Proc. 22nd Int. Conf. on Machine Learning, Bonn, Germany, 2005

Dirichlet Enhanced Relational Learning

Zhao Xu ZHAO.XU@CAMPUS.LMU.DE
Institute for Computer Science, University of Munich, Germany

Volker Tresp VOLKER.TRESP@SIEMENS.COM
Kai Yu KAI.YU@SIEMENS.COM
Corporate Technology, Siemens AG, Information and Communications, Munich, Germany

Shipeng Yu SPYU@DBS.INFORMATIK.UNI-MUENCHEN.DE
Hans-Peter Kriegel KRIEGEL@DBS.INFORMATIK.UNI-MUENCHEN.DE
Institute for Computer Science, University of Munich, Germany

Abstract

We apply nonparametric hierarchical Bayesian modelling to relational learning. In a hierarchical Bayesian approach, model parameters can be "personalized", i.e., owned by entities or relationships, and are coupled via a common prior distribution. Flexibility is added in a *nonparametric* hierarchical Bayesian approach, such that the learned knowledge can be truthfully represented. We apply our approach to a medical domain where we form a nonparametric hierarchical Bayesian model for relations involving hospitals, patients, procedures and diagnosis. The experiments show that the additional flexibility in a nonparametric hierarchical Bayes approach results in a more accurate model of the dependencies between procedures and diagnosis and gives significantly improved estimates of the probabilities of future procedures.

1. Introduction

Relational modelling has recently received increasing attention (Dzeroski & Lavrac, 2001; Raedt & Kersting, 2003) and plays an important role in modern data mining (Wrobel, 2001). The reason is that relevant information is not only contained in attributes describing properties of objects but also in relationships between objects. Two of the leading frameworks, i.e. the *probabilistic relational models* (PRM) framework (Friedman et al., 1999) and the *directed acyclic probabilistic entity relationship* DAPER framework (Heckerman et al., 2004), describe relational modelling in the context of relational data bases. They are motivated from different database structure representations: the PRM model is based on the *relational model* and the DAPER model is based on the the *entity-relationship model*(Ullman & Widom, 1997). The DAPER framework — the focus of this paper— is particularly elegant in a Bayesian context since it encourages an explicit representation of parameters and hyperparameters. A Bayesian approach is well suited for relational modelling. The reason is that parameters, instead of being global, can be personalized to entities and relationships leading to a hierarchical Bayesian (HB) framework (Gelman et al., 2003).

In an HB approach, the parameterization of the prior distribution obtains central importance since it must be able not only to represent ones prior belief but also must be flexible enough to represent the learned prior, which might not be in the same family of distributions. Thus it is advantageous to specify the prior distribution in a flexible nonparametric form, technically as a sample from a Dirichlet process (DP). Although we can still implement our vague prior belief in form of the base distribution of the DP, the learned prior can be very rich. Due to the central importance of the Dirichlet process, the re-parameterization of a prior distribution in form of a nonparametric highly flexible representation is sometimes referred to as Dirichlet enhancement (Escobar & West, 1998), thus we name the proposed framework "Dirichlet Enhanced Relational Learning" (DERL).

We apply our framework in the context of a medical

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

data base. The entities in the model are hospitals, patients, diagnosis and procedures. The existence of a diagnosis or procedure is dependent on patient features and hospital features and is modelled as reference uncertainty, which is a mechanism to represent the uncertainty in the relational structure itself (Getoor et al., 2003). The prior distributions for the multinomial parameters describing the reference uncertainties are now Dirichlet enhanced and are learned via nonparametric HB. In the learned nonparametric prior distribution, parameters for diagnosis and procedures are dependent allowing for inference from diagnosis to procedures and vice versa. We are investigating the task of predicting additional procedures and diagnosis based on hospital and patient attributes, the prime complaint and on previously administered procedures and diagnosis, thus emulating the process of a clinical workflow.

The paper is organized as follows. In the next section we will review Bayesian and HB modelling. In Section 3, we discuss relational modelling in the form of the DAPER model. In Section 4 we will introduce HB modelling in the context of relational data, and in Section 5 nonparametric HB modelling and Dirichlet enhancement. In Section 6 we will introduce efficient approximate learning and inference algorithms. In Section 7 we will describe the experiments and experimental results using the data from a medical data base. Section 8 contains our conclusions.

