

---

# Efficient probabilistic models for inference and learning

## *Individual Grant Review Report*

EPSRC grant GR/N07394

---

**Peter A. Flach**

Department of Computer Science, University of Bristol

### 1 Background and context

This project was concerned with enriching probabilistic models with structured knowledge representation. By a probabilistic model we mean any formalism that can be used to specify a complex probability distribution. For instance, a Bayesian network specifies a joint probability distribution over a tuple of random variables by means of a directed acyclic graph, in such a way that only the conditional probabilities of a variable given its immediate ancestors need to be considered. A hidden Markov model specifies a distribution over strings by extending a finite-state automaton that generates a language with probabilities attached to the transitions.

Such probabilistic models go a certain way to combine logic and probabilities, but there is considerable room for improvement. In particular, the logical knowledge most probabilistic models can handle is limited to simple propositions such as ‘the alarm went off’ or ‘the colour of this object is red’. First-order logic extends propositional logic by being able to reason about complex structured objects. For instance, in first-order logic we could say that this molecule contains an oxygen atom that has a bond with two hydrogen atoms, implying that a molecule is a complex object whose parts include atoms with their own properties. There is abundant structure in the world, and it is clear that first-order logic is in many cases a more precise and appropriate knowledge representation formalism than propositional logic, which can provide at best a crude approximation. The idea of integrating first-order logic and probability is not new. Roughly speaking, there are two approaches: either one starts from a first-order logic and extends it with probabilities (e.g., Muggleton’s stochastic logic programs), or one takes a probabilistic model and extends it with first-order features (e.g., Koller’s relational probabilistic models).

With respect to the goals of inference and learning, it is clear that there is a trade-off between expressiveness and efficiency. By progressing from propositional to first-order representations we are significantly increasing expressive-

ness, at the likely expense of decreasing efficiency for inference and learning. The challenge of devising first-order probabilistic models consists in striking the right balance between expressiveness and efficiency, by counteracting the richness of first-order representations with other restrictions in the model. Our conjecture was that approaches of the first kind mentioned above, such as stochastic logic programs, are too rich to be effectively learned and thus need to be restricted. On the other hand, approaches of the second kind, such as relational probabilistic models, seem to suffer from a somewhat awkward semantics which suggests that they do not exploit the real potential of first-order representations. This project aimed to investigate the middle ground between these two approaches. To this end, it was essential to keep both dual goals of inference and learning in mind.

The approach we have followed in this project was inspired by recent work on knowledge representation for inductive logic programming. Assuming that the universe of discourse consists of a homogenous collection of similar objects, the key idea is to concentrate on first-order representations that are based on a notion of individual, which is an aggregate of all there is to know about a particular object in the universe of discourse. Individual-centred representations are the key to understanding the relation between propositional learning and first- and higher-order learning.

The easiest way to perceive individual-centred representations is through typed programming languages. Each individual is represented by a term of a complex type, e.g. a tuple type, a set type, a list type, and so on. Complex terms have subterms, which themselves may be complex in general, individuals may be described by an arbitrary hierarchy of sets of tuples of lists of constants. It is readily seen that the joint distribution generated by a propositional Bayesian network corresponds to a cartesian product of atomic types, and the string distribution generated by a hidden Markov model corresponds to a list type constructed from a single atomic type. Employing a first-order representation means that the hierarchies can be more than one level deep, and that the complex terms can use any



formula to HBNs by considering the equivalent Bayesian network and renormalise (because there are fewer possible HBN structures than standard BN structures containing the same variables, given the type structure).

The next step is to find the HBN structure that maximises that expression. Since probabilistic links occur only between siblings in the type structure, this introduces a limit to the number of possible parents we need to consider for each node. Furthermore, if we apply an ordering on the nodes of the type structure (in fact, we only need an ordering for each subset of siblings) the number of possible structures decreases dramatically. Allowing p-links only between t-siblings reduces the maximum number of parents, but this is achieved in a domain-specific way, instead of simply imposing a hard limit for all the nodes. Our learning algorithm is a recursive search for the best possible p-link setup among a set of t-siblings (t-children of the same node in the type structure), beginning from the root of the type structure and proceeding towards the leaves.

