Combining Data and Knowledge by MaxEnt-Optimization of Probability Distributions

Wolfgang Ertel and Manfred Schramm

Fachhochschule Ravensburg-Weingarten,
Postfach 1261, 88241 Weingarten, GERMANY
<ertel|schramma>@fh-weingarten.de
www.fh-weingarten.de/~ertel
http://ti-voyager.fbe.fh-weingarten.de/schramma

Abstract We present a project for probabilistic reasoning based on the concept of maximum entropy and the induction of probabilistic knowledge from data. The basic knowledge source is a database of 15000 patient records which we use to compute probabilistic rules. These rules are combined with explicit probabilistic rules from medical experts which cover cases not represented in the database. Based on this set of rules the inference engine PIT (Probability Induction Tool), which uses the well-known principle of Maximum Entropy [5], provides a unique probability model while keeping the necessary additional assumptions as minimal and clear as possible. PIT is used in the medical diagnosis project LEXMED [4] for the identification of acute appendicitis. Based on the probability distribution computed by PIT, the expert system proposes treatments with minimal average cost. First clinical performance results are very encouraging.

1 Introduction

Probabilities deliver a well-researched method of reasoning with uncertain knowledge. They form a unified language to express knowledge inductively generated from data as well as expert knowledge. To build a system for reasoning with probabilities based on data and expert knowledge, we have to solve different problems:

- To infer a set of rules (probabilistic constraints) from data, where the number of rules has to be small enough to avoid over-fitting and to be large enough to avoid 'under-fitting'. For this task we use an algorithm, which generates a probabilistic network.
- To find probabilistic rules for groups of patients, not present in our database. In cooperation with our medical experts, we collect rules describing patients with acute abdominal pain, not taken to the theatre by the diagnosis of acute appendicitis and therefore not present in our database.

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 323-328, 1999.

- To construct a unique probability model from all our constraints. If we express our knowledge in a set of rules, this set usually does not allow to generate a unique probability model, necessary to get a definite answer for every probabilistic query (of our domain). We solve this task by the use of MaxEnt (see sec. 4), which delivers a precise semantics to complete probability distributions.
- A typical problem for classification (e.g. in medicine) is that different classification errors cause different costs. We solve this task by asking the experts to define the costs of wrong decisions, where these costs are not meant to be local to the hospital, but global in the sense of including all consequences the patient has to suffer from. The decisions of our system are found by minimizing these costs under a given probability distribution.

2 Automatic Generation of Rules from Data

From a database of 15000 patient records from all hospitals in Baden-Württemberg in 1995[1], the following procedure generates a set of probabilistic rules. To facilitate the description we simplify our application to the two class problem of deciding between appendicitis ($App = true$) and not appendicitis ($App = false$), assuming the basic condition of acute abdominal pain to be true in all our rules. In order to be independent of the environment of the particular clinic, the rules are conditioned on the diagnosis variable *App*, i.e. rules will have the form

$$P(A = a_i |\ App = true, B = b_j, \ldots) = x,$$

where a_i, b_i are values of binary symptom variables A and B (binary for simplification) and x is a real number or a real interval[2]. In order to abstract the data into probabilities, we use the concept of (conditional) independence[3] as widely known and accepted. We therefore draw an independence map ([6]) of the variables, i.e. an undirected graph G where the nodes represent variables and the edges and paths represent dependencies between variables. A missing edge between 2 variables denotes a conditional independence of the two variables, given the union of all other variables (see e.g. Sec. 4.5 in [9]). For most real world applications, however the number of elementary events, induced by the union of variables, is larger than the available data (in our application, the variables span 10^9 events where 'only' 15000 patient records are available). In order to avoid over-fitting, we have to use a 'local' approach for building an independence map, i.e. an approach which works on a small set of variables rather than the union. Our procedure works as follows:

[1] We are grateful to the ARGE - Qualitätssicherung der Landesärztekammer Baden Württemberg for providing this database.

[2] In case of an interval, $P(x) = [a, b]$ expresses the uncertainty that $P(x)$ can be any value in $[a, b] \subset [0, 1]$.

[3] Variables A and B are conditionally independent given C iff (knowing the value of C) the knowledge of the value of B has no influence on deciding about the value of A. In technical terms: $P(A = a_i|B = b_j, C = c_k) = P(A = a_i|C = c_k)$ for all i, j, k.

