2D-06

Learning Probabilistic Relational Models: A Novel Approach

Luiz Henrique Barbosa Mormille[†] Fabio Gagliardi Cozman[‡]
Soka University[†] Universidade de São Paulo[‡]

1. INTRODUCTION

Most large data collections are stored in relational database systems consisting of multiple tables; however, standard data mining techniques usually are only applicable to a single table.

Despite the success of Bayesian networks in a wide variety of real-world and research applications, it is hard to use them to model domains where we encounter several entities in different configurations [1], for they lack the concept of objects and relations.

Probabilistic Relational Models (PRM) are an extension of Bayesian Networks, introducing the concepts of objects and its properties, and the relations held between them, specifying a template for a probability distribution [2]. Thus, PRMs offer a rich relational structure, allowing a property of an object to depend on properties of the object itself and on properties of related objects. It can be said that PRMs are to Bayesian networks as relational logic is to propositional logic.

However, learning a PRM from relational data is a more complex task than learning a Bayesian Network from "flat" data. And, given the complications often faced at this task, a novel method for PRM learning is proposed and applied in a real large-scale problem.

2. THE PRM FRAMEWORK

A relational domain is usually represented by distinct tables (classes) in a database containing attributes and entities.

The vocabulary of a relational model consists of a set of *classes* $X_1, \dots, X_n$, and a set of *relations* $R_1, \dots, R_m$. Every class in the domain has a set of attributes $\mathcal{A}(X_i)$, and every attribute $A_j \in \mathcal{A}(X_i)$ has a space of possible values $V(A_j)$. One attribute A of a class X, is referred to as $X.A$, being that this vocabulary defines a *schema* for the relational model [3].

The logical description of the domain is called *relational schema*, and it shows how different classes relate to each other, through what is called *reference slots*.

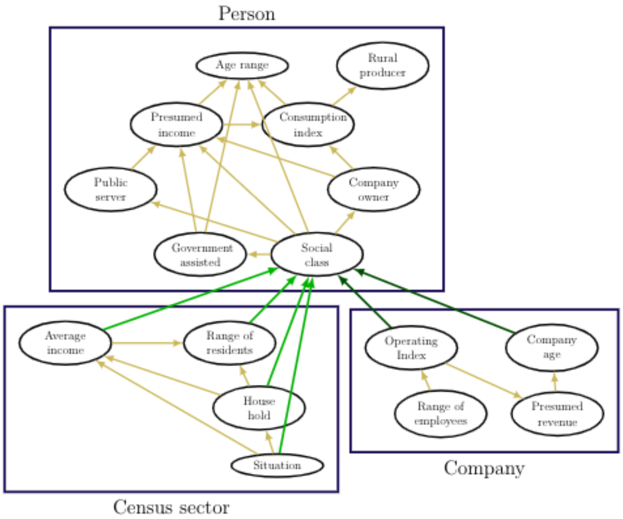

Figure 1. A PRM learned using the proposed method. The yellow arrows represent intra-class relations, and the green arrows represent inter-class relations.

The *relational skeleton* of a domain σ, defined as a partial specification of an instance of a schema [3]. It specifies for a set of *objects* $O^{\sigma}(X_i)$, a class, the value of the fixed attributes within this objects, and the relations held between them.

A PRM models the uncertainty over the possible values of the attributes of the skeleton. The model itself, consists of a dependency structure $\mathcal{S}$, and the conditional probability distribution θ_S associated with the dependency structure. Just like a Bayesian network, the dependency structure of a PRM is defined by associating a set of parents $pa(X.A)$ with each attribute $X.A$. However, in a PRM, an attribute $X.A$ can have as parent, either an intra-class attribute, denoted $X.B$, or an inter-class attribute, denoted as $X.\tau.B$, where τ is a *slot chain* representing the set of objects that are τ-*relatives* of an object $x \in X$. A simple PRM of our domain of study can be seen in Figure 1.

Aggregation functions are another important concept on PRMs, for they allow the representation of complex dependencies within the domain.

3. VANILLA-PRM: A NOVEL METHOD

The three main difficulties that arises while learning a PRM are: establishing what are the legal dependency structures for a given domain (we must avoid cycles in the structure); defining how to score the possible legal structures; and search for possible structures [1].

In face of these three challenges, we propose a method for searching for a PRM structure using scores

Learning Probabilistic Relational Models: A Novel Approach
†Luiz H. Mormille, Graduate School, Dept. of Information Systems Sciences, Faculty of Eng, Soka University, Japan.
‡Fabio G. Cozman, Full Professor at the Engineering School (Escola Politécnica) at Universidade de São Paulo, Brazil.