2. Bayesian and Hierarchical Bayesian modelling

A Bayesian model suitable for our discussion would consist of parameter vector θ with a prior probability distribution $P(\theta|h)$ that contains hyperparameters h with probability distribution $P(h)$. The training data D are generated following a likelihood model $P(D|\theta)$. In a Bayesian setting, the functional forms of all distributions must be specified *a priori*. Prior to the arrival of data, parameters are distributed as

$$P(\theta) = \int P(\theta|h)P(h)dh$$

and with known training data D one obtains using Bayes' rule

$$P(\theta|D) = \frac{\int P(D|\theta)P(\theta|h)P(h)dh}{\int P(D|\theta)P(\theta|h)P(h)dhd\theta}.$$

The posterior parameter distribution $P(\theta|D)$ now assumes the role of the new "learned prior", i.e., the available knowledge prior to the arrival of additional data. Note, that with an increasing size of the data set, the posterior distribution of θ becomes increasingly localized and converges eventually to a point distribution.

Although any Bayesian approach is essentially hierarchical, one speaks of a hierarchical approach in the narrow sense if data $\{D_i\}_{i=1}^M$ for *related but not identical* scenarios need to be modelled, e.g., outcome data from different hospitals. A reasonable assumption is that the data in each scenario D_i are generated with a specific parameter vector θ_i, but that the θ_i are generated from a common prior distribution leading to the joint model

$$P(h)\prod_{i=1}^{M} P(\theta_i|h)P(D_i|\theta_i).$$

After having observed data from M scenarios, the parameter distribution for a new scenario $M+1$ becomes

$$P(\theta_{M+1}|\{D_i\}_{i=1}^M) \propto \int P(h)P(\theta_{M+1}|h)\times \prod_{i=1}^{M} P(\theta_i|h)P(D_i|\theta_i)dhd\theta_1 \dots d\theta_M.$$

With $M \to \infty$ this distribution will converge to the actual *distribution of the parameters* which we will denote informally as the learned prior distribution in the paper.

In hierarchical Bayesian modelling, the prior distribution must be able to represent the actual distribution of parameters and not only one's vague prior belief. Thus it is very important that the prior distribution is very flexible. This can be achieved by a nonparametric representation of the prior distribution as discussed in Section 5.

3. The DAPER Model

A typical domain of interest might consist of objects (entities), their attributes and their relationships. Most machine learning approaches have tried to select a representation in which a relational representation could be avoided by constructing appropriate derived features (propositionalization). Thus, the full information contained, for example, in a relational data base, could not be represented and exploited. Over past years, a number of approaches have been developed, which consider relational information in a principled way. In this paper we assume the representation developed around the DAPER framework (Heckerman et al., 2004).

The DAPER model formulates a probabilistic framework for an entity relationship database model (a commonly used representation for the structure of a database). DAPER framework makes relationships

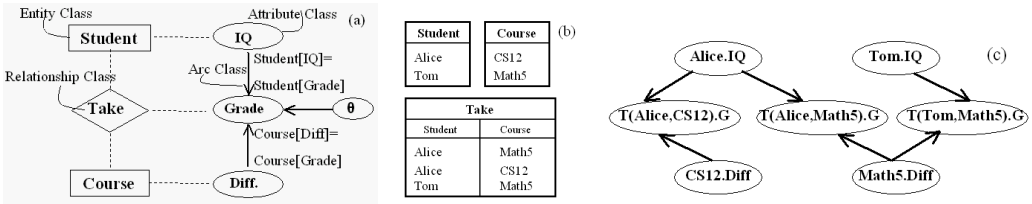

Figure 1. An example over university domain from Heckerman et al. (2004) (a) DAPER model (b) Instantiated entities and relationships.(c) Ground Bayesian network.

first class objects in the modelling language, and encourages an explicit representation of conditional probabilistic distribution. The DAPER model consists of entity classes, relationship classes, attribute classes and arc classes, as well as local distribution classes and constraint classes. Figure 1(a) shows an example of a DAPER model for a universe of students, courses and grades. The entity classes specify classes of objects in the real world, e.g. Student and Course shown as rectangles in Figure 1(a). The relationship class represents interaction among entity classes. It is shown as a diamond-shaped node with dashed lines linked to the related entity classes. For example, the relationship, Take(s, c) indicates that a student s takes a class c. Note, that the DAPER model assigns relationships the same importance as the entities. Attribute classes describe properties of entities or relationships. Attribute classes are connected to the corresponding entity/relationship class by a dashed line. For example, associated with courses is the attribute class Course.Difficulty and associated with the relationship class Take is the attribute class Take.Grade. *The attribute class* θ in Figure 1(a) represents the parameters specifying the probability of student's grade in different configurations (i.e. course's difficulty and student's IQ). It is denoted a global attribute class since it is not associated with an entity or relationship. The arc classes shown as solid arrows from "parent" to "child" represent probabilistic dependencies among corresponding attributes. For example, the solid arrow from $Student.IQ$ to $Course.Grade$ specifies the fact that student's grade probabilistically depends on student's IQ. A local distribution class for an attribute class is a specification from which local distributions for attributes corresponding to the attribute class can be constructed. As an example, the probabilistic distribution of Take.Grade given its parents is specified by a local distribution class (not shown) based on the global parameter vector θ.

