assumptions are inherent in score systems we translate them into probabilistic systems — a more powerful framework — thus making available their technical machinery for this analysis task.

Such a straightforward translation (as also given in [KR00]) reveals some properties of score systems, but leads to an exponential number of probabilistic rules. For this reason we also developed several further translations into probabilistic systems, which keep the simplicity of a score system (i.e. they use the same amount of rules as the score system). Moreover, the resulting probabilistic systems show their structure more explicitely than score systems, and they are also open to the addition of further knowledge.

1 Introduction

Medicine, industry and economy abound with problems of diagnosis (and decision). Illustrated by an example from medicine, such a problem consists of a knowledge base of propositions about the problem domain (e.g. 'Appendicitis is usually accompanied by strong stomach aches'), the values of certain symptoms in an actual (diagnosis) case (e.g. 'The patient has stomach aches'), a wanted diagnosis (e.g. 'Does the patient suffer from appendicitis?'), and finally, a decision ('surgical intervention?').

To solve such problems, **Score Systems** are frequently developed in research labs, e.g. in medicine ([OYF95]), or are in wide-range use, e.g. in economics ([KR00]), whenever uncertain knowledge plays an important role with the kind of problem to be solved.

> A Score System is based on a set of **attributes** (or **variables**) which have each a set of possible (attribute or variable) **values**.
>
> > For instance, in the medical domain, there may be the symptom/attribute 'body temperature' with (discrete) values 'low', 'normal', 'high' and 'very high'. To each attribute value a numerical value — its **weight** or **score**— is assigned (see Table 1).

When applying a score system to a concrete case the scores corresponding to the observed attribute values are added up. If the obtained sum falls in a certain **score interval**, the decision associated to this interval is proposed.

> Thus, for instance, a proposal for a medical treatment is established on the basis of symptoms found with a patient and which are represented by a list of attribute values.

Example 1:

symptom / attribute	score, if yes
tenderness in RLQ ‘	4.5
rebound tenderness	2.5
no micturition	2.0
continuous type of pain	2.0
number of leucocytes ≥ 10000	1.5
age < 50 years	1.5
relocation of pain to RLQ	1.0
rigidity	1.0

Table 1: 'Ohmann Score' [OFY$^+$95] for the diagnosis of appendicitis: In case of negative answers the scores are zero. Patients are diagnosed as having appendicitis if score sum ≥ 12, they are interned in case of $6 - 12$, and are sent home in case of ≤ 6. (RLQ: right lower quadrant of abdomen (as seen from the patient).)

When applying a score system to an actual case of decision finding, we feel intuitively that a score system seems to make some kind of assumption about the unrelatedness of the variables, because it provides no means to adapt the scores assigned to the values of one variable in dependence of the selected values of other variables. This makes us feel uneasy, because very often we are intuitively aware of influences between some of the variables, but by using a score system we are bound to the use of a tool where these relationships are not taken into account. Sometimes this feeling is expressed in the opinion that score systems implicitly assume marginal independence of variables/attributes.[1]

The first aim of this paper is to analyze rigorously our intuitive feelings about score systems and to explicit the hidden assumptions underlying them. To this end we translate scores systems into probabilistic systems as faithfully as possible thus being able to exploit the technical framework of probability theory for the investigation of independencies and indifferences. Being convinced of the faithfulness of our translation, we claim, that our discoveries in the obtained probabilistic systems shed light on the nature of score systems.

[1] In the "Handwörterbuch der Wirtschaftswissenschaft" (handbook of economy) the keyword "Produktplanung" (product planing) contains the following contribution by Dietrich Adam: "Die einfachste Form der Synthese der Urteile über einzelne Kriterien besteht in der Addition der gewichteten Punktzahlen aller Kriterien. Diese Art der Synthese hat jedoch unabhängige Kriterien zur Voraussetzung, d.h. ein positives Urteil bei einem Kriterium darf nicht mit dem Urteil bei anderen Kriterien korrelieren." (The simplest form of synthesizing the judgments of individual criteria consists in adding the weighted scores of all criteria. However, independent criteria are a prerequisite for this kind of synthesis, i.e. a positive judgment of one criterion must not correlate with the judgments of other criteria.)

This first aim is developed in sections 2 till 5: In section 2 we give a formal definition of score systems. In section 3 we present some general considerations on diagnosis systems and emphasize the need to model explicitly the relationship between diagnoses and symptoms. This leads naturally to the introduction of some terminology from probability theory, namely the terminology of event spaces, and culminates in the definition of a probabilistic diagnosis system in section 4. It will turn out that a score system is a little bit deficient in view of these considerations, but some slight extensions turn it quite directly into a probabilistic system. In section 5 this faithful translation $\mathcal{T}_{\vec{s},d}$ is presented and analyzed. It is defined via constraints on a probability measure, which are derived from a score system.
However, there is also a futher aim of this paper which is dealt with in section 6. The translation given in section 5.1 has the only purpose to analyse score systems. The resulting probabilistic system is not handy for practical use, since the number of rules grow exponentially. That this growth is not mandatory is shown by 2 further translations $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$, which are consistent and decision preserving, and where, in addition, the size of the rule base is linear in the cardinality of the Score System's variables and their respective cardinalities.
The paper concludes with a comparative view of the difference of score systems and probabilistic systems.

Examples in this paper are often taken from the field of medical diagnosis. Often our general explanations are tinged by medical language, which — in our opinion — better adds to clarity, than a more neutral, but less suggestive language.

The paper is a condensed version of [FS01] and [SF01] to which we refer for further details.

