PIT: A System for Reasoning with Probabilities

Manfred Schramm **Bertram Fronhöfer**
{schramm,fronhoefer}@pit-systems.de
PIT-Systems OHG
www.pit-systems.de

Abstract

PIT (Probabilistic Induction Tool) is a reasoning system based on Probability Theory which is intended for applications of decision support in cases where uncertain or incomplete knowledge plays an important role. In this paper we give an overview of the theoretical background of PIT and present the main features of its implementation. Finally we outline LEXMED (www.pit-systems.de/Lexmed), a successful PIT-based system for the diagnosis of appendicitis.

Keywords: Decision Support Systems, Probabilistic Reasoning, Maximum Entropy, Automated Diagnosis, Probabilistic Expert System Shell, Medical Knowledge Based Systems

1 Introduction

The dream of automated reasoning, stimulated by the stormy development of logic and computing machines in the 20th century was heavily impeded by the unforseen difficulties inherent in this enterprise. Classical logic, the natural basis — as well as many of the systems kindred to it — turned out to be too rigid for capturing common sense reasoning. Attempts of applying theorem provers to real world problems very often led to the disastrous situation that the rules/knowledge of an application domain, although at first asserted as certain, turned out as being not free from exceptions or as being only reliable to a certain extent.
A thorough analysis of this predicament left us in want of an inference system which allows to represent naturally degrees of certainty of knowledge (as for instance, provided by statistical investigations), to cope with partial lack of knowledge, to deal with combinations of knowledge other than the truth functional logical connectives, etc. When searching among established formal (mathematical, logical) theories for a suitable candidate, Probability Theory together with additional inference principles turned up as meeting all our demands. On the theoretical side the benchmarks of V. Lifschitz ([Lif89]) could be solved. On the practical side — after additional research in algorithms and implementation techniques — the system PIT (Probabilistic Induction Tool) was conceived and implemented. On its basis the very successful LEXMED system — for the diagnosis of appendicitis — was developed.
In this paper we give an introduction to PIT, its theory and practice. In Section 2 we present the theoretical foundations of PIT. In Section 3 we give an example of the necessity of context sensitive reasoning. In Section 4 we present several practical issues of PIT and in Section 5 we give a short outline of LEXMED.

2 Theoretical Background of PIT

Probability Theory is the theoretical starting point. Its insufficient deductive power leads to enhancing it by further inference principles, the most powerful of which is *Maximum Entropy*. Finally, restricting the basic knowledge representation language to *linear probabilistic constraints* makes inferences practically feasible, while still assuring sufficient expressiveness.

2.1 Probability Theory as Appropriate Method for Handling Uncertain and Incomplete Knowledge

The following features make Probability Theory an ideal starting point.

- **Natural Expression of Conditional Frequency Knowledge:** For most of us it is easy to express common sense knowledge by a set of 'conditional frequency expressions' (e.g. 'If I think of children showing symptom X, I expect them to suffer in many cases from illness Y') and to estimate the probability of some simple inferences without difficulties due to the natural human faculty to assess frequencies (see [Gig96]).

- **Context Sensitivity:** Probability Theory provides a language which allows to express consistently different information about a particular expression in dependence of different contexts.

  For instance, in the case of the most prominent form of specifying some information in Probability Theory, namely in the case of a conditional probability statement, we specify this information together with its context. E.g. with the probability statement $P(b|a) = x$ (or $P(a \longrightarrow\!\!\triangleright\ b) = x$ in our notation) a is the context in which the information $P(b) = x$ holds.

  We may now specify that b is highly probable in a (simple) context a (i.e. $P(a \longrightarrow\!\!\triangleright\ b)$ is close to 1) as well as in a further (simple) context c (i.e. $P(c \longrightarrow\!\!\triangleright\ b)$ is close to 1) and we may specify in addition that b is highly improbable in the context $a \wedge c$ or in the context $a \vee c$ (i.e. $P(a \wedge c \longrightarrow\!\!\triangleright\ b)$ resp. $P(a \vee c \longrightarrow\!\!\triangleright\ b)$ close to 0).

  In contrast to (classical) logic this is not considered inconsistent with Probability Theory. This 'behavior' of Probability Theory also coincides with trivial experiences in common sense reasoning: For example, the combination $a \wedge b$ of two poisons a and b may either strengthen their individual effects or cancel them (antidotes). Consequently, the poisons' combined effect cannot be forecast by a general inference rule and must be left open to specification.

- **Missing Truth Functionality:** Probability Theory also offers the same degrees of freedom on the right side of a conditional expression $a \longrightarrow\!\!\triangleright\ b$: If we have various pieces of information in the same context, we may not know how they combine and interfere, for which reason different kinds of combination of evidence must be provided.

  For instance, given the conditional probabilities $P(a \longrightarrow\!\!\triangleright\ b)$ and $P(a \longrightarrow\!\!\triangleright\ c)$, we cannot determine $P(a \longrightarrow\!\!\triangleright\ b \wedge c)$ nor $P(a \longrightarrow\!\!\triangleright\ b \vee c)$, because the properties b and c may either 'repel' or 'attract' each other in the context a. Imagine a society where an equal percentage of women (w) is married (m), has a job (j) or has children (c), i.e. $P(w \longrightarrow\!\!\triangleright\ m) = P(w \longrightarrow\!\!\triangleright\ j) = P(w \longrightarrow\!\!\triangleright\ c)$. However, $P(w \longrightarrow\!\!\triangleright\ m \wedge c)$ may be much greater than $P(w \longrightarrow\!\!\triangleright\ c \wedge j)$. In plain words this means that there may be a lot more women who are married and have children, than women who have children and a job.