Based on the DAPER model (e.g. Figure 1(a)) and the instantiated entities and relationships (e.g.Figure 1(b)), a ground Bayesian network (e.g.Figure 1(c)) can be generated in which probabilistic inference (e.g., belief propagation) can be performed. Constraint classes specify how to derive ground Bayesian network from the corresponding DAPER model over the given instantiated domain: For example, the constraint $course[Diff] = course[Grade]$ indicates that in the ground network an arc should be introduced between attribute $c.Diff$ and attribute Takes(s, c').Grade, only when $c = c'$. So it is forbidden to add a solid arrow from $CS12.Diff$ to $Take(Tom, Math5).Grade$.

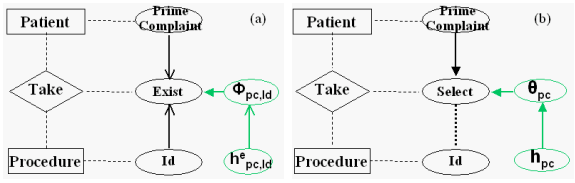

Figure 2. DAPER models with structure uncertainty over medical domain, which is modelled using the formalism of Getoor et al. (2003). (a)existence uncertainty modelling. (b)reference uncertainty modelling. The attribute Take.Select is modelled as a multinomial variable with as many states as there are procedures. $\theta_{s|pc}$ is the parameter vector for the multinomial distribution and h_{pc} are the parameters of the prior distribution.

In some real-world applications, the relational structure itself, is uncertain. Getoor et al. (2003) proposed two mechanisms to represent this type of uncertainty: one is existence uncertainty, the other is reference uncertainty. In existence uncertainty, a relationship class has a particular binary attribute Exist such that the probability for the existence of a relationship can be modelled. In particular, the attribute Exist has two states, Yes/No, and can be modelled as binomial. The Figure 2 gives an example in a medical domain. Patient.PrimeComplaint is an attribute describing the prime complaint of the patient. Proce-

dure.Id specifies the identifier of the procedure. The relationship Take(pa, pr) models the fact that a patient pa receives a procedure pr. In Figure 2(a), the uncertainty of which procedure is taken by a patient is modelled as existence uncertainty. The global attribute $\phi_{e|pc,Id}$ represents the parameters of distribution of Exist given prime complaint pc and procedure Id. $h^e_{pc,Id}$ are hyperparameters of $\phi_{e|pc,Id}$.

Reference uncertainty is used in situations where one part of a relationship might be certain, but there is uncertainty about the other part of the relationship. Illustratively, the relational link of the patient taking a procedure is known say "John" with a specific prime complaint, but the other side, i.e. the procedure, is unknown. In reference uncertainty, a relationship class is associated with an additional attribute Select (see Figure 2(b)) with as many states as there are possible procedures. The attribute Select is modelled as multinomial variable. The global attribute $\theta_{s|pc}$ are the parameters of the distribution of Select given prime complaint pc and hyperparameters h_{pc}. In the paper we focus on reference uncertainty to model the structural uncertainty.

4. Hierarchical Bayes for Relational Models

In the above medical domain example, parameters and hyperparameters specifying conditional distributions are explicitly modelled as global attributes. This has two important implications. First, the probability for a procedures is identical for all patients with the same prime complaint. Secondly, procedures are modelled as independent such that knowledge about a prescribed procedure does not influence the selection of subsequent procedures. Both implications are not realistic. Patients are truly unique which might be obvious to the attending physician but which is impossible to represent in a probabilistic model, which is always a simplification. Thus, given a prime complaint, a physician might select a personalized treatment strategy. Additionally, the procedures taken by a patient are related. The prescribed procedures influence the later procedures, the physician often make decision of the next procedure based on the previous ones. A principled approach to solve these problems is hierarchical Bayes (discussed in Section 2) where it is assumed that each patient should be an individual requiring individual procedure probabilities. In Figure 3(a) the probability of selecting a procedure is an attribute of a patient pa reflected by the attribute $\theta_{s|pc,pa}$. Naturally, we will almost never have sufficient data to estimate the individual patient parameters; this dilemma is solved by assuming that all patient models originate from a common prior distribution which can be learned and shared between patients: the hyperparameters are still modelled as global parameters, since it is still shared by all patients, not individual for each patient. Thus a common prior distribution can be learned and a new patient inherits an informed prior distribution biasing the model in a sensible manner. To fix the

