A Study of MEBN Learning for Relational Model

Cheol Young Park, Kathryn Blackmond Laskey, Paulo Costa, Shou Matsumoto
Volgenau School of Engineering
George Mason University
Fairfax, VA USA
[cparkf, klaskey, pcosta]@gmu.edu, smatsum2@masonlive.gmu.edu

***Abstract*— In the past decade, Statistical Relational Learning (SRL) has emerged as a new branch of machine learning for representing and learning a joint probability distribution over relational data. Relational representations have the necessary expressive power for important real-world problems, but until recently have not supported uncertainty. Statistical relational models fill this gap. Among the languages recently developed for statistical relational representations is Multi-Entity Bayesian Networks (MEBN). MEBN is the logical basis for Probabilistic OWL (PR-OWL), a language for uncertainty reasoning in the Semantic Web. However, until now there has been no implementation of MEBN learning. This paper describes the first implementation of MEBN learning. The algorithm learns a MEBN theory for a domain from data stored in a relational Database. Several issues are addressed such as aggregating influences, optimization problem, and so on. In this paper, as our contributions, we will provide a MEBN-RM (Relational Model) Model which is a bridge between MEBN and RM, and suggest a basic structure learning algorithm for MEBN. And the method was applied to a test case of a maritime domain in order to prove our basic method.**

Keywords: **Probabilistic ontology, Multi-Entity Bayesian networks, PR-OWL, Relational Model/Database, Machine Learning, Statistical Relational Learning**

I. INTRODUCTION

Statistical Relational Learning (SRL) is a new branch of machine learning for representing and learning a joint distribution over relational data [1, 2]. As its name suggests, it combines statistical and relational knowledge representations. A relational model represents a domain as a collection of objects that may have attributes and can participate in relationships with other objects. Relational representations are expressive enough for important real-world problems, but until recently have not supported uncertainty. This gap has been filled by SRL methods. Statistical relational knowledge representations combine statistical and relational approaches, allowing representation of a probability distribution over a relational model of a domain. SRL methods allow such representations to be learned from data.

Examples of representation languages for SRL include Probabilistic Relational Models (PRMs), Markov Logic Networks (MLNs), Relational Dependency Networks (RDNs), Bayesian Logic Programs (BLPs), Join Bayes Net (JBN), and Multi-Entity Bayesian Networks (MEBN) [2, 3, 4, 5, 6, and 7].

A comparison of some of the above models is given in [1]. Typically, SRL models provide a representation for relational knowledge, along with methods for both induction and deduction. Relational representations provide both class and instance models. A class model describes statistical information that applies to classes of objects. For example, a class model might describe the false positive and false negative rates for a class of sensor. The instance model is generated from the class model by a deduction method. For example, the instance model would be used to infer the probability that a given detection is a false positive. An induction method learns structure and parameters of a domain theory from observations. For example, induction would be used to learn the false positive and false negative rates from a data set annotated with ground truth.

SRLs have been applied to problems such as Object Classification, Object Type Prediction, Link Type Prediction, Predicting Link Existence, Link Cardinality Estimation, Entity Resolution, Group Detection, Sub-graph Discovery, Metadata Mining, and so on [2].

This paper is concerned with the Multi-Entity Bayesian Networks (MEBN), a relational language that forms the logical basis of Probabilistic OWL (PR-OWL), a language for uncertainty reasoning in the Semantic Web [7, 8]. PR-OWL has been extended to PR-OWL 2, which provides a tighter link between the deterministic and probabilistic aspects of the ontology [9]. MEBN extends Bayesian networks to a relational representation. A MEBN Theory, or MTheory, consists of a set of Bayesian network fragments, or MFrags, that together represent a joint distribution over instances of the random variables represented in the MTheory [7].

However, until now there has been no implementation of induction or learning for MEBN or PR-OWL. This paper describes such an implementation. We follow an approach used by other SRL models [1] and use Relational Database (RDB) to store the observations from which the representation is learned.

This paper focuses a basic learning algorithm that addresses the following issues:

1. Developing a bridge of MEBN and RDB;
2. Developing basic structure and parameter learning for MEBN.

Ultimately, a relational learning algorithm should address issues such as aggregation of data, reference uncertainty, type uncertainty, and continuous variable learning. These issues will be considered for future research.

