Abductive Stochastic Logic Programs for Metabolic Network Inhibition Learning

Jianzhong Chen[1], Stephen Muggleton[1], and José Santos[2]

[1] Department of Computing, Imperial College London, London SW7 2AZ, UK
{cjz, shm}@doc.ic.ac.uk
[2] Division of Molecular BioSciences, Imperial College London, London SW7 2AZ, UK
jose.santos06@imperial.ac.uk

Abstract. We revisit an application developed originally using Inductive Logic Programming (ILP) by replacing the underlying Logic Program (LP) description with Stochastic Logic Programs (SLPs), one of the underlying Probabilistic ILP (PILP) frameworks. In both the ILP and PILP cases a mixture of abduction and induction are used. The abductive ILP approach used a variant of ILP for modelling inhibition in metabolic networks. The example data was derived from studies of the effects of toxins on rats using Nuclear Magnetic Resonance (NMR) time-trace analysis of their biofluids together with background knowledge representing a subset of the Kyoto Encyclopedia of Genes and Genomes (KEGG). The ILP approach learned logic models from non-probabilistic examples. The PILP approach applied in this paper is based on a general approach to introducing probability labels within a standard scientific experimental setting involving control and treatment data. Our results demonstrate that the PILP approach not only leads to a significant decrease in error accompanied by improved insight from the learned result but also provides a way of learning probabilistic logic models from probabilistic examples.

1 Introduction

One of the key open questions of Artificial Intelligence concerns *Probabilistic Logic Learning* (PLL) [3], i.e. the integration of probabilistic reasoning with first order logic representations and machine learning. PLL is also called Probabilistic ILP (PILP) [7] as it naturally extends Inductive Logic Programming (ILP) [6] to probabilistic case that can explicitely deal with uncertainty such as missing and noisy information. Although more and more new developments and successful applications have been published, there are still many challenges in the PLL research. One of such challenging questions is '*should PLL/PILP always learn from categorical examples?*' In other word, the data sets used by most PLL/PILP systems or applications are non-probabilistic, like those used in ILP systems. A major reason for the problem is we lack the corresponding methods to extract or estimate empirical probabilities from raw data. We attempt to show a solution to the problem in this paper by introducing Abductive Stochastic Logic Programs (SLPs) for metabolic network learning.

2 Metabolic Network Inhibition Learning

One of the applied machine learning approaches, which have been conducted to model the inhibitory effect of various toxins in the metabolic network of rats, is abductive ILP [8], a variant of ILP. A group of rats are injected with hydrazine and the changes on the concentrations of a number of chemical compounds are monitored during a period of time. The binary information on up/down regulations of metabolite concentrations following toxin treatment is combined with background knowledge representing a subset of the KEGG metabolic diagrams. An abductive ILP program is used to learn the potential inhibition occured in the network, which contains a set of *observables*, *abducibles*, *background facts* and general *background rules* under which the effect of the toxin can increase or reduce the concentration of the meabolites. The key point in the study is it supports an integration of *abduction* and *induction* [4] in an ILP setting, through which, the abductive explanations of the observations are added to the theory in a generalized form given by a process of induction on them.

3 Abductive Stochastic Logic Programs

Abductive SLPs [1] are a framework that supports abduction in SLPs [5] to provide a probability distribution over the abductive hypotheses based on their *possible worlds*. An *abductive SLP* S_A is a first order Stochastic Logic Program that supports stochastic abduction in the following way. Suppose e is an observed arbitry first order ground atom in S_A, $\delta(e, S_A)$ is a stochastic ground derivation of e derived from S_A involving a set of ground *abducibles* and a is an arbitrary abducible, then the probability of a can be defined to be the sum of the probabilities of all the least models that have a in their abduced facts

$$P(a) = \sum_{\delta(e,S_A) | a \in \delta(e,S_A)} P(e)P(\delta(e, S_A)).$$

To learn an abductive SLP S_A, we assume a background knowledge theory B in the form of a complete SLP and a set of independently observed ground probabilistic examples E (ie. each $e \in E$ is associated with an empirical probability $P(e)$). The learning is to construct a set of labelled hypothesised abducibles $H = \{p : ha\}$ such that when added to S_A, we have $B \wedge H \models E$ and the labels $\{p\}$ are chosen to maximize the likelihood of H given E and B

$$L(H \mid E, B) = P(E \mid H, B) = \prod_{e \in E} \sum_{\delta \in SS(e,H,B)} P(e)P(\delta(e, H, B)),$$

where $SS(e, H, B)$ denotes the set of all stochastic SLD derivations of e from the model (H, B). In practice, we perform SLP parameter estimation algorithms, such as FAM [2], to learn the probabilities for a given set of abducibles.

```
1. Initialize a matrix $MR$ with column=2 and row=number of metabolites;
2. for each metabolite $\alpha$ do:
   2.1. $C_\alpha=\{concentration(\alpha)\}$, a set of $\alpha$ values observed in the **control** cases;
   2.2. $M_\alpha=\textsc{Mean}(C_\alpha)$,$SD_\alpha=\textsc{StandardDeviation}(C_\alpha)$;
   2.3. $T_\alpha=\{concentration'(\alpha)\}$, a set of $\tau_\alpha$ values observed in the **treatment** cases;
   2.4. $MR[\alpha,1] = M_\alpha < \textsc{Mean}(T_\alpha)$ ? Up : Down;
   2.5. $MR[\alpha,2]= \textsc{Mean}(\{\textsc{pnorm}(\tau_\alpha, M_\alpha, SD_\alpha)\})$;
3. Apply matrix $MR$ in the abductive SLP learning
```