2 Score Systems

Formally, a Score System can be defined as follows:
It consists of a set of **variables** (**attributes**) S_i $(i = 1, \dots, m)$. Each S_i can be identified with its set of variable **values** $\{s_{i1}, \dots, s_{ik_i}\}$ $(k_i > 1)$. We denote by $\vec{s}$ a tuple of values $\langle s_1, \dots, s_m \rangle$ with $s_i \in S_i$.
Moreover, for each variable S_i exists a set $W_i = \{w_{i1}, \dots, w_{ik_i}\}$ of nonnegative **weights** or **scores** and a bijective **score function** $w_i \colon S_i \longrightarrow W_i$. We also have a **(global) score function** w defined as $w(\vec{s}) := \sum_{i=1}^{n} w_i(s_i)$.
Finally, there are **score intervals** given by a set of **border values** $b_1 < \cdots < b_{k_T}$, a **decision variable** T with values $\{t_1, \dots, t_{k_T}\}$ and a **decision function** t which maps a sum of scores $w(\vec{s})$ to t_i iff $b_{i-1} < w(\vec{s}) \le b_i$ (with $b_0 := -\infty$).

Since decision functions are not the topic of this paper, we will skip any explicit treatment and just show that the border values can be adapted when translating score systems. Basically, within score systems decision functions are a very simple add-on to the kernel of a score system and consequently neglecting them in the sequel is no restriction. More sophisticated decision methods which are available with probabilistic systems cannot be adapted to score systems, because they require multiple diagnoses which score systems are not able to provide.

Example 2: A simple example of a score system, which we will reuse in the following, consists of 3 binary variables S_i ($i = 1, 2, 3$) — whose values we denote as $S_i = \{s_i, \overline{s_i}\}$ — together with the 3 score functions w_i which map S_i to a respective (binary) set of values $W_i = \{i, 0\}$ with $w_i(s_i) = i$ and $w_i(\overline{s_i}) = 0$. ■

For a further example including a decision function see Example 1 above.

When attempting to translate a formal system into another one, the crucial point is to obtain what might be called a homomorphic image, thus assuring faithfulness. For this purpose the 'source system' should be investigated for formal properties which should be conveyed into the 'target system' via the translation.
In score systems we only discovered the following **contribution invariance** property: For arbitrary $s_{i1}, s_{i2} \in S_i$ there exists a constant $\mathrm{const}_{\mathrm{s_{i1}\to s_{i2}}}$ such that

$$w(\vec{s}) - w(\vec{s}\,[s_{i1}{\to}s_{i2}]) = \mathrm{const}_{\mathrm{s_{i1}\to s_{i2}}} \tag{1}$$

holds for all $\vec{s}$ in which s_{i1} occurs. (We denote by $\vec{s}\,[s_{i1}{\to}s_{i2}]$ the tuple obtained from $\vec{s}$ when replacing s_{i1} by s_{i2}.) From the easy proof of this property, it is clear, that in a more 'axiomatic' formulization the subtraction is the appropriate instance of the inverse of the function which computes the overall score $w(\vec{s})$ from the scores of the symptom values, i.e. addition in the case of a score system.

3 Diagnosis Systems: Basic Considerations

Before we present a translation of score systems into probabilistic systems we will present some general thoughts on modeling systems for diagnosis and establish some general requirements.

3.1 Diagnosis-Attribute Relationship

Any diagnosis system must model the relationship between diagnoses (diseases) and attributes (symptoms) relevant for these diagnoses. In order to express knowledge about the evidence of a disease in view of certain symptoms as well as to express perhaps quite different knowledge about the occurrence of symptoms in case of a disease, diagnoses must be represented explicitly. This means that we have a finite set of **variables** — symptom variables and, in addition, diagnosis variables — with each having a finite set of **values**.[2] The **symptom variables** describe properties / symptoms / attributes relevant for the diagnosis task, e.g. examination results in the medical case.

> As before, we denote symptom variables by S_i ($1 \leq i \leq m$), and we identify them with the set of their values: $S_i = \{s_{ij} | 1 \leq j \leq k_i\}$.
> We sometimes refer to values of variable S_i just by s_i, s'_i ...

The values of a **diagnosis variable** define a classification of the possible diagnostic results, e.g. in kinds of diseases, based on the values of the symptom variables.

[2] Infinite numbers of symptoms or diagnoses cannot be taken into account in a real world application and an infinite range of values can be made finite in practice through appropriate discretization.

In the following we will only consider the case of a single diagnosis variable $D = \{d_1, \ldots, d_{k_D}\}$.[3]

The considerations presented above lead naturally to the view that a diagnosis system must model a relation on the tuple space $\Omega := S_1 \times \cdots \times S_m \times D$.
This relation can be specified by a **method of judgment**. Depending on this method we may either get a 'classical' relation or a somehow fuzzy one, i.e. we may have yes-no judgments on tuples $\langle s_{1j_1}, \ldots, s_{mj_m}, d_h \rangle$ or a more fine-grained judgment, e.g. scores or probability measures.

3.2 Uncertain Findings

It may occur quite often that the value of a certain symptom, say S_1, cannot be observed. This means that a method of judgment, which judges only individual tuples, is very restricted in view of practical applications, as we might be interested in getting a judgment of a set of tuples from Ω as, for instance,

$$\bigcup_{j_1=1}^{k_1} \langle s_{1j_1}, s_{2j_2}, \ldots, s_{mj_m}, d_h \rangle$$

which reflects ignorance of the value of S_1. Of course, there may also be cases where some values of S_1 can be excluded, while still not arriving at a single remaining variable value, thus obtaining a smaller set than the one given above, while still not getting a single tuple. Consequently, judgments on arbitrary subsets of Ω are desirable.