Our learning method is exact, and assumes discrete random variables, and complete data. It will be evaluated by the inference accuracy test.

In Section 2, we give a brief definition of MEBN and RM as background. In the Section 3, we introduce the MEBN-RM Model. In Section 4, we present the basic structure learning algorithm. The application of the algorithm is described in the Section 5.

II. Multi-Entity Bayesian Networks (MEBN) and Relational Model (RM)

A. Multi-Entity Bayesian Networks (MEBN)

MEBN extends Bayesian Networks (BNs) to represent relational information. BNs have been very successful as an approach to representing uncertainty about many interrelated variables. However, BNs are not expressive enough for relational domains. MEBN extends Bayesian networks to represent the repeated structure of relational domains.

MEBN represents knowledge about a domain as a collection of MFrags, an MFrag is a fragment of a graphical model that is a template of probabilistic relationships among instances of its random variables. Random variables in an MFrag can contain ordinary variables which can be filled in with domain entities. And MFrag includes context, input, and resident node for restriction of entity, reference of node, and random variable respectively. We can think of an MFrag as a class which can generate instances of BN fragments, which can then be assembled into a Bayesian network [7].

B. Relational Model (RM)

In 1969, Edgar F. Codd proposed RM as a database model based on first-order predicate logic [10]. RM is composed of Relation, Attribute, Key, Tuple, Instance, and Cell. Relational database which is the most popular database is based on RM.

III. MEBN-RM Model

As a bridge of MEBN and RM, we suggest MEBN-RM Model which provides a specification for how to match elements of MEBN to elements of RM. Key nodes in MEBN are the context and resident node. To understand this easily, we use the following example of the university relational model.

Course	
Key	Difficulty
c1	low
c2	high
c3	high
c4	low
c5	med
c6	low

Registration		
Course Key	Student Key	Grade
c1	s1	low
c1	s2	high
c2	s2	high
c2	s4	low
c3	s5	med
c4	s6	low

Student	
Key	Advisor
s1	p4
s2	p2
s3	p3
s4	p1
s5	p5
s6	null

Professor	
Key	Major
p1	SYST
p2	OR
p3	OR
p4	CS
p5	SYST
p6	OR

Table 1. Example of university relational model

A. Context Node

In MFrags, context terms (or nodes) are used to specify constraints under which the local distributions apply. Thus, it determines specific entities on an arbitrary situation of a context. In MEBN-RM model, we define four types of data structure corresponding to context nodes: Isa, Slot-filler, Value-Constraint, and Entity-Constraint type.

Type	Name	Example
1	Isa	Isa(Person, P), Isa(Car, C)
2	Value-Constraint	Height(P) = high
3	Slot-Filler	P = OwnerOf(C)
4	Entity-Constraint	Friend(A, B)

Table 2. Context Node Types on MEBN-RM Model

1) Isa

In MEBN, the Isa random variable represents the type of an entity. In a RM, an entity table represents a collection of entities of a given type. Thus, an entity table corresponds to an Isa random variable in MEBN. Note that a relationship table whose primary key is composed of foreign keys does *not* correspond to an Isa RV. A relationship table will correspond to the Entity-Constraint type of Context Node.

2) Value-Constraint

In a case, a value of attribute can limit keys which are related with only the value. For example, Consider Table 1, in which we have the course table with the difficulty attribute. (In our definition, Attribute is descriptive Attribute and Key is Primary Key)

The course table has instances of the key (e.g., c1, c2, c3, c4, c5, and c6). And if we want to focus on a case of the entity with "high" value of the attribute, it will be {c2, c3}. In this case, for the entity, any group of elements related with any attributes can be derived. We encode this into "Difficulty (Course) = high" in MEBN.

3) Slot-Filler

In the table 1, the professor key is used on the student table by a foreign key, Advisor. The foreign key is not primary key in the student table. In this case, the connection will be expressed by "Professor = Advisor (Student)" in MEBN. And its instance will be that s1's advisor is p4 and so on.

4) Entity-Constraint

The registration table is a relationship table which is a bridge between the course and student entity. In this case, obviously, the registration table will be an intersection group. And this is described as "Registration (Course, Student)" in MEBN.

B. Resident Node

In MFrags, Resident Node can be described as Function, Predicate, and Formula of FOL with a probability distribution. FOL Function consists of arguments and an output, while FOL Predicate consists of arguments and no output, but Boolean output. We define the following relationship between elements of RM and MEBN.