Hierarchical Bayesian Networks provide a means for probabilistic **inference** on the variables they contain. The aim is, given a valuation for the evidence variables, to compute a probability distribution over the query variables. Our way of estimating the probability  $P(Q|E)$  for given valuations of sets of query and evidence variables  $Q$  and  $E$  is by sampling the HBN. Since the graph resulting from higher-level relationships is always acyclic, there exist a total ordering of the variables such that every variable is preceded by its higher-level parents. We can then assign values to the variables in that order, according for each variable to the conditional probability given its higher-level parents, which will have already been given a value. If we generate a large number of such instantiations, the relative frequency of the cases where  $Q$  holds divided by the relative frequency of the cases where  $E$  holds will converge to  $P(Q|E)$ . Convergence is improved using logic sampling, as described in e.g. [10]. We have also adapted Gibbs sampling to HBNs.

More details about HBNs can be found in [7, 6]. An application of HBNs to real-world data is described in Section 2.4 of this report.

## 2.2 First-Order Bayesian Classification

In the next stage of the project, we concentrated on a particular kind of Bayesian network, *viz.* the network expressing the independence assumptions embodied in the naive Bayesian classifier (i.e., a network with arcs from the class node to all attribute nodes). We then proceeded to relax the deterministic type hierarchy assumed by HBNs, by allowing t-parents with a one-to-many relationship to their t-children. For instance, a molecule consists of a variable number of atoms. The question then becomes how to define probability distributions over sets of objects from probability distributions over objects. In [3] we considered a

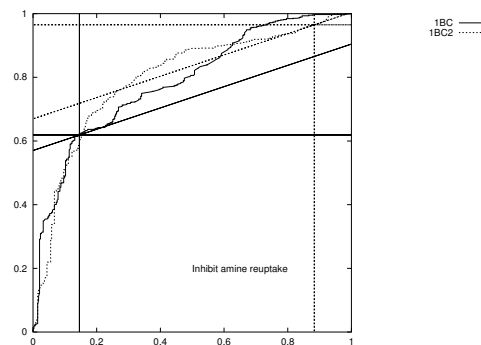


Figure 2: ROC curves in the Alzheimer’s disease domain for 1BC and 1BC2. The crosshairs denote the point chosen by the default probability threshold of 0.5, and the diagonal lines indicate the iso-accuracy lines at those points (higher lines are better).

number of possibilities, including a geometric distribution where a multi-sided die is rolled to generate objects until a stop symbol has been reached, and a distribution which iterates over all subsets of objects. The latter was found to be computationally expensive, as well as yielding virtually the same results as the geometric distribution in practice.

We then used these distributions to upgrade the 1BC system we developed earlier [2], which essentially applies a propositionalisation approach, to the 1BC2 first-order Bayesian classifier. **Learning** in 1BC2 involves estimating probability distributions over the leaf variables, as well as estimating the probability of encountering the stop symbol (the maximum likelihood estimate for this is  $\frac{1}{n+1}$ , where  $n$  is the average number of objects in a set); all these probabilities are taken conditional on the class. **Inference** involves calculating the probability distribution over the class variable given a valuation of the leaf variables.

We have obtained experimental results on several domains. Here we include results on a dataset involving drugs against Alzheimer’s disease; further experimental results can be found in [4]. Such drugs can have several desirable properties, one of which is to inhibit amine reuptake. Figure 2 shows ROC curves comparing the performance of the 1BC and 1BC2 classifiers.

More details about 1BC2 can be found in [8, 4].

## 2.3 Higher-Order Bayesian Networks

The final stage of the project was aimed at removing some of the unrealistic independence assumptions made by 1BC2. The approach was to incorporate a richer type structure by employing the higher-order logic of [9], which includes sets and multisets represented as lambda-expressions. This work is still ongoing as part of Elias Gyftodimos’ PhD thesis (expected submission deadline

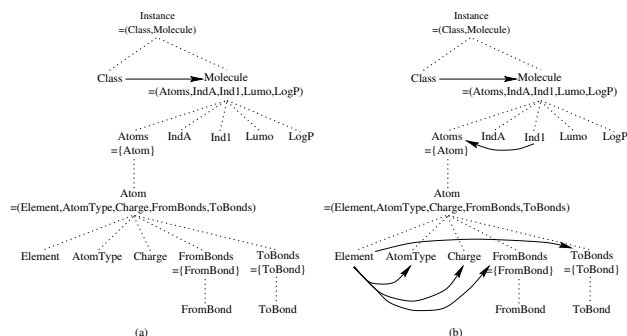


Figure 3: HOBN structures for the mutagenesis domain. (a) Under the naive Bayes assumption. (b) Extended structure.

