Learning Bayes Nets with Link Uncertainty for Relational Data Sets

Oliver Schulte and Zhensong Qian *

School of Computing Science
Simon Fraser University
Vancouver-Burnaby, Canada

Abstract

Many if not most big data sets are maintained in relational databases. We describe Bayes net learning methods that can discover knowledge about correlations among both link types and node attributes in big relational data.

1 Introduction: Link Correlations

Scalable link analysis for relational data with multiple link types is a challenging problem in network science. We describe a method for learning a Bayes net that captures simultaneously correlations between link types, link features, and attributes of nodes.

Building a Bayes net model is useful for big data analysis because such models provide a compact summary of the statistical relationships in the data. The model supports both descriptive and predictive analytics. Correlations are presented to the user in a graphical way, and queries about probabilistic relationships can be answered quickly using Bayes net inference rather than via database queries run against a large dataset.

Previous work on learning Bayes nets for relational data was restricted to correlations among attributes given the existence of links [4]. The larger class of correlations examined in our new algorithms includes two additional kinds:

1. Dependencies between two different types of links.
2. Dependencies among node attributes given the *absence* of a link between the node.

Contributions include the following:

1. To our knowledge this is the first implementation of Bayes net learning for modelling correlations among different types of links.
2. A comparison of a hierarchical vs. a single-table model search strategy.

*This research was supported by a Discovery Grant to Oliver Schulte from the Canadian Natural Sciences and Engineering Council. And Zhensong Qian was also supported by a grant from the China Scholarship Council. This is a preliminary version of a paper that will appear in the post proceedings of the IJCAI 2013 GKR workshop.

2 Background and Notation

Poole introduced the Parametrized Bayes net (PBN) formalism that combines Bayes nets with logical-relational syntax [2]. A **population** is a set of individuals. A **population variable** is capitalized. A **functor** represents a function or a Boolean predicate. A predicate with more than one argument is called a **relationship**; other functors are called **attributes**. A **Parametrized random variable** (PRV) is of the form $f(X_1, \ldots, X_a)$, where the populations associated with the variables are of the appropriate type for the functor. A **Parametrized Bayes Net (PBN)** structure is a directed acyclic graph whose nodes are PRVs.

We assume that data are represented in a standard **relational schema** containing a set of tables, each with key fields, descriptive attributes, and possibly foreign key pointers. The powerset of relationship tables can be ordered as a lattice (e.g., $\{Reg(S, C)\} \sqsubseteq \{Reg(S, C), Teaches(C, P)\}$. For each relationship set, there is a **data table** whose columns consist of: (1) the attributes of all entities/relationships involved in the set, and (2) *a Boolean relationship node* for each relationship, that records whether the relationship holds between two entities. For an illustration of these concepts see Figure 1.

Methods Compared We compared the following methods.

Flat Applies a single-table Bayes net learner to the maximal data table comprising all relationship sets in the database. The results of [3] provide a theoretical justification for this procedure.

LAJ The previous hierarchical learn-and-join method [4] without relationship nodes in the data table and hence without link correlations. Conducts bottom-up search through the lattice of relationship sets. Dependencies (Bayes net edges) discovered for smaller sets are propagated to larger sets.

LAJ+ The new LAJ method with relationship data that has the potential to find link correlations.

3 Evaluation

For the details of the system setup, the datasets, and the fast Möbius transform please see [4]. We report learning time,

(a)

Student		
s-id	Intelligence	Ranking
Jack	3	1
Kim	2	1
Paul	1	2

(b)

Registration			
s-id	C.id	Grade	Satisfaction
Jack	101	A	1
Jack	102	B	2
Kim	102	A	1
Kim	103	A	1
Paul	101	B	1
Paul	102	C	2

(c)

Course					
c-id	Prof	Rating	Difficulty	T-a-prof	P-prof
101	Oliver	3	1	1	3
102	David	2	2	1	2
103	Oliver	3	2	1	3

(d)

attribute-relation table								
Intelligence	Rank	Rating	Diff	P-prof	t-a-prof	Reg(S,C)	R.Grade	R.Sat
3	1	3	1	3	1	T	A	1
3	1	2	2	2	1	T	B	2
3	1	3	2	3	1	F	⊥	⊥
2	1	3	1	3	1	F	⊥	⊥
2	1	2	2	2	1	T	A	1
2	1	3	2	3	1	T	A	1
1	2	3	1	3	1	T	B	1
1	2	2	2	2	1	T	C	2
1	2	3	2	3	1	F	⊥	⊥

(e)