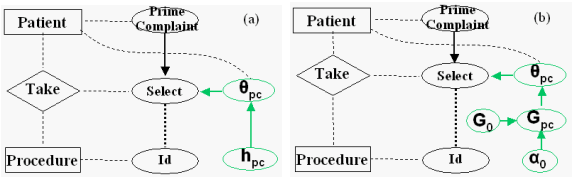

Figure 3. (a)Same as in Figure 2(b), except that the multinomial parameter $\theta_{s|pc,pa}$ is owned by the patient. This corresponds to a hierarchical Bayesian model. (b)A nonparametric hierarchical Bayesian model. The prior distribution G_{pc} is a sample from a Dirichlet process.

HB model over the medical domain, let's assume that the procedure probability is a multinomial distribution with Dirichlet prior. In particular, for each patient, an individual parameter vector $\theta_{s|pc,pa}$ is assumed which specifies the probability of procedure for the patient pa with prime complaint pc. The parameter vector $\theta_{s|pc,pa}$ is generated from a Dirichlet distribution with parameters $h_{pc} = \{\tau_{pc}, \alpha_{pc}\}$, which can be written as:

$$\mathrm{Dir}(\theta_{.|pc,pa}|\tau_{pc}, \alpha_{pc}) = \frac{1}{C}\prod_{k=1}^{K} \theta_{k,s|pc,pa}^{\tau_{pc}\alpha_{k,pc}-1}. \qquad (1)$$

where C is a normalization constant given by integration over all possible $\theta_{.|pc,pa}$, K is the total number of procedures, $\alpha_{pc} = \{\alpha_{1,pc}, \ldots, \alpha_{K,pc}\}$, $\alpha_{i,pc} \geq 0$, $\sum_i \alpha_{i,pc} = 1$, and $\tau_{pc} \geq 0$ named confidence parameters.

In hierarchical Bayes each patient obtains personalized procedure probabilities and shares a common parametric prior such that the two unrealistic assumptions are released. In more cases than not, the real prior distribution will not fall into the class of belief distributions that can be described by $P(.|h)$ for any h. One solution to these problems is to assume that the prior distribution assumes a very flexible nonparametric form which leads to the framework of nonparametric Bayesian modelling.

5. Nonparametric Hierarchical Bayes and Dirichlet Enhancement

Of central importance in nonparametric Bayesian modelling is the Dirichlet process, which can be thought of as an infinite-dimensional generalization of a Dirichlet distribution. In particular, one assumes that the prior parameter distribution is a sample from a Dirichlet process (Figure 3(b)) and writes (Escobar

& West, 1998):

$$G_{pc} \sim \mathrm{DP}(G_0, \alpha_0)$$

where G_0 is the base distribution, by which we can implement our vague prior belief. $\alpha_0 \geq 0$ is the concentration parameter specifying the degree of certainty in our prior belief. The nice feature of this approach is that, although we can still implement our vague prior belief in form of the parameters of the DP, i.e. G_0 and α_0, the prior G_{pc} can be very rich. The multinomial parameter $\theta_{s|pc,pa}$ is simply samples from G_{pc}

$$\theta_{\cdot|pc,pa} \sim G_{pc}.$$

We can explicitly write G_{pc} in the stick breaking representation (for a definition consult Teh et al. (2004)):

$$G_{pc} = \sum_{l=1}^{\infty} \pi_{l,pc} \delta_{\theta^*_{l,pc}}; \quad \theta^*_{l,pc} \sim G_0 \tag{2}$$

$$\pi'_{l,pc} \sim Beta(1, \alpha_0); \quad \pi_{l,pc} = \pi'_{l,pc} \prod_{k=1}^{l-1} (1 - \pi'_{k,pc}) \tag{3}$$

where $\theta^*_{l,pc}$ are samples independently and randomly selected from the base distribution G_0, $\delta_{\theta^*_{l,pc}}$ is a distribution concentrated at a single point $\theta^*_{l,pc}$. $\pi_{l,pc}$ are positive weights which sum to one. $\pi_{l,pc}$ only depend on the concentration parameter α_0 and are generated using Equation 3. For more information on Dirichlet processes, please consult Teh et al. (2004) or Tresp and Yu (2004). Note that despite the continuous nature of the base distribution, a sample from a Dirichlet process, e.g., G_{pc}, is discrete in nature.