RM	Resident Node
Attribute	Function/ Predicate
Key	Arguments
Cell of Attribute	Output

Table 3. Resident Node Types of MEBN-RM Model

For example, in the table 1, the grade of the registration table is the function having the course and student keys as

arguments. Its output will be the cell of the grade such as low, med, and high. On the other hand, if the domain type of the grade is Boolean, it will be the predicate in MEBN.

IV. The Basic Structure Learning For MEBN

To address the issues in Section 1, we suggest a basic structure learning algorithm for MEBN. The initial ingredients of the algorithm are a dataset of RM, a Bayesian Network Structure searching algorithm, and a size of chain. For the parameter learning, we only use Maximum Likelihood Estimation (MLE). The algorithm focuses on discrete variables with complete data. We utilize a standard Bayesian Network Structure searching algorithm to generate a local BN from the joined dataset of RM. To avoid infinite loops, we employed the size of chain. Thus, the process of searching structure will finish in the size of chain.

Firstly, the algorithm creates the default MTheory. All keys of DB are defined as entities of MEBN theory. One default reference MFrag is created. For the all of tables of DB, the dataset for each table is retrieved and, by using the BN structure searching algorithm, a graph is generated from the dataset. If the graph has a cycle and undirected edge, a knowledge expert for the domain sets the arc direction. Based on the revised graph, an MFrag is created. Until the size of chain is reached, the joined datasets which are derived by "Join" command in SQL are retrieved. The graphs related to the joined datasets are generated in the same way as the above. If any nodes of the new generated graph are not used in any MFrags, create the resident node having the name of the dataset of the graph on the default reference MFrag and the new MFrag for the dataset. If not, only make edges between resident nodes in the different MFrags. Lastly, for all resident nodes in the MTheory, LPDs are generated by MLE.

V. Case Study

To evaluate the algorithm, we used a dataset which came from the PROGNOS (Probabilistic OntoloGies for Net-centric Operation Systems) [11, 12]. The purpose of the system is to provide higher-level knowledge representation, fusion, and reasoning in the maritime domain.

The PROGNOS includes a simulation which provides the ground truth information for the system. The simulation uses a given single entity Bayesian Network (we use this term to discriminate the SSBN from Multi Entity Bayesian Networks) in order for sampling data. The simulation generates 85000 persons, 10000 ships, and 1000 organization entities with various values of attributes. The data for these entities are stored in the relational database.

For the evaluation of the model, the training and test dataset was generated by the simulation. Using the basic structure learning for MEBN, the PROGNOS MTheory was derived as shown in Figure 2. In the model, a total of four MFrags were generated such as the default reference, org_members, person, and ship MFrag.

To generate a SSBN from this MTheory, we assume that we have one person, ship, and organization. They are related as ship_crews (Ship S, Person P) and org_members(Organization O, Person P). We queried the isShipOfInterest node with the several evidence nodes located in the leaf nodes. Figure 3 presents the result SSBN in which the nodes of the ship and person entity are connected each other.

To compare the accuracies of the results, we used the single entity Bayesian Network which was used for the sampling. Thus, the single network provided another query result with the same evidence. Figure 1 shows the Receiver Operating Characteristic (ROC) Curve which describes accuracy of the result of the learned MTheory and single entity Bayesian Network. The areas under curves are shown in Table 4.

Model	AUC
Learned MTheory	0.874479929
Single Entity Bayesian Network	0.87323784

Table 4. AUC of Learned MTheory and Single Entity Bayesian Network

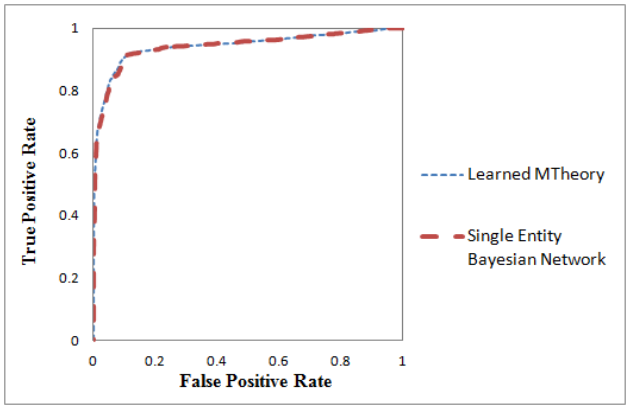

Figure 1. ROC of Learned MTheory and Single Entity Bayesian Network

As we can see from Figure 1 and Table 4, the results of accuracy of the learned MTheory and the single entity Bayesian Network are almost the same. This means that the learned MTheory well reflected the data of the relational database which was sampled using the single entity Bayesian Network.