- **Notion of Independence:** This notion, offered by Probability Theory, has a clear information theoretic meaning which permits qualitative inferences without having to deal with explicit numerical probability values. As we know from Bayes Nets it allows to distinguish different kinds of information flow and prepares the ground for modularizing reasoning processes.

Traditional objections to the use of Probability Theory can be overcome as follows:

- *Objection:* Probability Theory is of very limited deductive power, and the amount of inferences to be drawn from given information is so small, that Probability Theory (alone) is too weak for supporting decisions.
  Refutation: Although a true observation on Probability Theory, this objection is not relevant, because the deductive power of Probability Theory can be brought to a satisfying degree by adding cautious reasoning principles as, for instance, Maximum Entropy (see Section 2.2).

- *Objection:* Computations in Probability Theory are not practically feasible as they tend to grow exponentially with the size of the given knowledge.
  Refutation: This criticism was refuted by Lauritzen ([LS88]) who demonstrated efficient probabilistic reasoning via the use of dependence networks. (See as well Section 4.1 below on the way we compute Maximum Entropy models.)

- *Objection:* Knowledge Modeling in Probability Theory is impossible, as even the expert does not know precise probabilities in his/her field of expertise.
  Refutation: This objection is no more valid due to the introduction of probability intervals, which allow to represent uncertain probabilities. (We may specify e.g. $P(a) \in [x, y]$ which means that the probability $P(a)$ ranges somewhere between x and y.)

- *Objection:* Probability Theory is too complicated an approach for being comprehensible to the common customer.
  Refutation: We consider only finite domains which simplifies technical details of Probability Theory — at least for the user — to undergraduates' level of knowledge. This is of great convenience for the prudent customer who wants to gain some insight into the underlying principles of the systems he relies on. Moreover, as already stated above, [Gig96] has shown, that Probability Theory is well understood (by non-experts) in terms of frequency expressions.

2.2 Maximum Entropy as Appropriate Method to Complete Specifications

The main problem with knowledge specified by means of Probability Theory is its usual incompleteness. E.g. a problem involving n binary concepts (which may be true or false) yields a domain space of 2^n elements. Complete probabilistic information about a domain space of this size can only be specified by a manageable number of statements if simplifying regular patterns of information exist and are known to the expert; for instance, equal distributions of probabilities on large subdomains.
However, this is rarely the case. Consequently, the missing information must be added by means of some automatic completion procedure, because complete probability information is indispensable for supporting decisions. To achieve this aim we enhance Probability Theory by the (context sensitive) principle of **Maximum Entropy (ME)**. Using this principle means to choose (from the set of all probability models satisfying the given information) the model with maximal (information theoretic) entropy[1]. Main arguments for using ME are as follows:

[1] The entropy is defined as $H(\vec{v}) = -\sum_i v_i \cdot ln\, v_i$ with $\vec{v}$ a probability vector (see [Sha48]).

- **ME constitutes a consistent extension of Probability Theory**
  - **Model Selection:** Since ME selects one of the models provided by Probability Theory on the basis of the given knowledge it is obviously consistent.
    Of course, the choice of the ME model is context sensitive, i.e. if we change (enlarge) the given knowledge (considered as a set of constraints), we usually obtain a different ME model. (For the relation to 'non-monotonic reasoning' see [SF97]). ME also preserves the absence of truth functionality.
- **ME extends the Principles of Indifference and Independence**
  - **Extension of Indifference:** If there is no reason to consider two events as different, we don't want to introduce such a difference artificially and, consequently, we assign equal probability to the two events. This option for indifference (including a precise definition of the cases where we don't have any reason to make a difference) is respected by ME and used in PIT (see Section 4.1).
  - **Extension of Independence:** (Conditional) independences between parts of knowledge are extremely helpful for inferencing as they establish borders of context influence and thus simplify internal knowledge representation and reasoning. It is therefore of particular importance that ME does not augment the dependences among the specified knowledge, but assumes independence where no information to the contrary is available. To extend the Principle of Independence means to fulfill all independence statements given by the undirected graph (Markov network, see [Pea88]) of the constraints. (This graph can be drawn by taking the variables as nodes and connecting them according to their presence in the same constraint.) As the ME model fulfills all these independence statements, it extends the principle of independence. As these independence statements are learned (derived) from the constraints, ME does not assume independences to hold independently of the constraints (in contrast to standard methods).
- **ME selects a model characterized by *Minimal Information Increase:***
  Since maximum entropy is equivalent to minimal increase of information, the ME principle chooses the most uninformative model, or most neutral model. In other words, the given knowledge is supplemented by ME in the weakest possible way.
- **ME can be further justified by the following more theoretical considerations:**
  - **Concentration Phenomenon and the Wallis Derivation:**
    A further interesting property of the ME distribution is its strong relationship to counting models. This is explained by the so called Wallis Derivation of ME. Starting with N balls of m colors, we can count the number of urn models fulfilling a certain distribution (e.g. 3 green balls, 5 red balls,..) by using the multinomial formulae of Probability Theory. If we have to decide between two different probability distributions, we can compare the number of urn models behind them and choose the one with the greater number of models. Now the Wallis derivation says, that the ME distribution is the distribution based on most urn models ([Jay95]).
  - **Axiomatics of Information Measures:**
    As we said above, ME chooses the most uninformative model and thus avoids the introduction of any additional bias. Trying to determine axiomatically a function which measures the amount of lack of information (uncertainty) — an attempt

which goes back to Shannon's Information Theory — it can be shown that the entropy function is the only one which satisfies the specified axioms, i.e. it is the only function whose values measure 'uninformativeness' (see e.g. [Jay95, JV90]).