In the terminological tradition of probability theory, we say that we consider Ω as an **event space** with its power set as **set of events** or **event algebra**.
For a symptom value $s_i \in S_i$ of a symptom variable or an element $d_h \in D$ of a diagnosis variable we denote $\langle s_i \rangle := S_1 \times \cdots \times S_{i-1} \times s_i \times S_{i+1} \times \cdots \times S_m \times D$ resp. $\langle d_h \rangle := S_1 \times \cdots \times S_m \times d_h$. By extension, we denote by $\langle s_{i_1}, \ldots, s_{i_n} \rangle$ with $s_{i_j} \in S_{i_j}$ the intersection of the sets $\langle s_{i_j} \rangle$.
Consequently, an expression $\langle s_1, \ldots, s_m, d_h \rangle$ with $s_i \in S_i$ and $d_h \in D$ is an **elementary event** in Ω.[4] (All general **events** are sets of elementary events.)
For a value $s_i \in S_i$ we call $\langle s_i \rangle$ a **simple event**. We denote by $\vec{s}$ the event $\langle s_1, \ldots, s_m \rangle$ with $s_i \in S_i$, which corresponds to the set $\{s_1\} \times \ldots \times \{s_m\} \times D$, and it will be called an **elementary symptom event**. In addition, we denote by $\langle \vec{s}, d_h \rangle$ the elementary event $\langle s_1, \ldots, s_m, d_h \rangle$ with $d_h \in D$.
We also write $E \longrightarrow\!\!\!\triangleright\, E'$ for the **conditional event** $E'|E$ due to some similarity with common-sense implication. We make the convention to drop the parentheses if simple events occur in conditional events.

[3] This constitutes no restriction, because the case of a set of diagnosis variables can always be coded as a single one with values reflecting the possible combinations of the values from the set.

[4] Sometimes an elementary event is also called a **full conjunction** which refers to the suggestive notation of an elementary event as a conjunction of statements $S_1{=}s_1 \wedge \ldots \wedge S_m{=}s_m \wedge D{=}d_h$. Extending this view, we also write $E \wedge E'$, $E \vee E'$, $\neg E$ instead of $E \cap E'$, $E \cup E'$ and the complement $\overline{E}$ of events E and E'.

In addition to the event space, we require a **method of judgment** to be given, e.g. a **judgment function** on (all) the events.

3.3 Conditional Judgment

Let's assume now, just for simplicity, that in an actual case, e.g. a particular patient to be examined, we were able to determine our symptoms' values uniquely, i.e. we are given an event $\vec{s}$ and we want to get a judgment on whether this patient suffers from disease $d_h \in D$. (In the general case of missing uniqueness we just get a tuple set instead of a single tuple $\vec{s}$.)
In a concrete case where we face the need of a diagnosis of disease d_h

- our interest will not focus on the judgment of the event $\langle \vec{s}, d_h \rangle$ in comparison to all other (elementary) events in Ω — i.e. in comparison to events based on quite different combinations of symptom values which are already excluded in the case of the actual patient
- but our interest will focus on the judgment of the event $\langle \vec{s}, d_h \rangle$ in comparison to other diseases in view of the same symptom values $\vec{s}$.

This leads to our interest in the judgment of the conditional event $\vec{s} \longrightarrow\!\triangleright \langle \vec{s}, d_h \rangle$ in comparison to the judgment of a conditional event $\vec{s} \longrightarrow\!\triangleright \langle \vec{s}, d_g \rangle$ with another disease d_g. As a special case we consider $d_g = \overline{d_h}$ where $\overline{d_h}$ means the absence of disease d_h. Note that in the medical context the common reading of the desire to come to a judgment of $\vec{s} \longrightarrow\!\triangleright \langle \vec{s}, d_h \rangle$ is:

> 'If I know a patient showing the symptom values $\vec{s}$,
> what can I say — in view of this knowledge —
> about his risk of having the illness d_h ?'

Convention: For sake of simplicity we will just write $\vec{s} \longrightarrow\!\triangleright d_h$ instead of $\vec{s} \longrightarrow\!\triangleright \langle \vec{s}, d_h \rangle$.

3.4 From Diagnosis to Decision

The distinction between diagnosis and decision is by no means trivial, since uncertainties of the diagnosis as well as the desirable minimization of costs of the treatment have to be taken into account. For instance the common decision 'to keep the patient for further observation' would make no sense if the right and unique diagnosis were known to the doctor. Only a diagnosis where neither illness nor sanity obtain a decisively high judgment lead to this decision.
We skip the discussion of decision functions, because they lie out of focus of this paper (see Table 1 for an example).

4 Probabilistic Diagnosis Systems

Probabilistic Diagnosis Systems — or Probabilistic Systems for short — are a special case of the diagnosis systems just described.

- They use as method of judgment a function P, which assigns probabilities to events in compliance with the laws of probability theory.

  A **P-measure** P is a function on an event algebra which — in our finite case — can be specified by mapping every elementary event to a nonnegative real number such that the sum of function values over all elementary events is equal to 1. Since every event E is a (unique) union of elementary events $e_1, \ldots, e_n$ we define $P(E) := \sum_{i=1}^{n} P(e_i)$. A set together with a P-measure is called a **P-space**.

- In order to obtain a P-measure P, additional principles may be used to define it, if only incomplete knowledge about P is available. To obtain a unique P-measure is necessary in order to obtain unique probabilistic judgments.