In this paper, we only compared the learned MTheory to the true model which was the single entity Bayesian Network. This result proves that our approach reflects the true model correctly. However, the result of this paper is only the beginning and baseline for a full MEBN Learning method, because we didn't address the aggregating influence problem which is the important issue in SRL models.

VI. DISCUSSION AND FUTURE WORK

Because of a flood of complex and huge data, efficient and accurate methods are needed for learning expressive models incorporating uncertainty. In this paper, we have introduced a learning approach for MEBN. As a bridge between MEBN and RM, MEBN-RM Model was introduced. For induction, the Basic Structure Learning for MEBN was suggested.

Recently, we are studying about a heuristic approach which called as the Framework of Function Searching for LPD (FFS-LPD) to address the aggregating influence problem. We plan to expand the learning algorithm in order to include continuous random variables.

REFERENCES

[1] Hassan Khosravi, Bahareh Bina. A Survey on Statistical Relational Learning. In Proceedings of Canadian Conference on AI'2010. pp.256~268

[2] Getoor, L., Tasker, B.: Introduction to statistical relational learning. MIT Press, Cambridge, 2007

[3] Domingos, P., Richardson, M.: Markov logic: A unifying framework for statistical relational learning. In: Introduction to Statistical Relational Learning, ch. 12, pp. 339–367, 2007

[4] Nevile, J., Jensen, D.: Relational dependency networks. In: An Introduction to Statistical Relational Learning

[5] Kersting, K., de Raedt, L.: Bayesian logic programming: Theory and tool. In: Introduction to Statistical Relational Learning

[6] Oliver Schulte, Hassan Khosravi, Flavia Moser, and Martin Ester. Join bayes nets: A new type of bayes net for relational data. Technical Report 2008-17, Simon Fraser University, 2008. also in CS-Learning Preprint Archive.

[7] Laskey, K. B.,MEBN: A Language for First-Order Bayesian Knowledge Bases. Artificial Intelligence, 172(2-3), 2008

[8] Paulo C. G Costa, Bayesian Semantics for the Semantic Web. PhD Dissertation, George Mason University, July 2005. Brazilian Air Force.

[9] Rommel N. Carvalho, Probabilistic Ontology: Representation and Modeling Methodology, PhD Dissertation, George Mason University, July 2011.

[10] Codd, E.F. "A Relational Model of Data for Large Shared Data Banks". Communications of the ACM, 1970

[11] P.C.G. Costa, K.B. Laskey, and KC Chang, "PROGNOS: Applying Probabilistic Ontologies To Distributed Predictive Situation Assessment In Naval Operations." Proceedings of the 14th Int. Command And Control Research and Technology Symposium, Washington, D.C., USA, 2009.

[12] R. N. Carvalho, P. C. G. Costa, K. B. Laskey, and K. Chang, "PROGNOS: predictive situational awareness with probabilistic ontologies," in Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, Jul. 2010.