2.3 Advantages of Linear Probabilistic Constraints

As explained above, ME selects a special model which satisfies the specified knowledge. This means that this knowledge plays the role of a set of constraints on probability models. For practical reasons it is important to identify types of constraints which are easy to deal with computationally, and on the other hand, are also general enough to express most of the relevant expert knowledge. A good choice is a consistent set (conjunction) of **linear probabilistic (inequality) constraints** (which are linear combinations of probabilities of elementary events). They have the following nice properties:

- **Unique Decisions:** Allowing only linear probabilistic constraints assures the existence of a unique Maximum Entropy model. This is of great importance since a unique Probability Model implies that to every expression — for instance, $a \longrightarrow\!\!\triangleright\ b$ — a (unique) probability value is assigned. For instance, in a medical application, this means, that every expression consisting of an arbitrary combination of values of possible symptoms of a patient (including the case of unknown symptom values) we obtain a unique numerical value (and not just an interval) for the probability of a certain illness.

- **High Expressiveness:** On the other hand, despite their syntactically restricted form linear probabilistic constraints are still so general that they may serve as a kind of assembler language on which a very expressive High-level Probabilistic Knowledge Representation Language can be conceived (see Section 4.2).

- **Computational Efficiency:** Finally, instead of approximating the ME distribution by general methods of non-linear optimization, linear probabilistic constraints have the computational advantage that the ME distribution can be computed very efficiently.

3 Example: Context Sensitivity of Probabilistic Reasoning

As shown in the preceding section, using probabilistic specifications together with the principle of Maximum Entropy yields a powerful but fortunately still context sensitive logic. We will demonstrate this with the well-known Simpson Problem:

3.1 Preliminaries

A conditional probability[2] $P(a \longrightarrow\!\!\triangleright\ c) = x$ says that the probability of c in the context a (i.e. knowing that a is true) is x. But what happens if we change the context ?
In the analogical case of a material implication $a \longrightarrow c$ in classical (propositional) logic we can strengthen resp. weaken the context and obtain the following true inferences:

[2]Recall that $a \longrightarrow\!\!\triangleright\ b$ is an other notation for $b|a$, because we think the conditional easier to read with the arrow and switched arguments.

$$
\begin{aligned}
a \longrightarrow c &\implies a \wedge b \longrightarrow c \\
(a \longrightarrow c) \wedge (b \longrightarrow c) &\implies a \wedge b \longrightarrow c \\
(a \longrightarrow c) \wedge (b \longrightarrow c) &\implies a \vee b \longrightarrow c
\end{aligned}
$$

In case of expressions with probability 1 these inferences carry over to conditional probabilities if the contexts are not impossible (i.e. they have probability > 0).[3] We get:

$$
\begin{aligned}
P(a \longrightarrow c) = 1 &\implies P(a \wedge b \longrightarrow c) = 1 \quad \text{if} \quad P(a \wedge b) > 0 \\
P(a \longrightarrow c) = 1 \ \wedge \ P(b \longrightarrow c) = 1 &\implies P(a \wedge b \longrightarrow c) = 1 \quad \text{if} \quad P(a \wedge b) > 0 \\
P(a \longrightarrow c) = 1 \ \wedge \ P(b \longrightarrow c) = 1 &\implies P(a \vee b \longrightarrow c) = 1 \quad \text{if} \quad P(a \vee b) > 0
\end{aligned}
$$

However, this changes radically with probabilities < 1 in which case Probability Theory is too weak to derive the probabilities $P(a \wedge b \longrightarrow c)$ and $P(a \vee b \longrightarrow c)$ from the respective given assumptions. Indeed, every value $P(a \wedge b \longrightarrow c) \in (0, 1)$ is compatible with arbitrary given values $P(a \longrightarrow c) < 1$ and $P(b \longrightarrow c) < 1$, while for arbitrary given values $P(a \longrightarrow c) = 1-\varepsilon$ and $P(b \longrightarrow c) = 1 - \varepsilon$ we obtain $P(a \vee b \longrightarrow c) = 1 - \delta$ where for $\varepsilon \to 0$ also δ converges to 0. In particular, note that $P(a \longrightarrow s) > 0.5$ and $P(b \longrightarrow s) > 0.5$ does not imply $P(a \wedge b \longrightarrow s) > 0.5$ nor $P(a \vee b \longrightarrow s) > 0.5$.

3.2 Using Maximum Entropy

Using ME this probabilistic indeterminacy of expressions is overcome: ME will compute a concrete probability for each expression, which will vary (non monotonically) in dependence of additional information.
From $P(a \longrightarrow s) = 0.6$ and $P(b \longrightarrow s) = 0.6$ ME yields the conclusions

$$
\begin{aligned}
P(a \wedge b \longrightarrow s) &= 0.633 \qquad (1) \\
P(a \vee b \longrightarrow s) &= 0.589 \qquad (2)
\end{aligned}
$$

Conclusion 1 reflects the intuitive increase in certainty. (The decrease of certainty in Conclusion 2 is a necessary consequence of this).
As already mentioned the two conclusions above may change if additional information is added. This additional information may also consist of one of the two conclusions, which then influences the other one.