Summer 2004). Here, we give an example of a Higher-Order Bayesian Network (HOBN) taken from [6].

We have applied our approach on the Mutagenesis dataset. Instances in this domain are molecular structures, and each one is described by four propositional attributes and a set of atoms. The atoms themselves are characterised by three propositional attributes and two sets of “incoming” and “outgoing” chemical bonds. The task is to predict whether particular molecules are mutagenic or not. For our experiments, we have constructed several HOBNs based on the same type structure for instances, and tested different sets of p-links between the nodes. We have employed sets for non-determinate aggregation, and used a geometric distribution (as in 1BC2) to compute conditional probabilities involving sets. Figure 3 shows two such structures, one corresponding to the “naive” assumption of attribute independence given the class (equivalent to the 1BC2 approach), and the other containing a set of p-links that was found to increase the accuracy considerably.

## 2.4 Application to human skill modelling

We have explored how HBNs can be applied to the problem of modelling right arm motion in cello playing. This problem has been studied by [5], where Bayesian networks are proposed as a suitable descriptive model. Their approach is based on analysing data measurements acquired by an amateur and a professional cello player executing a short music extract. The task is to build a model for each performer’s behaviour and see how differences in their playing are reflected in those models. This problem is inherently hierarchical and therefore well-suited for modelling by HBNs. Models are built observing firstly the position of different joints and secondly muscular activity of the right arm during the execution of a short musical extract.

In order to apply our approach to this data, we calculated angular velocity and acceleration for each instance as the first and second derivatives, respectively, of the corresponding angle variable. Subsequently, both angular and EMG

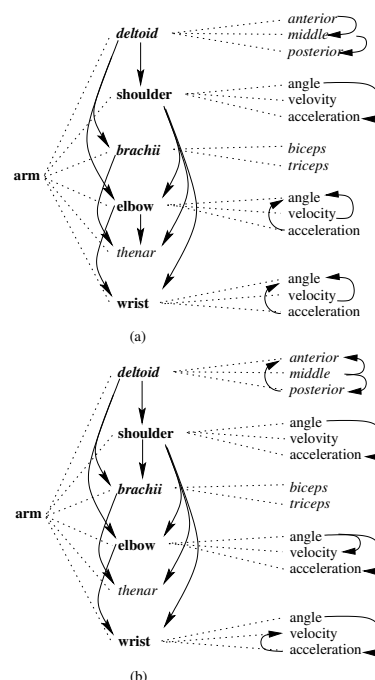


Figure 4: Musculoskeletal HBNs learned from data. (a) Amateur player. (b) Professional player.

data were quantised to a small number of possible values. Low pass filters were used throughout the process in order to eliminate noise. The derived values for each time point were used as independent instances of data.

We have conducted two series of experiments in order to learn HBNs that model the available data. In the first one we derived the probabilistic part for an HBN having the structure used by [5]. In the second series of experiments, only the type structure was given, and the task was to learn both the p-links and the probabilistic part. For efficiency reasons, an ordering on the nodes has been taken into account on the first level of the type structure, in the direction from the shoulder to the wrist. We make no further assumptions about dependencies between the nodes. The structures derived are shown in Figure 4, and deviate from the handcrafted Bayesian network structure used by [5]. Notable differences exist between the structures derived for the amateur and the professional. For instance, it can be seen that for the professional *shoulder* and *brachii* are independent given *deltoideid*, which does not hold for the amateur. There are also various differences at the lowest level of the type hierarchy.