1. For variables A and B and a vector of variables $\boldsymbol{S}$ (with a vector of values $\boldsymbol{s}$) with $A, B \notin S$) let $D_{A,S,B}$ denote the degree of dependence between the variables A and B given $\boldsymbol{S}$, which we calculate by the 'distance' [4] between $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{Y_{i,j,s}} := P(A = a|\boldsymbol{S} = \boldsymbol{s}) \cdot P(B = b|\boldsymbol{S} = \boldsymbol{s})$ and $\boldsymbol{X_{i,j,s}} := P(A = a_i, B = b_j)|\boldsymbol{S} = \boldsymbol{s})$. Let $D_{A,0,B}$ denote the degree of dependence between A and B, which we calculate by the distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $Y_{i,j} := P(A = a_i) * P(B = b_j)$ and $X_{i,j} := P(A = a_i, B = b_j)$.
2. We build an undirected graph by the following rules: Draw a node for every variable A, including the special diagnosis variable *App* (with values 'true' and 'false'). Draw an edge (A, App) iff $D_{A,0,App}$ is above a heuristically determined value t (see below). For the pair of variables A and B with the largest value of $D_{A,App,B}$ we add an edge (A,B) to the graph if for the minimal separating set $\boldsymbol{S}$ for the nodes A and B [5] the distance $D_{A,\boldsymbol{S},B}$ is above t.
3. If the procedure is completed, a graph G has been generated. As already mentioned, medical knowledge is typically conditioned on illnesses, expressing the assumption that this type of rules is more context independent than others (see footnote 8 in sec. 3). We therefore adopt the graph to this type of rules. For this goal we direct the edges (App, A) towards A and calculate rules of the form $P(A = a_i|\ App = true) = x$. Directions for the remaining edges are selected arbitrarily with the result of defining a Bayesian network[6] of rules like $P(A = a_i|App = true, B = b_j, \ldots) = x$, where App, B and possibly other variables are 'inputs' to the variable A.

   Remember that the number of edges is limited by the size of the threshold t: If the number of variables in a rule is too large in relation to the available data, t has to be increased (to avoid over-fitting); if the density of edges is too small (if the inductive power of the probabilistic rules is too weak) t has to be decreased (to avoid under-fitting).

   This set of rules is incomplete (i.e. it does not specify a unique probability model) because we do not construct rules for the class $(App = false)$ from our database (see Sec. 3). Additional rules are specified by our experts. But as the resulting set of rules is still incomplete (for e.g. using intervals in our rules), we need the method of Maximum Entropy (see 4) to complete the probability model.

[4] We use the cross entropy-function for this task, which is similar, but not equivalent to the correlation coefficient: it is defined as $CR(\boldsymbol{x}, \boldsymbol{y}) = \sum_i x_i \cdot \log(x_i/y_i)$.

[5] A separating set $\boldsymbol{S}$ for the nodes A and B disconnects A and B, i.e. there are no paths between A and B if the variables in $\boldsymbol{S}$ and their edges are removed from the graph. A minimal separating set is minimal in the number of variables it contains. If there is more than one minimal separating set $\boldsymbol{S}$, we take the set $\boldsymbol{S}$ with the lowest distance $D_{A,\boldsymbol{S},B}$. Remark: By construction, the minimal separating set will always contain *App*.

[6] The missing distribution of *App* is given by our experts

3 Expert-Rules

All patients in our database have been operated under the diagnosis 'acute appendicitis', suffering from 'acute abdominal pain'. Thus our database can not provide a model of patients with have been sent home (with the diagnosis 'non specific abdominal pain') or which have been forwarded to other departments (assuming other causes for their pain). In order to get a model of these classes of patients, we use the explicit knowledge of our medical experts[7] and the literature (see e.g. [1]) to receive rules like: [8]

$$P(A = a_i | App = false \wedge \ldots) = [x, y] \quad .$$

4 Generating a unique probability distribution from rules by the method of Maximum Entropy

In order to support interesting decisions in cases of incomplete knowledge, we have to add more constraints. In order to add no (false) ad hoc knowledge, the constraints have to be selected such that they maximize the ability to decide and minimize the probability of an error. The method of Maximum Entropy which chooses the probability model with maximal entropy H [$H(\boldsymbol{v}) := -\sum_i v_i \ * \ \log(v_i)$] is known to solve these problems:

- it maximizes the ability to decide, because it is known to choose a single (unique) probability model in the case of linear constraints.
- it minimizes the probability of an error, because the distribution of models is known to be concentrated around the MaxEnt model ([3]).

Computing the MaxEnt-Model is not a new idea but very expensive in the worst case. The main problem is that the number of interpretations (elementary events) grows exponentially with the number of variables. To avoid this effect in the average case, the principles of independence and indifference are used to reduce the complexity of the computations. These two principles are both used in our system Pit (**P**robability **I**nduction **T**ool) for a more efficient calculation of the MaxEnt model ([8]).

5 Generating Decisions from Probabilities

Once the rule base is constructed, a run of Pit computes the MaxEnt model and any query can be answered by standard probabilistic computations. However, the expected result of reasoning in Lexmed is not a probability but a decision (diagnosis). How are probabilities related to decisions? In our application

[7] We are grateful to our medical experts Dr. W. Rampf and Dr. B. Hontschik for their support in the knowledge acquisition and patience in answering our questions.

[8] This type of knowledge surely depends on the particular application scenario. For example in a pediatric clinic in Germany there are other typical causes for abdominal pain than in a hospital in a tropical country or in a military hospital.