Example 1 [4] *Given the 3 expressions 'taking a shower' (sh), 'using a hair dryer' (b) and 'no risk for health' (nr), we specify the following information:*

- $P(sh \longrightarrow nr) = 0.99$;
  (There should be no risk in taking a shower)

[3] Of course, when translating logical formulae, which are assumed as true, as certain probabilistic statements, we get:

$$
\begin{aligned}
P(a \longrightarrow c) = 1 &\implies P(a \wedge b \longrightarrow c) = 1 \\
P(a \longrightarrow c) = 1 \ \wedge \ P(b \longrightarrow c) = 1 &\implies P(a \wedge b \longrightarrow c) = 1 \\
P(a \longrightarrow c) = 1 \ \wedge \ P(b \longrightarrow c) = 1 &\implies P(a \vee b \longrightarrow c) = 1
\end{aligned}
$$

where $a \longrightarrow b$ — in contrast to $a \longrightarrow\!\!\triangleright\, b$ — is equivalent to $\neg a \vee b$.

[4] following an idea of D. Dubois

- $P(hd \longrightarrow\!\!\triangleright\ nr) = 0.99$;
  (There should be no risk in using a hair dryer)

- $P(sh \wedge hd \longrightarrow\!\!\triangleright\ nr) = 0.05$;
  (It is not recommended to use a hair dryer and to take a shower (at the same time))

Now the probability of nr in the disjunctive context increased, as ME computes

- $P(sh \vee hd \longrightarrow\!\!\triangleright\ nr) \approx 0.99$;

Example 2 *Given the 3 expressions 'having a full-time job' (including education of children) (ft), 'working in a technical domain' (t) and 'male' (sex of the person, which works) (m), we specify the following information:*[5]

- $P(ft \longrightarrow\!\!\triangleright\ t) = 0.55$;
  (More than half of the full-time-jobs are technical)

- $P(m \longrightarrow\!\!\triangleright\ t) = 0.55$;
  (More than half of the males are working in a technical domain)

- $P(ft \vee m \longrightarrow\!\!\triangleright\ t) = 0.45$;
  (If a person has a full-time job or (not exclusive) is male ,
  less than half are working in a technical domain)

Now the probability of t in the other context, the conjunctive one, increased, as ME computes a value greater than 0.55 :

- $P(ft \wedge m \longrightarrow\!\!\triangleright\ t) \approx 0.84$;

4 From Theory to Software Systems: Coping with Real World Applications

4.1 Implementational Achievements

Since the computational effort for determining the ME model may grow exponentially with the number of concepts in a problem domain, the ME procedure must be optimized in order to meet time and resource constraints of real world applications.
For this purpose our implementation strongly exploits the properties of ME completion like Indifference and Independence to achieve exponential reductions of the computational effort (in the average case). This way the problem domain is preprocessed and subdivided into more tractable parts, thus facilitating the task of ME completion — without influencing the outcome of the ME completion — and increasing the scope of feasible problems.

- *Using the Principle of Indifference* precludes the assignment of different probabilities to those elementary events, which cannot be distinguished in view of a given set of variables and constraints. Therefore, PIT does not reason with elementary events, but with equivalence classes of elementary events, which are encoded as sets of propositional models using BDDs (Binary Decision Diagrams). This efficient handling reduces the storage space and speeds up the necessary calculations by reducing the sets of objects to be dealt with.

[5] The real syntax of PIT is slightly different, due to supporting variables with more than 2 values.

- *Using the Principle of Independence* allows (in most cases) to split the calculation of the ME model into a network of smaller distributions which are easier to compute and which are related by conditional independence. The resulting data structure is called a decomposition tree ([Tru00]).

- *Direct calculation of the Lagrange factors for expressions of our HPL.*
  In every iteration step of calculating the ME model, we have to adjust a current probability model to satisfy a currently selected constraint. Therefore, it is of great importance, how fast this update can be performed. For most expressions of our specification language (HPL) the necessary update is done in one step, for the remaining ones it is carried out by efficient search.

- Last, but not least, we have developed very *efficient data structures* and *algorithms for quick query response times.* In our medical application LEXMED query evaluation on a knowledge base of more than 500 (nested) conditional probabilities can be done in less than a second on a standard PC.

4.2 High-level Probabilistic Language (HPL)

To express one's knowledge in the form of linear probabilistic constraints is not convenient for the average expert. Therefore a **High-level Probabilistic Language (HPL)** was designed on their basis, which permits a user-friendly specification of the available information and which is automatically translated into linear probabilistic constraints.
The central feature of this HPL are the usual statements of conditional probabilities — which in addition, may range over intervals — for expressing probabilistic if-then-propositions. Accepting such an expressive language with probability intervals is an outstanding feature of PIT.
A further salient feature of PIT are so-called **relational constraints** which allow to model statements like 'In the context of symptom a, I expect to encounter disease d between 3 and 10 times more often than disease e.' More formally: $P(a \longrightarrow\!\!\triangleright\, d)/P(a \longrightarrow\!\!\triangleright\, e) \in [3, 10]$.
Our HPL also supports multivalued concepts as is seen with the following examples of HPL expressions:

- Support of multivalued (textual) variables

$$P([\text{Bloodpressure} = \text{medium}] \wedge [\text{Fever} = \text{low}]) = 0.5;$$

- Support of uncertain probabilities (interval probabilities)

$$P([\text{Bloodpressure} = \text{very high}] \longrightarrow\!\!\triangleright\, [\text{Fever} = \text{low}]) = [0.2, 0.3];$$

- Support of relational constraints

$$\frac{P([\text{Bloodpressure} = \text{high}] \longrightarrow\!\!\triangleright\, [\text{Fever} = \text{low}])}{P([\text{Bloodpressure} = \text{high}] \longrightarrow\!\!\triangleright\, [\text{Fever} = \text{high}])} = [3, 5];$$

4.3 Different Semantics of Constraints

All the HPL expressions/constraints can be used with different semantics:

- Marked Constraints:
  Marking a constraint as 'marginal' — denoted by P_m instead of simply P — means that the constraint's application will not change the distribution of its premise (as otherwise typical for ME). Example: $P_m(a \wedge b \longrightarrow\!\!\!\!\triangleright\ c) = 0.9$ will not influence the distribution of $< a, b >$. These marginal constraints allows to model marginal independence, when for instance translating Bayes nets.