6. Approximate Inference and Learning

Traditionally, learning in nonparametric Bayesian modelling is performed via Gibbs sampling. The most common variations are the Polya urn or Chinese restaurant sampling approach (Teh et al., 2004; Tresp & Yu, 2004). These approaches are computationally quite involved; thus in our paper we focus on a computationally efficient approach described in Yu et al. (2004).

The goal is to estimate G_{pc} for each possible pc in the data base, using the marginal likelihood:

$$\hat{G}_{pc} = \arg\max_{G} \ \mathrm{DP}(G|G_0, \alpha_0) \times \prod_{\{pa\}_{pc}} \int \mathrm{Mul}(\{pr\}_{pa}|\theta_{\cdot|pc,pa}) G(\theta_{\cdot|pc,pa}) d\theta_{\cdot|pc,pa}$$

where $\{pa\}_{pc}$ is the set of patients with the same prime complaint pc, and $\{pr\}_{pa}$ is the set of procedures of patient pa. Unfortunately, calculating the marginal likelihood is intractable and we rely on a mean field approximation. The mean field approximation is motivated by the stick breaking representation of Equation 2, which can be written as:

$$G_{pc} \approx \sum_{pa=1}^{n} \pi_{pa,pc} \delta_{\theta^*_{\cdot|pa,pc}}$$

where $n = |\{pa\}_{pc}|$, is the number of patients with the prime complaint pc. The learning process is divided into two steps: 1) to calculate the location $\theta^*_{\cdot|pa,pc}$ of the concentrated distribution, 2) to estimate the weight $\pi_{pa,pc}$. The location of the discrete term is approximated by the maximum a posteriori (MAP) estimates of $\theta_{\cdot|pc,pa}$ defined as:

$$\theta^{MAP}_{\cdot|pc,pa} = \arg\max P(\theta_{\cdot|pc,pa}|\{pr\}_{pa})$$

where the Dirichlet distribution of Equation 1 is used as prior.

In the second step, to estimate the weight $\pi_{pa,pc}$, the assumption is made that:

$$\hat{P}(\theta_{\cdot|pc,pa}|\{pr\}_{pa}) \approx q_{pa}(\theta_{\cdot|pc,pa}) = \sum_{\{\tilde{pa}\}_{pc}} \xi_{\tilde{pa},pc,pa} \delta_{\theta^{MAP}_{\cdot|pc,pa}}$$

where $\xi_{\tilde{pa},pc,pa}$ are the variational parameters with $\xi_{\tilde{pa},pc,pa} \geq 0$ and $\sum_{pa} \xi_{\tilde{pa},pc,pa} = 1$.

We obtain as variational E-step, for $t = 1, 2, \ldots$:

$$\xi^t_{\tilde{pa},pc,pa} = \frac{P(\{pr\}_{\tilde{pa}}|\delta^{MAP}_{\theta_{\cdot|pc,pa}}) \ \hat{G}^{(t)}_{pc}(\delta_{\theta^{MAP}_{\cdot|pc,pa}})}{\sum_{\tilde{pa}} P(\{pr\}_{\tilde{pa}}|\delta^{MAP}_{\theta_{\cdot|pc,pa}}) \ \hat{G}^{(t)}_{pc}(\theta^{MAP}_{\cdot|pc,pa})} \tag{4}$$

and with $\xi_{pc,pa} = \sum_{\tilde{pa}} \xi_{\tilde{pa},pc,pa}$, the M-step

$$\hat{G}^{(t+1)}_{pc}(\theta_{\cdot,pc,pa}) = \frac{\alpha_0 G_0(\theta_{\cdot,pc,pa}) + \sum_{\{pa\}_{pc}} \xi_{pc,pa} \delta_{\theta^{MAP}_{\cdot|pc,pa}}}{\alpha_0 + |\{pa\}_{pc}|}. \tag{5}$$

After convergence, the "learned" prior assumes the form of Equation 5. With $\alpha_0 \to \infty$ the learned prior corresponds to the uninformed prior. With a finite α_0 we obtain a nonparametric hierarchical Bayes solution.