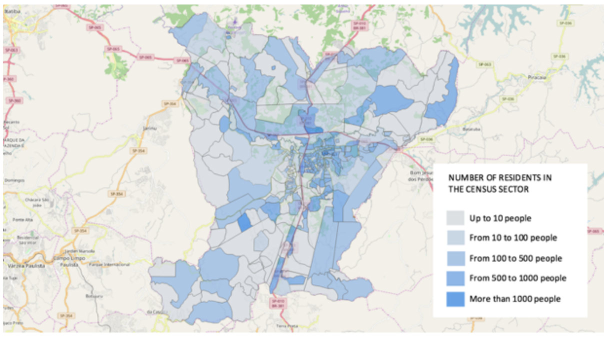

Figure 2. All census sectors in *Atibaia* plotted in a map. The color of each sector represents the range of residents inside the sector [4].

typically used to learn Bayesian networks, which are easier to implement.

In Bayesian networks and in PRMs, an attribute cannot be an ancestor of itself, that is, the probabilistic dependencies must be acyclic. For that, two graphs are considered: the *instance dependency graph* G_σ, representing the dependencies between attributes of objects in the skeleton; and the *class dependency graph* G_π, considering potential dependencies on a class level.

Introduced in Ref.[4], our method has evolved and is currently based on four distinct assumptions that ensure acyclicity, thus resulting in a legal PRM structure.

The first assumption of the method, that restricts the space of possible structures, is that the dependency structure $\mathcal{S}$ is not allowed to have cycles, not only at an attribute level, on G_σ, but at class level as well, on G_π.

Assumption 1: the *class dependency graph* G_π is not allowed to have cycles.

A second assumption is that in the G_σ cycles at a class level are not allowed to occur, even if they ultimately do not result in a cycle at an attribute level.

Assumption 2: the *instance dependency graph* G_σ is not allowed to have cycles at a class level.

A third assumption is that, in practical cases, when implementing a PRM over a domain, one might have a subset of *attributes of* interest for prediction, that usually belongs to a single distinguished class, that is referred to as the *leaf class*.

Assumption 3: given the attribute(s) of interest, the user must specify the *leaf class* beforehand.

The fourth assumption states that attributes in the leaf class cannot parent any attribute from a distinct class.

Assumption 4: attributes on the leaf class are not allowed to parent attributes from distinct classes.

After the structure is learned, the joint distribution over these assignments can be factored by taking the product, over all $x.A$ of the probability in the conditional probability distribution of the specific value assigned by the instance to the attribute given the values assigned to its parents [5]. The formal expression can be written as follows:

$$P(\mathcal{I}|\sigma, \mathcal{S}, \theta_{\mathcal{S}}) = \prod_{X_i} \prod_{A \in \mathcal{A}(x)} \prod_{x \in \sigma(X_i)} P(\mathcal{I}_{x.A}|\mathcal{I}_{Pa(x.A)}).$$

4. THE DOMAIN OF STUDY

The domain of the case study is a relatively small town named Atibaia, in the state of São Paulo, Brazil. For this study, three classes of objects were considered, representing the *citizens* of the city, the *business* located in Atibaia, and aspects of the *census sectors* that form the city territory.

Therefore, the database of the domain was comprised of three tables: the first table represented the inhabitants of the city, with 110823 objects and 23 attributes; the second table represented the local business in the city, with 20162 objects and 8 attributes; and the third table represents the *census sectors* (the smaller territorial unit considered in the demographic census), with 327 objects and 16 variables, as shown in Figure 2.

5. EXPERIMENTS

The goal of the experiment was to learn a PRM structure to predict the *presumed revenue attribute* of a business located in Atibaia. For the same domain, two PRMs were learned. The first used the proposed method, and we denote it Π_c; The second PRM was leaned using the method proposed by Koller and Friedman in Ref. [1] and [3], and we denote it Π_g.

Both Π_c and Π_g were used to predict the target variable in a *10-fold cross validation*. The overall accuracy of Π_c was 0.8241, while the overall accuracy of Π_g was 0.7701.

6. CONCLUSION

To lessen the difficulties that usually arise when learning a PRM structure, we proposed a novel method based on four assumptions. Even though our method could result in the restriction of good candidate structures, when applied to a real large-scale problem it outperformed the current state of the art method introduced by Koller and Friedman in Ref. [1] and [3].

REFERENCES

[1] KOLLER, D. Probabilistic relational models. In: SPRINGER. International Conference on Inductive Logic Programming. 1999. p. 3–13.

[2] KOLLER, D.; FRIEDMAN, N. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[3] FRIEDMAN, N. et al. Learning probabilistic relational models. In: IJCAI. 1999. v. 99, p. 1300–1309.

[4] MORMILLE, L.; COZMAN, F. Learning Probabilistic Relational Models: A Simplified Framework, a Case Study, and a Package. In: Symposium on Knowledge Discovery, Mining and Learning, KDMILE. 2017.

[5] GETOOR, L. Multi-relational data mining using probabilistic relational models: research summary. In Proceedings of the First Workshop in Multi-relational Data Mining. 2001.