- A priori Constraints:
  Usually ME starts with a uniform distribution on the domain of concepts. This can be avoided by using a priori constraints, which allow to define the start distribution.

- Uncertain Queries / Constraints:
  By default, the answer to queries assumes the context of the query to be known with certainty. By understanding a given probability model as resulting from a priori constraints, it is possible to enter vague knowledge about the query context as further constraints to which ME is applied. E.g. this way we may cope with a physician's lack of sureness about the result of a specific medical examination (say, 70% yes, 30% no).

4.4 Combining Knowledge from Different Sources

While developing our application LEXMED, it was necessary to **use and combine knowledge from different sources**, like data, human experts' knowledge and literature, because — as in most cases — one source of information was not sufficient.

- **Experts** are not able to specify hundreds of rules, including complex dependences between 3 or 4 variables, as their time is strictly limited and their dependence knowledge is not precise enough for coping with a task of this size. Moreover, different experts often contradict themselves.
  (Recall that this problem was underestimated when the first expert systems were built.)

- Available **data** alone are in most cases not representative for the actual purpose and the **literature** often does not contain enough systematic information of the necessary precision.

As our HPL (resp. some sublanguage of it) is close to the output language of several data mining tools (e.g. Bayesian Nets, with some restrictions, decision trees, some neural nets and, of course, our own cross-entropy based Data Mining Tool) and on the other side, frequency statements are quite common in the literature ([Dom91]), it is relatively easy to combine the knowledge of different sources.
Furthermore, constraint based systems offer a knowledge base which is easy to read, to understand, to check and to maintain, while not impairing efficiency. This is important, as we want to be informed about what the system knows resp. what is has learned.

4.5 Data Mining Tool

From our background in Probability Theory, Maximum Entropy (resp. the variant of Minimum Cross Entropy) and the concepts of independence, it's only a small step to generate

corresponding data mining algorithms. These algorithms — a basic concept (multi information function) can be found in [SV98] — look for correlations between different attributes and construct a dependence graph, which represents the most important dependences between the attributes. This dependence graph then serves as a basis for the determination of the (probabilistic) 'rules' which are necessary for constructing an appropriate probabilistic model (which then represents the most important dependences of the data.) For a detailed description of this algorithm see [ES99].

4.6 Cost Matrix

Basically, when asking a query, the answer provided by PIT will be a probability. But what we are eventually aiming at are decisions. How are probabilities related to decisions? Surprisingly this answer is not trivial. Different diagnoses and the respectively obtained probabilities must be compared in view of (treatment) decisions, while taking into account different types of decision errors as well as their different costs.
For instance, in our application LEXMED it makes a difference

- whether we make a surgical intervention due to the fallacious diagnosis 'perforated appendicitis' in case of a patient with 'no appendicitis', or
- whether we release a patient due to the fallacious diagnosis 'no appendicitis', in case where a 'perforated appendicitis' would have compelled a surgical intervention.

Although very unlikely to happen, the latter one is of course a much bigger mistake, or in other words, much more expensive in view of the follow-up costs. While the first erroneous decision entails the unnecessary costs of a surgical intervention, etc. the second erroneous decision is fatal for the patient with high probability. Note in addition, that apart from definitely adequate treatments (e.g. surgical intervention, releasing the patient) in case where the diagnosis is *(nearly) certain*, there are also treatment decisions (e.g. the decision to monitor the patient) which are optimal just for *uncertain* cases.
We therefore use a general method for stepping from diagnoses to decisions, so-called cost matrices (see Figure 4 for an example). A cost matrix shows the (mis)classification costs for a set of diagnoses and a set of (treatment) decisions. Multiplying this matrix with the probabilities yielded for the different diagnoses in case of an actual patient, the error costs of each decision are calculated. Of course, the best decision is the one with minimal error costs. Due to its construction process, this decision takes into account the originally provided uncertain knowledge (leading to probabilities for different diagnoses) and different types of errors (in calculating the minimum average cost, based on these probabilities). For further details see the explanation provided with Figure 4.

4.7 Online GUI

A special version of PIT[6] is available online via the home page of PIT-Systems OHG (`www.pit-systems.de`), where it's possible to play with a couple of provided examples (e.g. change their values) and to experiment specifications of one's own.

[6] Our online version is limited to 16 (binary) variables and does not use the principle of independence to speed up the calculations by separating the cliques.