7. Experiments

We apply our model in the context of a medical data base. Its entity-relational model shown in Figure 4(a). The domain includes four entity classes (hospitals, patients, diagnoses and procedures) and three relationship classes (In:patient being in a hospital, Make:patient making a diagnosis and Take:patient taking a procedure). A patient pa is in exactly one hospital ho and typically has both multiple procedures

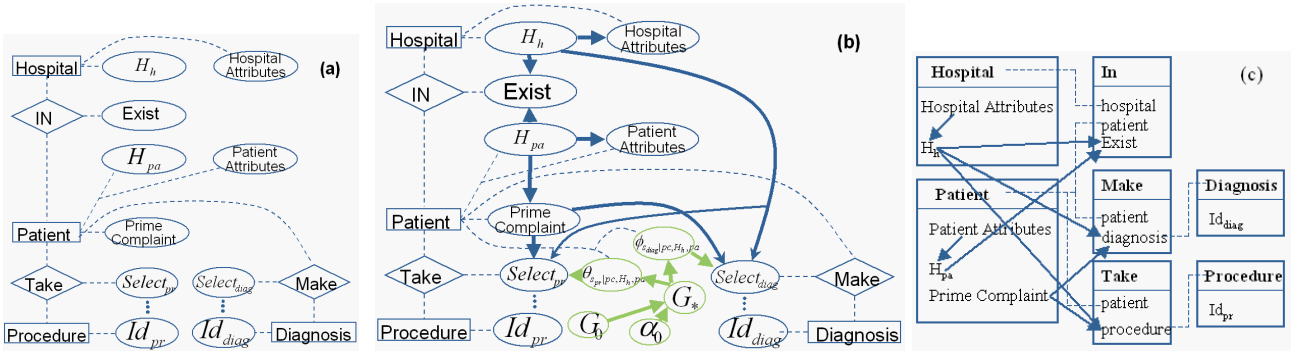

Figure 4. (a) Medical data base structure represented by entity relational model (b) DERL model (c) PRM model

pr and diagnoses dia. Hospital class has attribute classes such as hospital bedsize, teaching status (teaching/nonteaching), hospital location (urban/rural), etc. Patient class has attribute classes including gender, age, admission source, etc. To reduce complexity of the Figure 4(a), hospital and patient characteristics are grouped together as HospitalAttributes and PatientAttributes respectively (these attributes are not aggregated in learning and inference). In addition, patient class has the attribute class PrimeComplaint, which states the prime complaint of the patient at the time of admission. For both the hospital characteristics and patient characteristics we learned multinomial mixture models using hidden mixture attributes Hospital.H_h and Patient.H_{pa}. Both the relationship between patients and procedures and the relationships between patients and diagnoses are modelled as reference uncertainty. Thus the two relationships have additional attributes $Select_{pr}$ and $Select_{dia}$, respectively. The states of $Select_{pr}$ and $Select_{dia}$ indicate which procedure, resp. diagnosis is given by the physician. We used data from 9980 patients for training and 4082 patients for testing. In the data, there were 703 different diagnoses and 367 different procedures. The system was optimized to have 60 patient clusters and 3 hospital clusters.

The DERL model is shown in Figure 4(b). The parameters of the multinomial variables $Select_{pr}$ and $Select_{dia}$ are $\theta_{s_{pr}|pc,H_h,pa}$ and $\phi_{s_{dia}|pc,H_h,pa}$, which are individual for each patient. The two parameters share a common prior G_{pc,H_h}, which is a sample from a Dirichlet process. Note that the base distribution G_0 of the Dirichlet process is a product of two independent Dirichlet distributions:

$$G_0 = \mathrm{Dir}(\theta_{.|pc,H_h,pa}|\tau_{pr},\alpha_{pr}) \times \mathrm{Dir}(\phi_{.|pc,H_h,pa}|\tau_{dia},\alpha_{dia})$$

$$\alpha_{pr} = \frac{1}{N_{pr}}(1,1,1,\ldots)^T; \alpha_{dia} = \frac{1}{N_{dia}}(1,1,1,\ldots)^T$$

where N_{pr} and N_{dia} are the number of procedures and diagnosis, respectively (i.e. 367 and 703 in the case). The base distribution implies unbiased priors. τ_{pr}, τ_{dia} are confidence parameters and were optimized via cross-validation. Our model implies that a priori, procedures and diagnosis are modelled as being independent[1]. A posterior the Dirichlet enhanced model is able to represent *dependencies* between procedures and diagnosis. We have to realize separate Dirichlet processes for each configuration of the states of the parents of the select variables. This immediately brings up the issue of overfitting, since for any particular combination of the states of prime complaint and H_h, there might be only few or no data in the training data set. For example, if consider the situation where there is no patient with the prime complaint *circulatory* in the hospitals clustered 2 in training data, then we have $G_{circulatory,2} = 0$. That means the probability of any procedure of a new patient with that configuration is always zero, which is obviously incorrect. The typical approach to dealing with this problem is to smooth the probability, assigning positive value, no matter whether the configuration occurs in the training data. Thus we employ Linear-interpolation-smoothing from language modelling (Jelinek, 1997). For the procedure Select s_{pr} we obtain

$$P(s_{pr}|H_h,pc) = \lambda_0 P(s_{pr}) + \lambda_1 P(s_{pr}|H_h) +$$

$$\lambda_2 P(s_{pr}|pc) + \lambda_3 P(s_{pr}|H_h,pc)$$

and a corresponding expression for diagnosis Select s_{diag}. The weights λ_i are estimated using an EM algorithm. The (conditional) probabilities $P(s_{pr})$, $P(s_{pr}|H_h)$, $P(s_{pr}|pc)$, and $P(s_{pr}|H_h,pc)$ are all modelled as separate Dirichlet enhancement models. LM-smoothing can be implemented in the DERL model