  Two very interesting such principles are **indifference** (on subspaces), **independence** and the more powerful principle of **maximum entropy**[5] [PV90], the latter extending the principles of indifference and independence ([SG95]).

- The probabilistic character of the obtained judgments allows their combination with powerful decision procedures.

Notation:
As we have knowledge from different sources, we want to denote where our rules come from. We therefore use

- the sign *'c'* to mark a (quantitative) statement as constraint, obtained via a translation from a score, and collected in a certain knowledge base (see e.g. eq. (9)).
- the sign *'me'* to mark a (qualitative) statement, valid if the Maximum Entropy Method is applied to a certain knowledge base If this kind of knowledge is used in a proof, we refer to the number of the corresponding equation.
- the sign * to mark (quantitative) statements, which are valid under the maximum entropy distribution in an example for a certain knowledge base (see e.g. eq. Table 2 (right)).

5 Probabilistic Reconstruction of Score Systems

Score systems don't comply well with the considerations of section 3.

- They consider only a single disease which is not even explicitly represented, thus restricting their ability to model the relationship between symptoms and a disease (cf. sec. 3.1). They don't not even incorporate the opposite case — absence of disease — not to speak of comparisons with alternative diseases.
- Moreover, with $D = \{d\}$ the difference between $\langle \vec{s}, d \rangle$ and $\vec{s}$ vanishes, and consequently, they loose the possibility to judge the conditional event $\vec{s} \longrightarrow\!\!\triangleright \langle \vec{s}, d \rangle$ (cf. sec. 3.3).

[5]For short, the Maximum Entropy Method chooses a probability model with maximal (Shannon) entropy, which is defined on probability vectors v as $H(v) = -\sum_i v_i \cdot \log v_i$. (The base of the logarithm does not matter, in most cases ln is taken) . For constraints, linear in v (as in our cases here), the maximum of $H(v)$ is known to be unique.

- Finally, they just provide judgments of elementary symptom events and not of arbitrary ones (cf. 3.2). (See [FS01] on overcoming this restriction.)

5.1 Faithful Translation $\mathcal{T}_{\vec{s},d}$

To overcome the mentioned deficiencies of a score system we embed it into slightly larger probabilistic systems by deriving from score systems a set of constraints to be satisfied by a respective P-measure.

(1) In order to achieve a distinction between $\langle \vec{s}, d\rangle$ and $\vec{s}$ we introduce apart from disease d also the contrary diagnosis $\overline{d}$ — absence of disease d — and we extend the symptom space $\Sigma := S_1 \times \cdots \times S_m$ to $\Omega := S_1 \times \cdots \times S_m \times D$ with $D = \{d, \overline{d}\}$.

(2) We define a set of P-measures on Ω by understanding the score value of a (complete) tuple of symptom values $\vec{s}$ as a judgment of the conditional event $\vec{s} \longrightarrow\!\!\triangleright\, d$. Together with a normalization this leads to the **constraint** (indicated by ^{c})

$$^{c}P_{\vec{S},D}(\vec{s} \longrightarrow\!\!\triangleright\, d) := \frac{w(\vec{s})}{\widehat{w}_{max}} \tag{2}$$

for all $\vec{s} \in \Sigma$, where $\widehat{w}_{max} := \max\{w(\vec{s}) | \vec{s} \in \Sigma\}$. Since these constraints only make sense if

$$^{me}P_{\vec{S},D}(\vec{s}) > 0 \tag{3}$$

we require this **constraint** for all $\vec{s} \in \Sigma$ as well.

5.2 P-measure $P_{\vec{S},D}{}'$: Consistency of our Translation

Of course, the constraints (2) and (3) don't determine a unique P-measure. That they are consistent can be seen with the P-measure $P_{\vec{S},D}{}'$ defined as follows:

1. We extend the score function w to a function $\widehat{w}$ on the tuples of Ω by defining

$$\begin{aligned} \widehat{w}(\langle \vec{s}, d\rangle) &:= w(\vec{s}) && (4)\\ \widehat{w}(\langle \vec{s}, \overline{d}\rangle) &:= \widehat{w}_{max} - w(\vec{s}) && (5)\end{aligned}$$

2. By normalizing the function $\widehat{w}$ with the sum $\widehat{w}_{total} := \sum_{e\in\Omega} \widehat{w}(e)$ we get for all $e \in \Omega$

$$P_{\vec{S},D}{}'(e) := \frac{\widehat{w}(e)}{\widehat{w}_{total}} \tag{6}$$

Obviously, $P_{\vec{S},D}{}'$ is a P-measure since $P_{\vec{S},D}{}'(e) \geq 0$ for all $e \in \Omega$ and $\sum_{e\in\Omega} P_{\vec{S},D}{}'(e) = 1$. We get $P_{\vec{S},D}{}'(\vec{s}) = \widehat{w}_{max} > 0$ [6] and in addition

$$P_{\vec{S},D}{}'(\vec{s} \longrightarrow\!\!\triangleright\, d) = \frac{P_{\vec{S},D}{}'(\vec{s} \cap d)}{P_{\vec{S},D}{}'(\vec{s})} = \frac{P_{\vec{S},D}{}'(\langle \vec{s}, d\rangle)}{P_{\vec{S},D}{}'(\langle \vec{s}, d\rangle) + P_{\vec{S},D}{}'(\langle \vec{s}, \overline{d}\rangle)} = \frac{\frac{w(\vec{s})}{\widehat{w}_{total}}}{\frac{w(\vec{s})}{\widehat{w}_{total}} + \frac{\widehat{w}_{max} - w(\vec{s})}{\widehat{w}_{total}}} = \frac{w(\vec{s})}{\widehat{w}_{max}} \tag{7}$$

This means that the P-measure $P_{\vec{S},D}{}'$ fulfills the constraints (2) and (3).