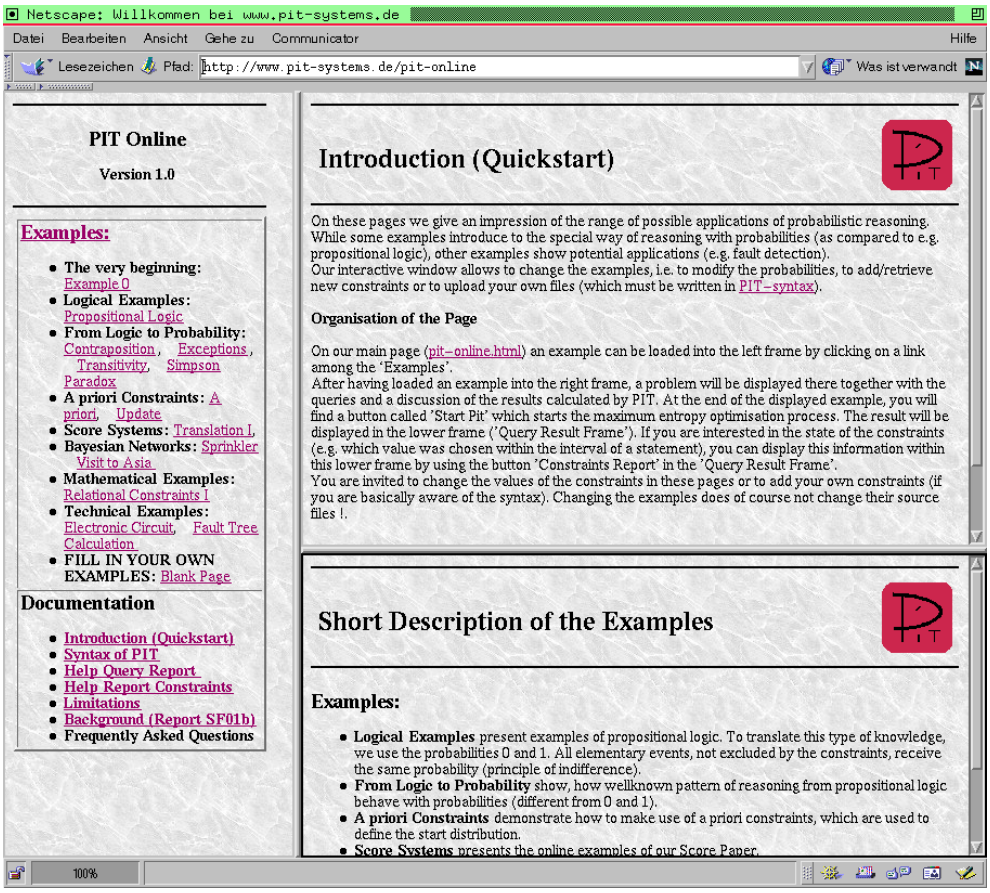

Figure 1: The Pit Online Page.

5 From Software to real world applications: LEXMED: A Medical Application of PIT

During the last twenty years the diagnosis of acute appendicitis has been improved with respect to the misclassification rate [Hon94, Dom91]. However, depending on the particular way of sampling and on the hospital, the rate of misclassification among physicians, depending on their clinical experience, still ranges between 15 and 30% which is not satisfactory ([Hon94]). A number of expert systems for this task have been developed, some with high accuracy ([Dom91]), but there is still no breakthrough in clinical use of such systems. The system LEXMED is a very promising step forward, outperforming the average physician and currently being evaluated in a hospital.

LEXMED is a **l**earning **ex**pert system for **me**dical **d**iagnosis based on the Maximum Entropy principle. Viewed as a black box, LEXMED maps a vector of clinical symptoms (discrete attribute-values) to the probabilities of different diagnoses.

Inside LEXMED (see Figure 3) apart from the PIT inference engine the central component is the Rule Base containing a set of about 500 probabilistic rules about roughly 30 binary variables.

In the LEXMED project a large database of 15000 patient records was available, which contains all the patients whose appendix has been removed in 1995 in the state Baden-Württemberg. For statistical inference on "typical" patients this database is not very useful, because it is not a representative sample of the set of patients with suspected appendicitis. Consequently, it is only possible to induce from this database a model of the patient who