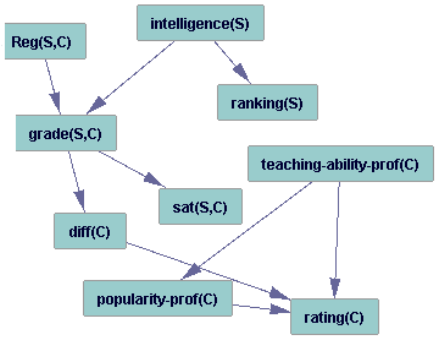

Figure 1: Database Table Instances: (a) *Student*, (b) *Registered* (c) *Course*. To simplify, we added the information about professors to the courses that they teach. (d) The data table for $Registered(S, C)$, which lists for each pair of entities their descriptive attributes, whether they are linked by *Registered*, and the attributes of a link if it exists. $\perp$ means "not applicable." (e) A Parametrized Bayes Net for the university schema.

Dataset	**Flat**	**LAJ+**	**LAJ**
University	1.916	1.183	0.291
Movielens	38.767	18.204	1.769
Mutagenesis	3.231	3.448	0.982
Small-Hepatitis	9429.884	8.949	10.617

Table 1: Model Structure Learning Time in seconds.

log-likelihood, Bayes Information Criterion (BIC), and the Akaike Information Criterion (AIC) [1].

Learning Times Table 1 provides the model search time for each of the link analysis methods. This does not include the time for computing table joins since this is essentially the same for all methods (the cost of the full table join). On the smaller and simpler datasets, all search strategies are fast, but on the medium-size and more complex datasets (Hepatitis, MovieLens), hierarchical search is much faster due to its use of constraints.

Statistical Scores On the medium-sized dataset MovieLens, which has a simple structure, all three methods score similarly. LAJ and LAJ+ return the same model. The most complex dataset, Hepatitis, is a challenge for flat search, which overfits severely. Because of the complex structure of the Hepatitis schema, the hierarchical constraints are effective in combating overfitting. The situation is reversed on the Mutagenesis dataset where flat search does much better than hierarchical search. The reason for that is that, unusually, links in Mutagenesis are dense. As a result, we find strong correlations between attributes conditional on *the absence of relationships*. Our current version of the LAJ+ algorithm cannot detect such correlations; we leave an appropriate extension for future work.

University	BIC	AIC	log-likelihood	# Parameter
Flat	-17638.27	-12496.72	-10702.72	1767
LAJ+	-13495.34	-11540.75	-10858.75	655
LAJ	-13043.17	-11469.75	-10920.75	522

MovieLens	BIC	AIC	log-likelihood	# Parameter
Flat	-4912286.87	-4911176.01	-4910995.01	169
LAJ+	-4911339.74	-4910320.94	-4910154.94	154
LAJ	-4911339.74	-4910320.94	-4910154.94	154

Mutagenesis	BIC	AIC	log-likelihood	# Parameter
Flat	-21844.67	-17481.03	-16155.03	1289
LAJ+	-47185.43	-28480.33	-22796.33	5647
LAJ	-30534.26	-25890.89	-24479.89	1374

Hepatitis	BIC	AIC	log-likelihood	# Parameter
Flat	-7334391.72	-1667015.81	-301600.81	1365357
LAJ+	-457594.18	-447740.51	-445366.51	2316
LAJ	-461802.76	-452306.05	-450018.05	2230

Table 2: Performance of different Model Search Algorithms by dataset.

4 Conclusion

We described different methods for extending relational Bayes net learning to correlations involving links. Statistical measures indicate that Bayes net methods succeed in finding relevant correlations. There is a trade-off between statistical power and computational feasibility (full table search vs constrained search). Hierarchical search often does well on both dimensions, but needs to be extended to handle correlations conditional on the absence of relationships.

To improve scalability, computing sufficient statistics needs to be feasible for cross product sizes in the millions or more. A promising solution is to utilize virtual join methods that precompute sufficient statistics without materializing table joins such as the Fast Möbius Transform [4] and tuple ID propagation [5].

References

[1] D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2003.

[2] David Poole. First-order probabilistic inference. In *IJCAI*, pages 985–991, 2003.

[3] Oliver Schulte. A tractable pseudo-likelihood function for Bayes nets applied to relational data. In *SIAM SDM*, pages 462–473, 2011.

[4] Oliver Schulte and Hassan Khosravi. Learning graphical models for relational data via lattice search. *Machine Learning*, 88(3):331–368, 2012.

[5] Xiaoxin Yin, Jiawei Han, Jiong Yang, and Philip S. Yu. Crossmine: Efficient classification across multiple database relations. In *ICDE*, pages 399–410. IEEE Computer Society, 2004.