[6] Recall that we excluded 'degenerated' score systems with $W_i = \{0\}$ for all $i \in \{1, \ldots, m\}$.

5.3 Preservation of decisions

From Constraint (2) follows that with our translation the probabilities of the diagnoses are the old scores normed by $\widehat{w}_{max}$, which implies the additional constraint

$$w(\vec{s}) > w(\vec{s}\,') \Longleftrightarrow {}^{c}P_{\vec{S},D}(\vec{s} \longrightarrow\!\!\triangleright\, d) > {}^{c}P_{\vec{S},D}(\vec{s}\,' \longrightarrow\!\!\triangleright\, d) \qquad (8)$$

for arbitrary events $\vec{s}$ and $\vec{s}\,'$ of symptom values. With $t_i/\widehat{w}_{max}$ as new border values a decision function of the score system can be easily adapted and we have equivalence of the decisions proposed by the score system and those proposed by a probabilistic system resulting from our translation.

5.4 Properties of our Translation

Our translation preserves the contribution invariance property: For $s_{i1}, s_{i2} \in S_i$ and $\forall\, \vec{s} \in \Omega$ in which s_{i1} occurs, follows with the contribution invariance (1) of score systems the following additional constraint

$${}^{c}P_{\vec{S},D}(\vec{s} \longrightarrow\!\!\triangleright\, d) - {}^{c}P_{\vec{S},D}(\vec{s}\,[s_{i1}\!\rightarrow\! s_{i2}] \longrightarrow\!\!\triangleright\, d) = \frac{w(\vec{s})}{\widehat{w}_{max}} - \frac{w(\vec{s}\,[s_{i1}\!\rightarrow\! s_{i2}])}{\widehat{w}_{max}} = \frac{\mathrm{const}_{\mathrm{s_{i1}\rightarrow s_{i2}}}}{\widehat{w}_{max}}$$

i.e. we get again a constant difference depending on s_{i1} and s_{i2} alone.

5.5 Independence Considerations

Since we translate score systems into probabilistic systems in a rather canonical way, it is reasonable to assume that findings about the resulting probabilistic systems shed considerable light on the nature of score systems.

An interesting topic which might be elucidated this way are the claims about existing independencies. This is a rather moot question open to speculation, as score systems provide no properly defined (in)dependence concepts nor is there any general definition of (in)dependence from which (in)dependence concepts for e.g. score systems can be derived by specialization. On the other hand, probability theory offers appropriate concepts and tools.

Of particular interest is the sometimes made affirmation that the symptoms must be marginally independent. However, this can be refuted.

Since marginal independence of symptoms cannot be established for ordinary score systems, and consequently cannot be responsible for this uneasiness, we may ask what else is the origin of the intuitive awareness of some kind of independence in score systems? We believe that this feeling stems from property (1) — contribution invariance. It seems that in part of the literature on score systems independence of symptoms has commonly been confused with some kind of independence of judgments about them.

Contribution invariance implies that, for instance, the number of leucocytes cannot be 'scored' differently in dependence of the age of a patient despite the fact that elder people have ordinarily lower leucocyte numbers than younger ones. Independent 'scoring' of leucocyte numbers leads to their underestimation as indication of a disease with elder

patients and to overestimation with younger ones. (Note as well that leucocyte number and age are *not* independent.)
Additional light is shed on this issue when considering diagnosis as a classifications task. An overall score can be understood as a *linear combination* of criteria and we know from Machine Learning that good classification can only be achieved if the data, we want to learn from, can be separated by *hyperplanes*. However, with many problems of diagnosis this setting is too simple: The criteria should either be combined in a non-linear way or non-linear separating surfaces should be modeled. Probability theory is an excellent tool for pursuing these directions.
For a detailed presentation and further investigations into questions of dependence and independence we refer to [FS01].

6 Non-Exponential Translations

While a score system has a linear number of rules ($\sum_i k_i$), the rules base of $\mathcal{T}_{\vec{s},d}$ contains an **exponential** number of rules ($\prod_i k_i$) and can therefore not be recommended for practical use. But as shown in e.g. [FS01], we cannot exploit any general independence relations in this translation (which is also true if the specification is completed via Maximum Entropy) to reduce the amount of rules. This makes $\mathcal{T}_{\vec{s},d}$ problematic for practical use.
We will therefore present in this section 2 further translations which do not have this defect, because in their case the exploitation of independence relations allows to reduce the number of necessary rules. These 2 translations lead to a consistent set of rules and preserve all decisions of the corresponding score system (compare eq. 8). For details of this statement we refer to [SF01].

6.1 Translation $\mathcal{T}_{d,s}$

The Translation Rule Translation $\mathcal{T}_{d,s}$ generates for each symptom value $s_{ij} \in S_i$ the following constraint [7]:

$$^{c}P_{D,S}(d \longrightarrow\!\!\triangleright\, s_i) \quad = \quad \frac{2^{w_i(s_i)}}{\sum_{s_{ij} \in S_i} 2^{w_i(s_{ij})}} \tag{9}$$

With these rules the P-measure $P_{D,S}$ is of course not uniquely determined and we need additional principles such as Maximum Entropy[8] for completing the specification.