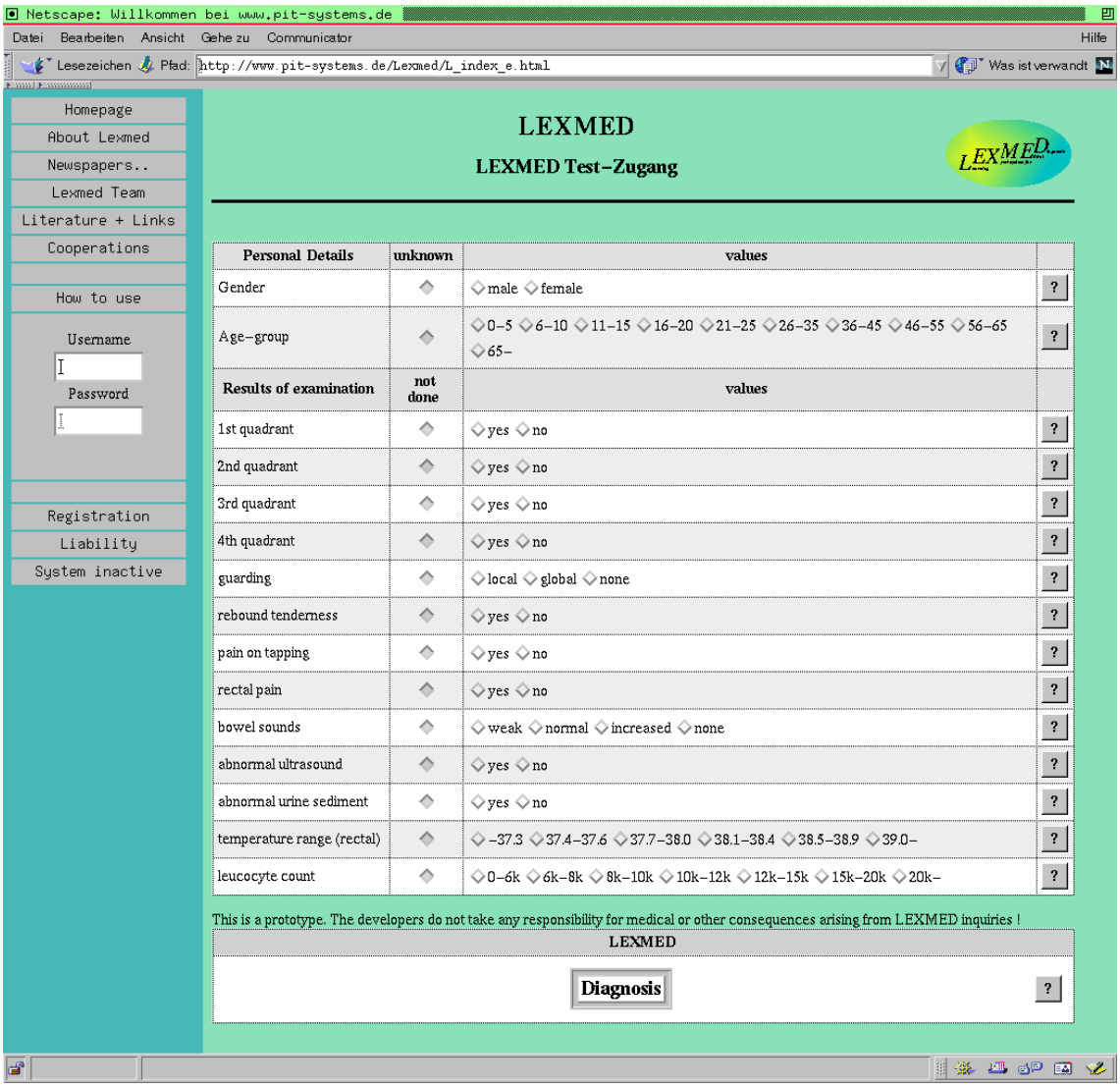

Figure 2: LEXMED User Interface

really suffers from appendicitis (assuming that nearly all cases of appendicitis are collected in the database), but it is not possible to induce a model of the 'healthy' patient, suffering only from non specific abdominal pain (NSAP). Lacking a representative database for the latter case, we decided to acquire probabilistic rules about NSAP from medical experts.
The integration of knowledge from two different sources in a joint rule base may cause severe problems, in particular if the formal knowledge representation of the two sources is different, for example if the inductive component is a neural net and the explicit knowledge is represented in first order logic. In our system, however, the language of probabilities provides a uniform and powerful knowledge representation mechanism for both expert rules and rules derived from data. And Maximum Entropy is an inference engine which does not require a complete set of rules ('complete' means that the rules are sufficient to induce a unique probability distribution by means of Probability Theory alone).
It turned out that our experts had no problems in specifying concrete probability values. Reasons may be the statistical education of physicians and the easy interpretation of conditional probabilities as relative frequencies on the respective subset. Of course in some cases we got inconsistent values from different experts which had to be discussed in subsequent interviews. If no agreement was achieved we used average values or intervals covering all

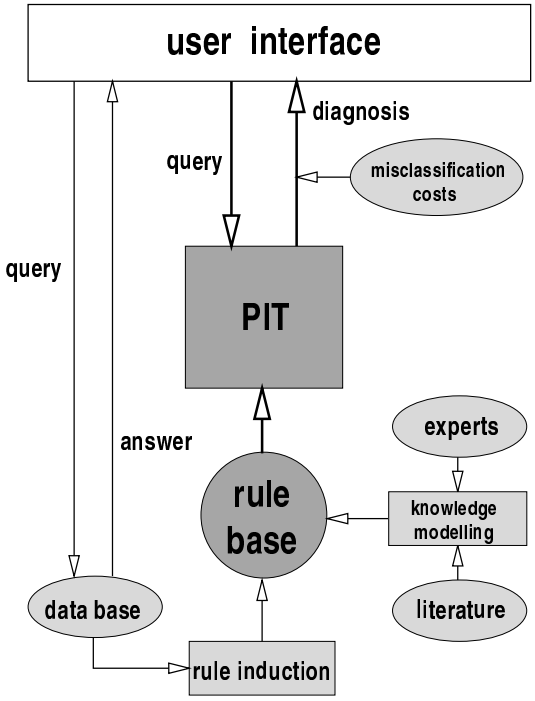

Figure 3: Overview of the LEXMED architecture.

values.
Once the rule base is constructed, a run of PIT computes the Maximum Entropy model and any query can be answered by standard probabilistic computations. However, the result of reasoning in LEXMED is not a probability of a diagnosis, but a treatment decision provided via the cost matrix ([SE99]).

5.1 Experimental Results

The running times for query evaluation in LEXMED are about 1–2 seconds. This means satisfactory efficiency.
Since June 1999 LEXMED has been running in the municipal hospital '14 Nothelfer' at Weingarten, south Germany, and in addition, has been accessible for use free of charge via internet (e.g. at `www.pit-systems.de/Lexmed`). The judgment of the physicians who work with LEXMED is very positive.
LEXMED is also continuously tested at the mentioned hospital during clinical routine, currently (March 2001) with 228 cases and an actual percentage of 88% correct decisions (sensitivity and specificity)[7].