We thus derived four different models: for each of the players we built two HBNs, one over a fixed HBN structure that was supplied as prior knowledge ( $HBN_P$ ) and one where the HBN structure was derived from the observations ( $HBN_D$ ). In order to evaluate those models further we applied the inference method described in Section 2.1 to compute the probabilities of a set of queries that reflect

Query	$HBN_{P,ama}$	$HBN_{P,pro}$	$HBN_{D,ama}$	$HBN_{D,pro}$
Q1	0.128	0.695	0.409	0.788
Q2	0.491	0.721	0.416	0.564
Q3	0.483	0.687	0.457	0.546
Q4	0.244	0.275	0.234	0.117
Q5	0.210	0.264	0.230	0.247

Table 1: Probabilities associated with various queries.

some intuitive rules for the domain.

- Q1  $P(\text{wrist.velocity} \neq 0)$ : This is a measure of the extent that the player was using his wrist.
- Q2  $P(\text{elbow.velocity} \neq 0 | \text{shoulder.angle} = \text{Closed})$ : Quantifies the movement of the elbow when the arm is close to the body.
- Q3  $P(\text{elbow.velocity} \neq 0 | \text{shoulder.angle} = \text{Open})$ : Quantifies the movement of the elbow when the arm is away from the body.
- Q4  $P(\text{triceps} = \text{High} | \text{elbow.velocity} > 0)$ : Measure of the triceps muscle activity when the elbow joint is opening.
- Q5  $P(\text{deltoid.anterior} = \text{High} | \text{elbow.velocity} > 0)$ : Measure of the anterior deltoid muscle activity when the elbow joint is opening.

Results displayed in Table 1 show that a lot of meaningful information is captured regarding the professional’s skill. From Q1 we see that the experienced player is making significantly more use of his wrist. Q2 and Q3 suggest that the professional is also moving his elbow more, for different positions of the shoulder joint. These observations reflect a known fact among cello tutors, namely that experienced students make better use of their wrist and elbow, whereas beginners tend to rely a lot on their shoulder for moving the bow. Q4 and Q5 show that the professional is more likely to be using his deltoid muscle than the brachii triceps when opening the elbow (“down bowing” movement); for the amateur, the probabilities are slightly higher in favour of using the brachii triceps. Note that this difference is better reflected in the  $HBN_D$  models.

### 3 Project plan review

We have stuck to and achieved the original objectives of the project: (1) to devise high-level (e.g. graphical) representations for individual-centred first-order probabilistic models; (2) to develop efficient inference methods for these first-order probabilistic models; (3) to develop techniques for learning first-order probabilistic models from data; and (4) to demonstrate the usefulness of first-order probabilistic models for practical inference and learning tasks.

The original project plan was roughly as follows: year 1 – graphical representation and inference methods; year 2 – learning methods; year 3 – experiments, integration and dissemination. Relatively early on it became clear that a more iterative approach was required, so that experience

with e.g. learning simple models could feed back into the graphical representation. We therefore abandoned the original project plan in favour of a three-stage plan investigating probabilistic models of increasing complexity, as outlined in the previous section. In each stage we addressed all of graphical representation, inference, learning, and experimental validation. I feel that this has worked well and has led to more significant results than would have been possible with the original project plan.

Some delay has occurred in that it has not been possible for the PhD student on the project to complete his PhD thesis in three years. Consequently, the work on Higher-Order Bayesian Networks is still ongoing (although unfunded) and some further results are expected.

### 4 Research impact and benefits to society

The project has resulted in two workshop papers [7, 6], a conference paper [8], and two journal papers [3, 4] (the latter one awaiting final approval). In addition Elias Gyftodimos is expected to submit his PhD thesis Summer 2004. I believe this is a considerable output given the modest resources of the project. All methods and algorithms described in this report have been implemented and have been or will be released in the public domain.<sup>2</sup>

There has been interest from other groups working on first-order probabilistic models. In November last year I have been invited to speak at the 4th Augustus de Morgan workshop on *Combining Logic and Probability* (King’s College, London). Elias Gyftodimos currently spends four months in the group of Prof Luc De Raedt at the University of Freiburg in Germany, on a Marie Curie Fellowship. I am also in touch with Dr David Page at the University of Wisconsin at Madison (USA), who is interested in applying our methods in bioinformatics.