Rules valid under the use of Maximum Entropy The following rules (indicated by ^{me}) are valid if we complete the specification by Maximum Entropy (and therefore will not be mentioned in our knowledge base):

- $^{me}P_{D,S}(d) > 0$ (Positivity)
  which is necessary to guarantee, that Eq. 9 is always defined.

[7]The proposed translation rule may be modified by varying the base value without affecting the following.

[8]See e.g. [PIT] for theory and implementation of the Maximum Entropy Method.

- As the rules of Eq. (9) introduce a difference between $P_{D,S}(\langle d, s_{ij}\rangle)$ and $P_{D,S}(\langle d, \overline{s_{ij}}\rangle)$, but no difference between $P_{D,S}(\langle \overline{d}, s_{ij}\rangle)$ and $P_{D,S}(\langle \overline{d}, \overline{s_{ij}}\rangle)$, the Maximum Entropy P-model obtained from the knowledge base will show 'conditional' indifference of the k_i values s_{ij}, i.e.
$$^{me}P_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, s_{ij}) = 1/k_i \tag{10}$$
- As the rule base does not contain a rule including more than one symptom at the same time (and thus does not draw a direct link between two or more symptoms in the dependence graph) the Maximum Entropy P-model obtained from the knowledge base (derived via Eq. (9)) satisfies (see e.g. [Sch96]) **conditional independence** of the symptom variables, given a value for D. Formally this means for $s_i \in S_i$

$$\begin{aligned} ^{me}P_{D,S}(d \longrightarrow\!\!\triangleright\, \vec{s}) &= \prod_{i=1}^{m} P_{D,S}(d \longrightarrow\!\!\triangleright\, s_i) \\ &=_{(9)} \frac{2^{\sum_{i=1}^{m} w_i(s_i)}}{\prod_{i=1}^{m}\sum_{s_{ij}\in S_i} 2^{w_i(s_{ij})}} = \frac{2^{w(\vec{s})}}{\prod_{i=1}^{m}\sum_{s_{ij}\in S_i} 2^{w_i(s_{ij})}} \end{aligned} \tag{11}$$

and for $\overline{d}$

$$^{me}P_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, \vec{s}) = \prod_{i=1}^{m} P_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, s_i) =_{(10)} 1/\prod_{i=1}^{m} k_i \tag{12}$$

Size of Rule Base When using $\mathcal{T}_{d,s}$ we obtain k_i rules for each variable S_i. Since each of these rules can be derived from the k_i-1 other rules, only k_i-1 rules must be specified for variable S_i, which means that we need to state $\sum_{i=1}^{m}(k_i-1)$ rules in case of m variables (which makes m rules for the case of m binary symptom variables as in our example of Tab. 2).

Example Continuing Example 2 we obtain (using Eq. (9)) the knowledge base of Table 2 (left). The probabilities for the queries $P^*_{D,S}(\vec{s} \longrightarrow\!\!\triangleright\, d)$ (calculated in the Maxent-Model $P^*_{D,S}$) are shown in Table 2 (right), additionally related to $w(\vec{s})$. [9] With Maximum Entropy we get in addition $P^*_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, s_i) = P^*_{D,S}(\overline{d} \longrightarrow\!\!\triangleright\, \overline{s_i}) = 0.500$ $(1 \le i \le 3)$ and $P^*_{D,S}(d) \approx 0.356$.

6.2 Translation $\mathcal{T}_{s,d}$

The Translation Rule With translation $\mathcal{T}_{s,d}$ we get for each symptom value $s_i \in S_i$ the following rule[10]:

$$^{c}P_{S,D}(s_i \longrightarrow\!\!\triangleright\, d) = \frac{2^{w_i(s_i)}}{1+2^{w_i(s_i)}} \tag{13}$$

[9]Please note that overlining is part of our succinct notation of values of binary variables and does not refer to a general concept of set complement. For the general case consider the translation rule (9).

[10]The proposed translation rule may be modified by varying the base value without affecting the following. For adjusting such a system in practice, also a shift to negative weigths is possible.

Knowledge Base

${}^{c}P_{D,S}(d \longrightarrow\!\!\triangleright\, s_1) = \frac{2^1}{2^1+2^0}) = \frac{2}{3} = 0.667$
${}^{c}P_{D,S}(d \longrightarrow\!\!\triangleright\, s_2) = \frac{2^2}{2^2+2^0}) = \frac{4}{5} = 0.800$
${}^{c}P_{D,S}(d \longrightarrow\!\!\triangleright\, s_3) = \frac{2^3}{2^3+2^0}) = \frac{8}{9} = 0.889$

(For the number of rules see (6.1))

Possible Queries : $P_{D,S}(\vec{s} \longrightarrow\!\!\triangleright\, d)$		$w(\vec{s})$
$P^*_{D,S}(\langle \overline{s_1}, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.032	0
$P^*_{D,S}(\langle s_1, \overline{s_2}, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.061	1
$P^*_{D,S}(\langle \overline{s_1}, s_2, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.116	2
$P^*_{D,S}(\langle s_1, s_2, \overline{s_3}\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.207	3
$P^*_{D,S}(\langle \overline{s_1}, \overline{s_2}, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.207	3
$P^*_{D,S}(\langle s_1, \overline{s_2}, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.344	4
$P^*_{D,S}(\langle \overline{s_1}, s_2, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.512	5
$P^*_{D,S}(\langle s_1, s_2, s_3\rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.677	6

Table 2: Knowledge Base and Comparison of Score Sums and Probabilities in case of $\mathcal{T}_{d,s}$

With these rules the P-measure $P_{S,D}$ is not uniquely determined and we need additional principles such as Maximum Entropy[11] for completing the specification.