6 Conclusion

When building systems for decision support we will very often be faced with

1. **uncertain and incomplete information** (e.g. 'most birds fly' resp. it may be unknown whether a certain object X has property Y)

[7]Sensitivity is the percentage of ill patients recognized as ill, while specificity is the percentage of healthy patients recognized as healthy.

	probability for			
	appendix inflamed	appendix perforated	negative (NSAP)	
	0.25	0.15	0.60	
therapies				average cost
operation	0	500	5800	3555
emerg.-op.	500	0	6300	3905
release	12000	150000	0	25500
monitoring	3500	7000	400	**2165**

Figure 4: The cost matrix of LEXMED together with an example probability vector `[0.25, 0.15, 0.60]`. It shows on the right the resulting average costs for the different decisions/therapies: operation, emergency operation, release, monitoring. The figures in the center include real health care costs as well as cost of estimated risk for the patient and economic loss due to illness. For example the figure 150000 represents the expected cost if a patient with perforated appendix is being released from the hospital. Due to the minimal average cost 2165 the decision is *monitoring*.

2. **heterogeneous information sources** (e.g. frequency data extractable from a database as well as rule knowledge obtainable from experts)

3. **resource constraints, e.g, time and cost** (e.g. 'a decision has to be taken within 30 minutes' resp. 'the cheapest solution has to be found')

Of course, in simple cases of decision support either straightforward procedures are available or even ad hoc solutions will do. However, increasing complexity of decision support problems entails increasing demand for more general and more powerful solutions which are justified by application independent, intelligible theoretical foundations and which also allow for efficient implementation. Therefore, the problems mentioned above have been an issue of intensive research.

PIT is a system developed to cope with these problems.

In order to cope with point (1) above we pursue the approach consisting of the combination of Probability Theory (Section 2.1) and the Principle of Maximum Entropy Completion (Section 2.2) on the basis of linear probabilistic Constraints (Section 2.3). Due to this combination our approach doesn't labor under the shortcomings of most of its competitors like decision trees, score systems, fuzzy logic, Bayesian networks, neural nets, non-monotonic logics and others.
Apart from interviewing experts we use our Entropy based Data Mining Methods (Section 4.5) to enlarge the scope of accessible knowledge (point 2).
Point (3) is taken into account by our use of Cost Matrices (Section 4.6) which translate probabilities into the financial consequences of decisions and by our efficient implementation.
For allowing user-oriented knowledge representation, on top of linear probabilistic constraints an expressive knowledge representation language (Section 4.2) was developed for the easy specification of the available information.
For increasing the scope of feasible problems, methods for anticipating some of the effects of ME Completion were developed — exploiting independence and indifference (Section 2.1) — thus allowing to modularize the completion computation. In addition, special Data structures were designed for handling large probability distributions.

Finally, the convincing performance of LEXMED promises the successful future application of PIT to other problems of medical diagnosis and to other application domains.

References

[Dom91] F.T. Dombal. *Diagnosis of Acute Abdominal Pain.* Churchill Livingstone, 1991.

[ES99] W. Ertel and M. Schramm. Combining Data and Knowledge by MaxEnt-Optimization of Probability Distributions. In *PKDD'99 (3rd European Conference on Principles and Practice of Knowledge Discovery in Databases)*, LNCS, Prague, 1999. Springer Verlag.

[Gig96] G. Gigerenzer. The Psychology of Good Judment: Frequency Formats and Simple Algorithms. *Medical Decision Making*, 1996.

[Hon94] Bernd Hontschik. *Theorie und Praxis der Appendektomie: Eine historische, psychosoziale und klinische Studie.* PhD thesis, 1994. Mabuse Verlag.

[Jay95] E.T. Jaynes. *Probability Theory: The Logic of Science.* 1995. Fragmentary edition, http://bayes.wustl.edu/etj/prob.html.

[JV90] J.B.Paris and A. Vencovská. A note on the inevitability of maximum entropy. *Int. Journal of approximate reasoning*, 4:183–223, 1990.

[Lif89] Vladimir Lifschitz. Benchmarks Problems for Formal Nonmonotonic Reasoning (V. 2.0). In Reinfrank & al, editor, *Lecture Notes in Artificial Intelligence: Non-Monotonic Reasoning*, volume 346, pages 202–219, 1989. Proc. of the workshop of non-monotonic reasoning.

[LS88] S. Lauritzen and D.J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and their application to expert Systems. In *J. Royal Statistical Society , B*, volume 50 of *B*, pages 157–224, 1988.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Kaufmann, San Mateo, CA, 1988.

[SE99] M. Schramm and W. Ertel. Reasoning with Probabilities and Maximum Entropy: The System PIT and its Application in LEXMED. In *Operations Research Proceedings 1999, ed.: Inderfurth K., Springer*, 1999.

[SF97] Manfred Schramm and Volker Fischer. Probabilistic Reasoning with Maximum Entropy — The System PIT. In *Proceedings for the twelfth workshop on logic Programming*, page 8, 1997. http://winx21.informatik.uni-wuerzburg.de/wlp97.

[Sha48] Claude E. Shannon. The Mathematical Theory of Communication. *The Bell Systems Technology Journal*, 27:349–423, 1948.

[SV98] M. Studeny and J. Vejnarova. The multiinformation function as a tool for measuring stochastic dependence. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. Kluwer Academic Publishers, 1998.

[Tru00] P. Trunk. Applying Graphical Models to Reasoning with Maximum Entropy. Diplomarbeit, TU München, February 2000.