Table 1. Algorithm of estimating empirical probabilities from control/treatment data for metabolic network inhibition

4 Extracting Probabilistic Examples from Scientific Data

Assume we have a scientific data set involving a set of data values collected from some control cases as well as a set of data points from some treated cases. All the data are mutually independent. Table 1 presents an algorithm applied to our rat metabolic network inhibition data set for extracting the probabilistic examples from empirical data. The average of the integral (using the PNORM function in the R Language) $MR[\alpha, 2]$ is considered to be the estimated empirical probability of α happened in the treatment cases against the control cases.

5 Experiments - Learning Metabolic Network Inhibition

The experiments include two learning tasks – learning abductive SLP_C from categorical (non-probabilistic) examples and learning abductive SLP_P from probabilistic examples. In particular, each observation inputted into SLP_P is associated with an estimated empirical probability ρ we have obtained in last section.
Null hypotheses: The predictive accuracy of an SLP_P model **does not** outperform an SLP_C model for predicting the concentration level of metabolites in a given rat metabolic network inhibition experiment.
Materials and Methods: The (estimated) empirical probabilities are extracted from the raw data consisting of the concentration level of 20 metabolites on 20 rats (10 control cases and 10 treated cases) after 8 hours of the injection of hydrazine. The initial SLP uses background knowledge derived from the ILP model. We apply a leave-one-out cross validation process to do the prediction and evaluation. The learning tasks are performed by playing FAM (using Pe-pl software) under Yap 5.1.1.
Results: The cross validation models are high likely overfitting the training data due to the fact that the number of parameters (150) is much more than the number of examples (20); the (log)likelihood generated by FAM and the predictive accuracy of the models are not correlated; the number of iterations and the depth of recursion affected both the running time and the accuracy of the models. Therefore, the evaluation of the prediction models is made by calculating the average predictive accuracy of SLP_C and SLP_P over n iterations for a given

recursion depth and against the estimated empirical probabilities respectively, with a statistical significance calculated over n iterations. In particular, when evaluating only with the categorical observations, the SLP_C models correctly predicted 12 out of 20 metabolites, which is close to the result of ILP models (about 62% for all the time points); moreover, when evaluating with the probabilistic examples in the case of recursion depth 1, SLP_P is consistently more accurate than SLP_C in all the 10 iterations; SLP_P outperforms SLP_C by 69.0% against 63.64% in average predictive accuracy with a significance level of 0.03%.
Interpretability: By comparing the learned SLP model with the previous ILP model, at least two promising new findings have been discovered in the SLP model, which can be explained by the introduced empirical probabilities. In addition, the SLP models learned not only the patterns but also the degree of belief of the patterns which improve the insight from the learned models.

6 Discussion and Conclusions

In conclusion, the null hypotheses we have set in the paper and experiments were rejected on the bases of the theoretical and experimental results. Our results demonstrate that the PILP approach not only leads to a significant decrease in error accompanied by improved insight from the learned result but also provides a way of learning probabilistic logic models from probabilistic examples. Future work include further study of the abductive SLPs in theory and more robust experiments and comparisons in practice.

References

1. A. Arvanitis, S.H. Muggleton, J. Chen, and H. Watanabe. Abduction with stochastic logic programs based on a possible worlds semantics. In *Short Paper Proceedings of the 16th International Conference on Inductive Logic Programming*. University of Corunna, 2006.
2. J. Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.
3. L. De Raedt and K. Kersting. Probabilistic Logic Learning. *ACM-SIGKDD Explorations: Special issue on Multi-Relational Data Mining*, 5(1):31–48, 2003.
4. P. Flach and A. Kakas (editors). *Abductive and Inductive Reasoning.* Pure and Applied Logic. Kluwer, 2000.
5. S.H. Muggleton. Stochastic logic programs. In L. de Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, 1996.
6. S.H. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19,20:629–679, 1994.
7. L. De Raedt and K. Kersting. Probabilistic inductive logic programming. In S. Ben-David, J. Case, and A. Maruoka, editors, *Proceedings of the 15th International Conference on Algorithmic Learning Theory*, volume 3244 of *Lecture Notes in Computer Science*. Springer-Verlag, 2004.
8. A. Tamaddoni-Nezhad, R. Chaleil, A. Kakas, and S.H. Muggleton. Application of abductive ILP to learning metabolic network inhibition from temporal data. *Machine Learning*, 64:209–230, 2006. DOI: 10.1007/s10994-006-8988-x.