(as in many others), misclassifications do have very different consequences. The diagnosis 'perforated appendicitis', where 'no appendicitis' would be correct, is very different to the diagnosis 'no appendicitis' where 'perforated appendicitis' would be correct. The latter case is of course a much bigger mistake, or in other words, much more expensive. Therefore we are interested in a diagnosis which causes minimum overall cost. Including such a cost calculation in the diagnosis process is very simple (c.f. Figure 1). Let C_{ij} be the additional costs if the real diagnosis is class j, but the physician would decide for i. Given a matrix C_{ij} of such misclassification costs and the probability p_i for each real diagnosis i, the query evaluation of LEXMED computes the average misclassification cost $\bar{C}_j$ to

$$\bar{C}_j = \sum_{i=1}^{n} C_{ij} \cdot p_i.$$

and then selects the class $j^* = \text{argmin}\{\bar{C}_j | j = 1, \ldots, n\}$ with minimum average cost.

6 Diagnosis of Acute Appendicitis

During the last twenty years the diagnosis of acute appendicitis has been improved with respect to the misclassification rate [2, 1]. However, depending on the particular way of sampling and the hospital the rate of misclassification among surgeons still ranges between 15 and 30%, which is not satisfactory ([2]). A number of expert systems for this task have been realized, some with high accuracy ([1]), but still there is no breakthrough in clinical applications of such systems.

LEXMED is a **le**arning **ex**pert system for **me**dical **d**iagnosis based on the MAXENT method. Viewed as a black box, LEXMED maps a vector of clinical symptoms (discrete variable-values) to the probability for different diagnoses. The central component inside LEXMED is the rule base containing a set of probabilistic rules as shown in Figure 1. The acquisition of rules is performed by the inductive part (see 2) and the acquisition of explicit knowledge (see 3). The integration of knowledge from two different sources in one rule base may cause severe problems, at least 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

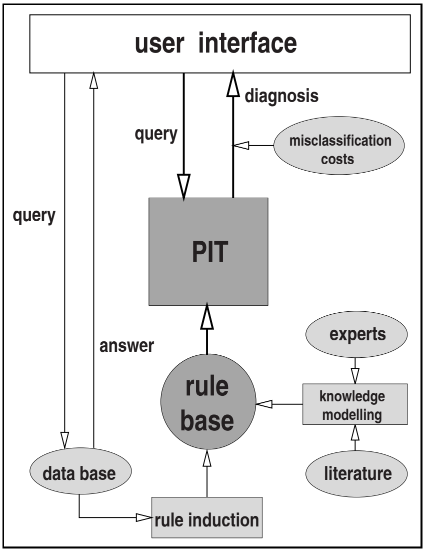

Figure1: Overview of the LEXMED architecture.

a uniform and powerful knowledge representation mechanism. And MaxEnt is an inference engine which does not require a complete[9] set of rules.

7 Results

Running times of the system for query evaluation on the appendicitis application with about 400 probabilistic rules are about 1–2 seconds. The average cost of the decisions was measured for Lexmed *without expert rules* (as described in Section 2) and for the decision tree induction system C5.0 [7] with 10-fold cross validation on the database in Table 1. For completeness reasons we also performed runs of both systems without cost information and computed the classification error.

	Lexmed	C5.0
Average cost	1196 DM	1292 DM
Classification error	22.6 %	23.5%

Table1. Average cost and error of C5.0 and Lexmed without expert rules.

The figures show that on the database the purely inductive part of Lexmed and C5.0 have similar performance. However in a real test in the hospital the expert rules in addition to the inductively generated rules will be very important for good performance, because the database is not representative of the patients in the hospital as mentioned in Section 3. Thus, in the real application we expect Lexmed to perform much better than a pure inductive learner like C5.0.

Apart from the numeric performance results the first presentation of the system in the municipal hospital of Weingarten was very encouraging. Since June 1999 the doctors in the hospital use Lexmed via internet [4].

References

1. De Dombal. *Diagnosis of Acute Abdominal Pain.* Churchill Livingstone, 1991.
2. B. Hontschik. *Theorie und Praxis der Appendektomie.* Mabuse Verlag, 1994.
3. E.T. Jaynes. Concentration of distributions at entropy maxima. In Rosenkrantz, editor, *Papers on Probability, Statistics and statistical Physics.* D. Reidel Publishing Company, 1982.
4. Homepage of Lexmed. http://lexmed.fh-weingarten.de, 1999.
5. J.B. Paris and A. Vencovska. A Note on the Inevitability of Maximum Entropy. *International Journal of Approximate Reasoning*, 3:183–223, 1990.
6. J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.
7. J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993. C5.0, online available at www.rulequest.com.
8. M. Schramm and W. Ertel. Reasoning with Probabilities and Maximum Entropy: The System PIT and its Application in LEXMED. In *accepted at: Symposium on Operations Research 1999*, 1999.
9. J. Whittaker. *Graphical Models in applied multivariate Statistics.* John Wiley, 1990.

[9] 'complete' means that the rules are sufficient to induce a unique probability distribution.