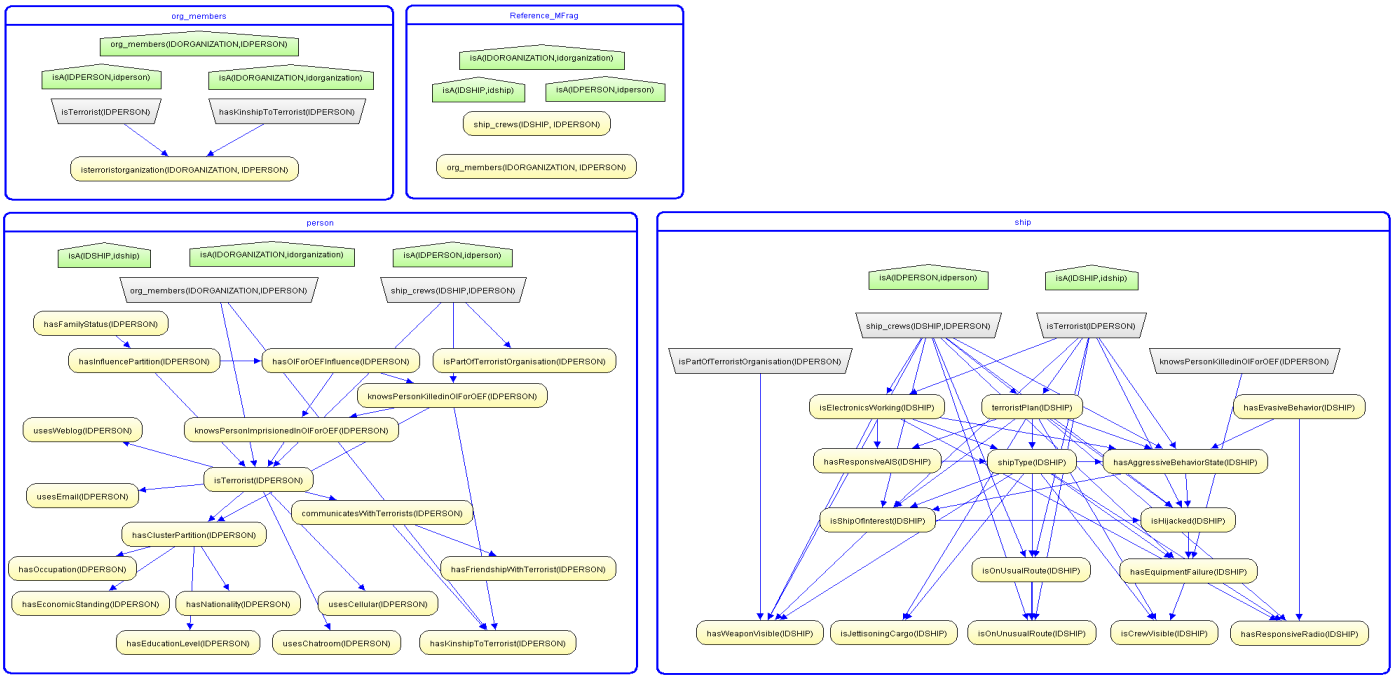

Figure 2. Generated PROGNOS MTheory

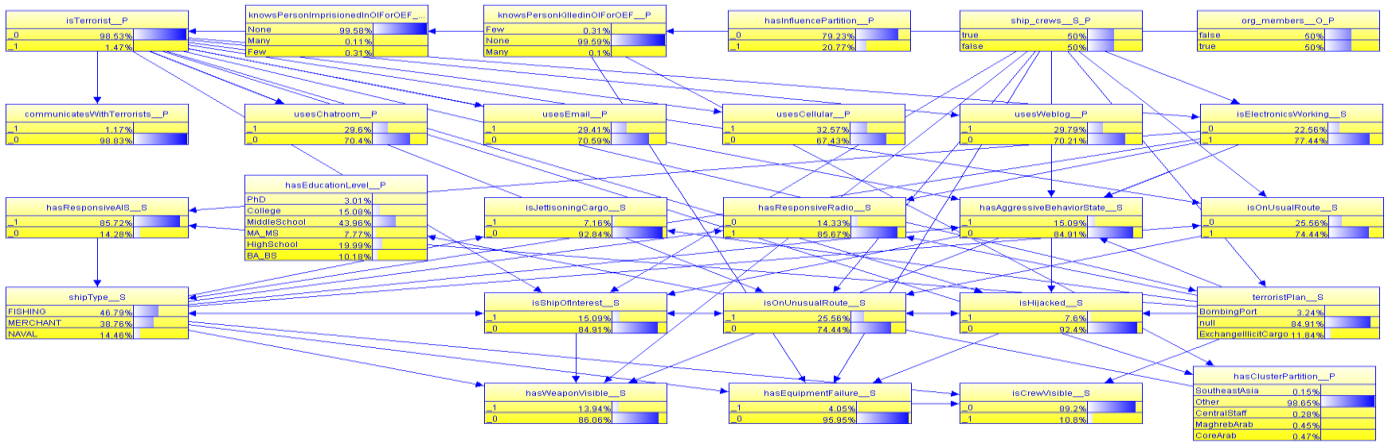

Figure 3. Generated SSBN of PROGNOS MTheory