Rules valid under the use of Maximum Entropy for $\mathcal{T}_{s,d}$ The following properties are valid if we complete the specification by the Maxent principle (and will not be mentioned in our knowledge base):

- ${}^{me}P_{S,D}(s_i) > 0.$ (Positivity)
  which is necessary to guarantee, that Eq. 13 is always defined.
- Similar to $\mathcal{T}_{d,s}$, no rule from (13) connects different symptoms. So we have again in the Maxent Model — obtained from the knowledge base — **Conditional Independence** of the symptom variables given a value for d:

$$ {}^{me}P_{S,D}(d \longrightarrow\!\!\triangleright\, \vec{s}) = \prod_{i=1}^{m} P_{S,D}(d \longrightarrow\!\!\triangleright\, s_i) \tag{14}$$

  for $s_i \in S_i$ (and analogously for $\overline{d}$).

Size of the Rule Base When using $\mathcal{T}_{s,d}$ we obtain k_i rules for each variable S_i , which means that we need to state $\sum_{i=1}^{m} k_i$ rules in case of m variables.

Example For the score system from Example 2 we obtain with translation $\mathcal{T}_{s,d}$ the Tab. 3 showing the rule base (left) and (with Maximum Entropy) the values for our possible queries $P^*_{S,D}(\vec{s} \longrightarrow\!\!\triangleright\, d)$ ((right), in comparism to the values of $w(\vec{s})$). [12]
For this example, Maximum Entropy chooses a value $P^*_{S,D}(d) \approx 0.607$, where this value was free to choose from $(0.5, \frac{2}{3})$.

[11]See [PIT] for theory and implementation of the Maximum Entropy Method.
[12]Please note that overlining is part of our succinct notation of values of binary variables. For many valued variables (more than two values) consider the translation rule (13).

Knowledge Base
${}^cP_{S,D}(s_1 \longrightarrow\!\!\triangleright\, d) = \frac{2^1}{1+2^1} = \frac{2}{3} = 0.667$
${}^cP_{S,D}(s_2 \longrightarrow\!\!\triangleright\, d) = \frac{2^2}{1+2^2} = \frac{4}{5} = 0.800$
${}^cP_{S,D}(s_3 \longrightarrow\!\!\triangleright\, d) = \frac{2^3}{1+2^3} = \frac{8}{9} = 0.889$
${}^cP_{S,D}(\overline{s_1} \longrightarrow\!\!\triangleright\, d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$
${}^cP_{S,D}(\overline{s_2} \longrightarrow\!\!\triangleright\, d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$
${}^cP_{S,D}(\overline{s_3} \longrightarrow\!\!\triangleright\, d) = \frac{2^0}{1+2^0} = \frac{1}{2} = 0.500$

Possible Queries : $P_{S,D}(\vec{s} \longrightarrow\!\!\triangleright\, d)$		$w(\vec{s})$
$P^*_{S,D}(\langle \overline{s_1}, \overline{s_2}, \overline{s_3} \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.296	0
$P^*_{S,D}(\langle s_1, \overline{s_2}, \overline{s_3} \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.456	1
$P^*_{S,D}(\langle \overline{s_1}, s_2, \overline{s_3} \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.627	2
$P^*_{S,D}(\langle s_1, s_2, \overline{s_3} \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.771	3
$P^*_{S,D}(\langle \overline{s_1}, \overline{s_2}, s_3 \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.771	3
$P^*_{S,D}(\langle s_1, \overline{s_2}, s_3 \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.871	4
$P^*_{S,D}(\langle \overline{s_1}, s_2, s_3 \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.931	5
$P^*_{S,D}(\langle s_1, s_2, s_3 \rangle \longrightarrow\!\!\triangleright\, d)$	≈ 0.964	6

Table 3: Knowledge Base and Comparison of Score Sums and Probabilities

7 Discussion and Conclusion

By presenting 3 translations of a score system into a probabilistic system, we want to emphasize the superiority of probabilistic systems, which allow for a richer representation of the available knowledge while avoiding unnecessary complexity. Especially we find the following features worth mentioning:

- **Size of the Representation:** Probabilistic systems offer a rich language to relate symptoms and diagnoses. Of course, not every problem needs a complex and large set of rules. The simplicity of score systems can be reestablished in probabilistic systems of type $\mathcal{T}_{d,s}$ or $\mathcal{T}_{s,d}$: While $\mathcal{T}_{\vec{s},d}$ uses an exponential number of rules to map the decisions of a corresponding score system, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ use a linear number. For simplicity of the rules, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ do not reflect the 'difference property' of score systems (eq. 1), but use conditional independence, to reduce the necessary number of rules.
- **Ability to add additional information**
  - Probabilistic systems allow to state **complex relations between symptoms and illnesses**. Therefore, e.g. the well-known 'exclusive or' [13] is no problem for probabilistic systems (but for Score Systems)[14]
    $P(\langle s_1, s_2 \rangle \longrightarrow\!\!\triangleright\, d) = P(\langle \overline{s_1}, \overline{s_2} \rangle \longrightarrow\!\!\triangleright\, d) = \mathit{high}$ but
    $P(\langle \overline{s_1}, s_2 \rangle \longrightarrow\!\!\triangleright\, d) = P(\langle s_1, \overline{s_2} \rangle \longrightarrow\!\!\triangleright\, d) = \mathit{low}$.
    By the way of translation, most of this ability is lost in $\mathcal{T}_{\vec{s},d}$. On the other side, $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$ allow the specification of e.g. 'contrary' knowledge.
    Example: Without creating an inconsistency the rule bases of both $\mathcal{T}_{d,s}$ and

[13] which asks for a simple, but non linear separating function on the data (in terms of the machine learning community).