More generally speaking, it is clear that the integration of logic and probability is one of the main outstanding problems in Artificial Intelligence, and that many application areas such as bioinformatics, information retrieval, skill modelling, etc. benefit from the advancement of this integration. For instance, two of the benchmark problems that we used for evaluation are from molecular biology, and we also worked on a significant application in human skill modelling. I believe that this project has contributed to this advancement.

### 5 Explanation of expenditure

The budget of £51,360 has been spent. On some categories we have spent more than budgeted, but never more than 15% more. On equipment we have spent consider-

<sup>2</sup><http://www.cs.bris.ac.uk/Research/MachineLearning/IBC/> and <http://www.cs.bris.ac.uk/Research/MachineLearning/fopm.html>.

ably less, because of the decreasing hardware prices and the availability of other equipment that could be used for this project.

## 6 Further research and dissemination activities

As already indicated, the work on Higher-Order Bayesian Networks is currently being extended as part of Elias Gyftodimos' PhD. We expect that a further 1-2 conference publications and 1 journal article will be published within the next 1-2 years. The HBN/HOBN software is currently being prepared for release in the public domain; there are plans at the University of Freiburg for an academic repository of software for first-order probabilistic modelling, and we have already been invited to contribute our software.

In the middle-long term we plan to apply this work to further problems in bioinformatics and information retrieval. We also plan to further study the relationship between HBN/HOBN and other first-order probabilistic models such as stochastic logic programs, Bayesian logic programs, and relational probabilistic models. Finally, we plan to investigate in more detail the independence assumptions that are made in HOBN, both in theory (which independence assumptions can we make?) and in practice (how do they affect the accuracy of inference?).

## References

- [1] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
- [2] P. Flach and N. Lachiche. 1BC: A first-order Bayesian classifier. In S. Džeroski and P. Flach, editors, *Proceedings of the 9th International Workshop on Inductive Logic Programming*, volume 1634 of *Lecture Notes in Artificial Intelligence*, pages 92–103. Springer-Verlag, 1999.
- [3] Peter A. Flach, Elias Gyftodimos, and Nicolas Lachiche. Probabilistic reasoning with terms. *Linkoping Electronic Articles in Computer and Information Science*, to appear. Available at <http://www.ida.liu.se/ext/epa/cis/2002/011/tcover.html>.
- [4] Peter A. Flach and Nicolas Lachiche. Naive Bayesian classification of structured data. *Machine Learning*, to appear. Conditionally accepted for publication.
- [5] K. Furukawa, S. Igarashi, K. Ueno, T. Ozaki, S. Morita, N. Tamagawa, T. Okuyama, and I. Kobayashi. Modeling human skill in bayesian network. *Linkoping Electronic Articles in Computer and Information Science*, to appear. Available at <http://www.ida.liu.se/ext/epa/cis/2002/012/tcover.html>.
- [6] E. Gyftodimos and P.A. Flach. Hierarchical Bayesian networks: an approach to classification and learning for structured data. In T. Horváth and A. Yamamoto, editors, *Proceedings of the Work-in-Progress Track at the 13th International Conference on Inductive Logic Programming*, pages 12–21. Department of Informatics, University of Szeged, September 2003.
- [7] Elias Gyftodimos and Peter A. Flach. Hierarchical bayesian networks: A probabilistic reasoning model for structured domains. In Edwin de Jong and Tim Oates, editors, *Proceedings of the ICML-2002 Workshop on Development of Representations*. University of New South Wales, Sydney, 2002.
- [8] N. Lachiche and P. A. Flach. 1BC2: a true first-order Bayesian classifier. In S. Matwin and C. Sammut, editors, *Proceedings of the 12th International Conference on Inductive Logic Programming*, volume 2583 of *Lecture Notes in Artificial Intelligence*, pages 133–148. Springer-Verlag, 2003.
- [9] J.W. Lloyd. *Logic for learning: learning comprehensible theories from structured data*. Springer-Verlag, 2003.
- [10] Stuart J. Russell and Peter Norvig. *Artificial intelligence, a modern approach*. Prentice Hall, 2nd edition, 1995.