[1]Model selection showed that we obtain a better predictive model by using prime complaint as a parent and not H_{pa}

with an additional hidden variable. Since this would make the model less readable, we did not draw this variable in Figure 4(b).

We compare our model with standard PRM (e.g. Friedman et al. (1999)), which is shown in Figure 4(c). The difference from our model is that the multinomial distributions of selection of procedures (and diagnoses) are global, not individual for each patient.

We test model performances by predicting the application of procedures. In the first experiment we predicted any of the procedures that a patient has received given hospital properties, patient properties and given prime complaint. The corresponding ROC curve (averaged over all patients) for DERL model is shown as E2 in Figure 5. In the experiment we selected the top N procedures recommended by the model. Sensitivity indicates how many percent of the actually being performed procedures were correctly proposed by the model. (1-specificity) indicates how many of the procedures that were not actually performed were recommended by the model. Along the curves, the N was varied from left to right as $N = 5, 10, \ldots, 50$. $E1$ shows the experimental result of the standard PRM model (Figure 4(c)) given the same information as E2. It is essentially identical to the result of E2. The situation changes when additional information is available such as past procedures or diagnosis: the standard PRM model would not change the proposal probabilities. In the DERL model, in contrast, the prediction of a subsequent procedure is improved if the first diagnosis is available (E3) or both the the first diagnosis and the first procedure are available (E4). We can see, for example, that if we would propose 15 procedures, after we know the prime complaint, the first diagnosis, and the first procedure, we would cover approximately 83% of the actually prescribed procedures. Figures 6 shows the corresponding plots for patients with prime complaint *respiratory problem* exhibiting similar trends.

In the second set of experiments we investigated how the procedure probabilities sequentially change when information becomes available. Figure 7 shows the selection probabilities for 20 procedures which are relevant for myocardial infarction. The top ten procedures are listed in Table 1. The first column indicates the predicted probabilities for the case that only patient property and hospital property are available. The second column shows the procedure probabilities when, in addition, the prime complaint *circulatory problem* becomes available. The third column shows the situation when, in addition, the first diagnosis *acute myocardial infarction* becomes available. The fourth column shows the situation when, in addition, the procedure *single vessel percutaneous transluminal coronary angioplasty* has been performed. One sees that the procedure probabilities for procedures relevant for myocardial infarction increase when prime complaint becomes available. The tendency is that if more information becomes available, the model becomes more certain about coming procedures for a patient. Figure 8 shows that hospital properties are quite relevant since the proposed procedures for a given patient can vary greatly between hospitals. We assign hospitals to the most likely cluster component H_h. Shown are procedure probabilities for the "diagnosis" *single live-born in hospital*. As one can see, the procedures in the different hospital clusters vary significantly.

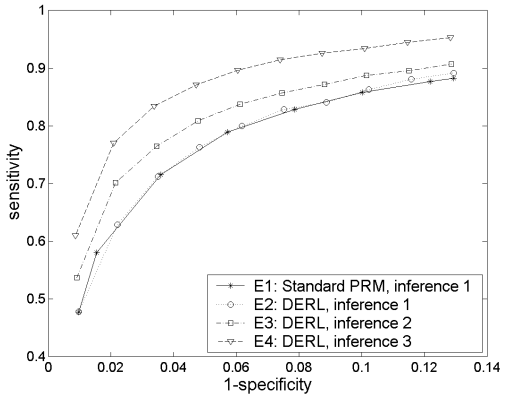

Figure 5. ROC curves for predicting procedures, given prime complaint and patient and hospital characteristics, average over all test patients.

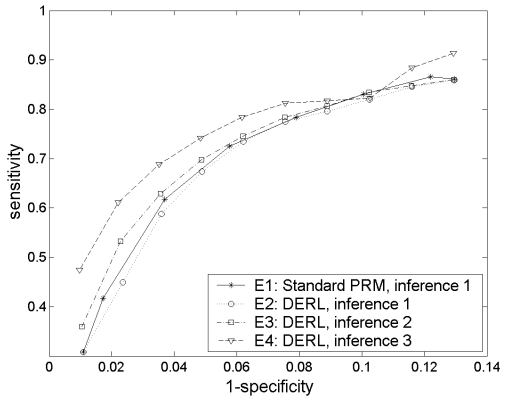

Figure 6. ROC curves for predicting procedures, given prime complaint *respiratory problem* and patient and hospital characteristics.