[14] For just a few cases this can be treated in score systems via defining a new 'combined' symptom including the appropriate table of weights. But as these tables grow exponentially with the number of symptoms included, they hurt the aim of simplicity of score systems and do not offer a general solution. Probabilistic systems do not suffer from an exponential growth of the knowledge base in such cases.

$\mathcal{T}_{s,d}$ can be extended by the rule

$$P(\langle s_1, s_2, s_3 \rangle \longrightarrow\!\!\triangleright\ d) = 0.2;$$

- All three translations[15] offer a choice of the (a priori) **probability of the disease** d — which may be different at different locations — and the freedom to add information about the dependence of symptoms. Of course, the conditional independence, which is now present in $\mathcal{T}_{d,s}$ and $\mathcal{T}_{s,d}$, will get lost if we add relations between symptoms[16].
- **Multiple diseases:** Probabilistic systems are open to (consistently) combine knowledge about different illnesses in one system. As score systems work by adding more or less points, they are able to model within the same system different kinds of severity of a single disease, but are not able to model different diseases with partly similar, partly different symptoms. Moreover, it is not clear how different score systems might be integrated into a combined decision system.

- **Transparency:** Probabilistic systems are more transparent than score systems, as the elements of their knowledge base may be understood (and checked) in terms of relative frequencies.
- **Adaptivity:** Probabilistic systems provide a language, with is appropriate to formalize common sense knowledge about frequencies ([Gig96]). While $\mathcal{T}_{d,s}$ uses knowledge similar to 'cause $\rightarrow$ effect' relations as e.g. available in medical books, $\mathcal{T}_{s,d}$ uses 'symptom $\rightarrow$ cause' relations, as e.g. available from experienced practitioners or by mining in databases. Moreover, if carefully specified, these two kinds may both be used in one knowledge base, thus making available different kinds of knowledge from different sources.

But probabilistic systems enhanced by Maximum Entropy methods do not only deliver a theoretically well justified method to cope with incomplete and uncertain knowledge (e.g. [PV90, Sch96]). For several years they have been available in powerful implementations ([PIT, SPI]) and have demonstrated their practical benefit in real world applications (e.g. [LEX99, SER01]).
We therefore recommend the use of probabilistic systems, which may be as simple as score systems, but may also grow and become as complex as necessary for the application.

8 Acknowledgements

We want to thank Friedhelm Kulmann for providing the citation from the "Handwörterbuch der Wirtschaftswissenschaft" and Gabriele Kern-Isberner for intensive reading of earlier versions and a heap of useful hints.

[15] $\mathcal{T}_{s,d}$ in a restricted way, but remember that we can change the base or shift the scores, by which the full range for $P(d)$ can be reached.

[16] We remember, that using Maximum Entropy means to look for independencies (and indifferences) which can be derived from the available knowledge (in the sense that the amount of additional information, which leads to those models, is only minimal)

References

[FS01] B. Fronhöfer and M. Schramm. A Probability Theoretic Analysis of Score Systems. In *Proc. of the Workshop Uncertainty in Artificial Intelligence at KI, Wien*, 2001.

[Gig96] G. Gigerenzer. The psychology of good judment: Frequency formats and simple algorithms. *Medical Decision Making*, 1996.

[KR00] F. Kulmann and W. Rödder. Probabilistische Modellbildung auf der Basis von Scoring-Schemata. In *Symp. on Operations Research (SOR), Dresden*. LNCS, 2000.

[LEX99] LEXMED: Computer Aided Diagnosis of Appendicitis .
http://www.pit-systems.de/Lexmed or http://www.lexmed.de, June 1999.

[OFY+95] C. Ohmann, C. Franke, Q. Yang, M.Margulies, M.Chan, P.J. van Elk, F.T. de Dombal, and H.-D. Röher. Diagnosescore für akute Appendizitis. *Der Chirurg*, 66:135–141, 1995.

[OYF95] C. Ohmann, Q. Yang, and C. Franke. Diagnostic scores for acute appendicitis. *Eur. J. Surg.*, 161:273–281, 1995.

[PIT] Homepage of PIT. http://www.pit-systems.de.

[PV90] J.B. Paris and A. Vencovska. A Note on the Inevitability of Maximum Entropy. *International Journal of Approximate Reasoning*, 3:183–223, 1990.

[Sch96] M. Schramm. *Indifferenz, Unabhängigkeit und maximale Entropie: Eine Wahrscheinlichkeitstheoretische Semantik für nichtmonotones Schließen*. Number 4 in Dissertationen zur Informatik. CS Press, München, 1996.

[Sch00] M. Schramm. Simulation von Scoresystemen durch probabilistische Systeme. Technical report, FH Ravensburg-Weingarten, May 2000.

[SER01] Manfred Schramm, Wolfgang Ertel, and Walter Rampf. Bestimmung der Wahrscheinlichkeit einer Appendizitis mit LEXMED. *Biomedical Journal*, 57:9–11, 2001.

[SF01] M. Schramm and B. Fronhöfer. On efficient Decision Preserving Translations of Score Systems into Probabilistic systems. Technical report, PIT-Systems OHG, 2001. Available from `http://www.pit-systems.de/`.

[SG95] M. Schramm and M. Greiner. Foundations: Indifference, Independence & Maxent. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods in Science and Engeneering (Proc. of the MaxEnt'94)*. Kluwer Academic Publishers, 1995.

[SPI] Homepage of SPIRIT. http://www.xspirit.de.