8. Conclusions

In this paper we have shown how nonparametric hierarchical Bayesian modelling can be very useful in re-

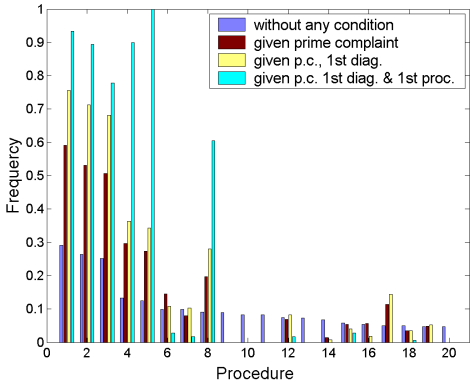

Figure 7. Procedures probabilities (see text).

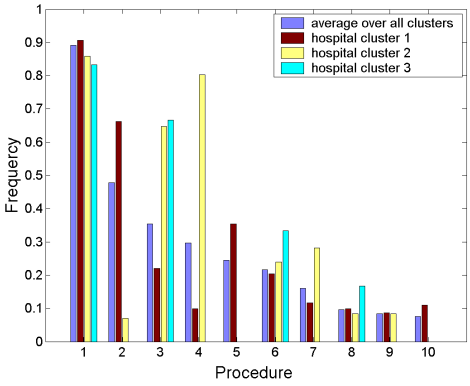

Figure 8. Procedures probabilities for different hospital clusters.

Table 1. Most frequent procedures for disease No. 410.71.

Rank	Code	Description
1	88.56	coronary arteriography using two catheters
2	37.22	left heart cardiac catheterization
3	88.53	angiocardiography of left heart structures
4	36.06	insertion of coronary artery stent(s)
5	36.01	single vessel percutaneous transluminal coronary angioplasty
6	99.20	injection or infusion of platelet inhibitor
7	36.15	single internal mammary-coronary artery bypass
8	39.61	extracorporeal circulation auxiliary to open heart surgery
9	88.72	diagnostic ultrasound of heart
10	99.04	transfusion of packed cells

lational learning. Parameters describing dependencies can be attributes of entities or relationships and can thus be non-global. The learned distribution can exhibit a rich structure and represent parameter dependencies which are impossible to represent in a parametric formulation. We demonstrated the advantages of our approach using data from a medical database. We used the nonparametric representation to model the reference uncertainty between patients and diagnosis and patients and procedures. Despite the fact that the base distribution exhibited parameter independence, the learned parameter distribution displayed parameter dependence. As a result the couplings between diagnosis and procedures could truthfully be modelled.

References

Dzeroski, S., & Lavrac, N. (Eds.). (2001). *Relational data mining*. Berlin: Springer.

Escobar, M. D., & West, M. (1998). *Computing bayesian nonparametric hierarchical models* (Technical Report 92-A20). Duke University, ISDS.

Friedman, N., Getoor, L. Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. *Proc. 16th JCAI* (pp. 1300–1309). Morgan Kaufmann.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. CRC press.

Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2003). Learning probabilistic models of link structure. *Journal of Machine Learning Research*, *3*, 679–707.

Heckerman, D., Meek, C., & Koller, D. (2004). *Probabilistic models for relational data* (Technical Report MSR-TR-2004-30). Microsoft.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press.

Raedt, L. D., & Kersting, K. (2003). Probabilistic logic learning. *SIGKDD Explor. Newsl.*, *5*, 31–48.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). *Hierarchical dirichlet processes* (Technical Report 653). UC Berkeley Statistics.

Tresp, V., & Yu, K. (2004). An introduction to nonparametric hierarchical bayesian modelling with a focus on multi-agent learning. In *Proc. the hamilton summer school on switching and learning in feedback systems*, 290–312. Springer.

Ullman, J. D., & Widom, J. (1997). *A first course in database systems*. Upper Saddle River,NJ,USA: Prentice Hall.

Wrobel, S. (2001). Inductive logic programming for knowledge discovery in databases. In S. Dzeroski and N. Lavrac (Eds.), *Relational data mining*, 74–101. Springer.

Yu, K., Tresp, V., & Yu, S. (2004). A nonparametric hierarchical bayesian framework for information filtering. *Proc. 27th SIGIR* (pp. 353–360). ACM.