Chapter 1

Introduction

Contents

Statistical Relational Learning (SRL) has proved successful to learn efficiently from the large amount of interlinked information available, for instance on the Web. SRL approaches are capable to deal with the inherent noise of large heterogeneous relational datasets, which includes partial inconsistencies, ambiguities, or duplicate entities. Factorization models have proven to be powerful models for relational learning, providing highly competitive prediction performance while being able to scale to large dataset sizes. However, new paradigms are still needed towards statistical and computational inference based on relational data.

1.1 Multi-relational learning

Statistical machine learning models (Hastie *et al.*, 2009) assume that data points in a dataset are all sampled independently from each other but from the same distribution, which is known as the independent and identically distributed (iid) assumption and that data instances are represented as points in a high-dimensional space. This largely simplifies statistical inference and has enabled many practical

applications of machine learning models However. One of its implications is that the models are able to exploit attribute information about the instances but ignores any information about the relationships between them. However many real world datasets contain rich relational structure and knowing how different data points are related, does provide useful information about them. For instance, when predicting the topic of a Web document, it is useful to know the topics of the documents it is connected to through hyperlinks; also, in a social network environment, the interests of the friends of a given user, are a good indicator of his/her own interests. To see how this is important, take as an example the task of predicting the blood type of a person. Knowing the blood type of a person x does not provide a priori any information about the blood type of another person y. However if it is known that x is the father of y, then the blood type of x does provide some indication on the likely blood type of y. In statistical terms, let P be a probability distribution, $B(x) \in \{a, b, o\}$ a variable denoting the blood type of x and $F(x, y) \in \{0, 1\}$ a variable denoting whether x is the father of y. Then, for a typical iid model,

$$P(B(y)|B(x)) = P(B(y))$$

since the model assumes that $B(y)$ and $B(x)$ are independently sampled.

A relational model is a model capable of considering relationships between instances like the `father-of` relation denoted by the variable $F(x, y)$. Since the blood type is a genetic characteristic, knowing that $F(x, y) = 1$ means that $B(y)$ and $B(x)$ are not independent anymore:

$$P(B(y)|B(x), F(x, y) = 1) \neq P(B(y)) \ .$$

This of course models the data in a better way. The independence assumption of iid models makes it easier to compute the joint distribution $P(B(x), B(y))$:

$$P(B(x), B(y)|F(x, y)) = P(B(x)|F(x, y))P(B(y)|F(x, y)) = P(B(x))P(B(y))$$

whereas for the relational model the joint distribution is given by:

$$P(B(x), B(y), F(x, y)) = P(B(y)|B(x), F(x, y))P(B(x)|F(x, y))P(F(x, y)) \ .$$

Note that instead of simply multiplying the marginals $P(B(x))$ and $P(B(y))$, the conditionals have to be defined and computed. This difference might not look big on this toy example but for moderate scale datasets it might render such distribution infeasible to represent and compute. The challenge is to design a

model that compactly represents information like $F(x, y)$ for a large dataset. Also the conditionals $P(B(y)|B(x), F(x, y))$ and $P(B(x)|F(x, y))$ need to be defined, as well as scalable strategies to learn their parameters and make inference about them. This is sensible in real world datasets because a single entity instance can be related to a number of different instances through a variety of relationships. To have an idea of how crucial this is for machine learning models let us take a look at a simple likelihood function of a parameter vector Θ given a dataset $\mathcal{D}$. The likelihood can be written as:

$$\mathcal{L}(\Theta|\mathcal{D}) = P(\mathcal{D}|\Theta)P(\Theta) = \left(\prod_{d\in\mathcal{D}} P(d|\Theta)\right) P(\Theta) \ .$$

The probability $P(\mathcal{D}|\Theta)$ could only be simplified in the product above because of the iid assumption on the data points $d \in \mathcal{D}$. Assume now that we have a non-iid dataset $\mathcal{D} = \{d_1, d_2, d_3\}$, then $P(\mathcal{D}|\Theta)$ is now:

$$P(\mathcal{D}|\Theta) = P(d_1|d_2, d_3, \Theta)P(d_2|d_3, \Theta)P(d_3|\Theta)$$

One can easily see that, for a small dataset it is infeasible to compute the likelihood. From this discussion, two basic problems arise when dealing with relational data: (i) how to compactly represent a joint distribution of model and data with relational information and (ii) develop models for which inference is feasible.

The first problem was originally approached by using first order logic as a representation formalism (Muggleton & De Raedt, 1994). The Inductive Logic Programming (Muggleton & De Raedt, 1994) is a general approach to learn first order logic inference rules from relational databases.

While this approach has the advantage that the models are easily interpretable and understandable by humans, it has the drawback that logic based methods are not able to deal with incomplete or noisy data. This seriously limits their application to real world problems. In order to overcome this problem, Statistical Relational Learning (SRL) models (Friedman *et al.*, 1999; Getoor & Taskar, 2007; Kersting, 2006; Neville *et al.*, 2003) combine knowledge representation formalisms like first-order logic with probabilistic graphical models. A number SRL models have been proposed like the Bayesian Logic Programs (Kersting, 2006) and the Markov Logic Networks (Richardson & Domingos, 2006). At the same time, non-parametric bayesian approaches like IRM (Kemp *et al.*, 2006) and IHRM (Xu *et al.*, 2006) have been proposed for relational learning.

1.1.1 Factorization models for Multi-relational data

Although powerful, general and expressive, SRL models still suffer from scalability issues. Recently, multi-relational factorization models have shown to scale well while providing good predictive performance and are currently considered as the state-of-the-art for SRL tasks (Jenatton *et al.*, 2012; Nickel *et al.*, 2011; Singh & Gordon, 2008b). Factorization models for multi-relational data associate entities and relations with latent feature vectors and define predictions about new relationships through operations on these vectors (e.g., dot products). Nickel *et al.* (2011) showed that these models are strongly competitive against MLNs while they have much better scalability. Singh & Gordon (2010) provide some insight about why factorization models work well with relational data. Basically, a factorization model assumes that data points are a priori related and that they are only independent given the latent features. Using the bayes theorem this results in a model which considers relationships between entities in the data but can also be learned using the machinery developed for models under the iid assumption. To be more clear, in the blood type example, a factorization model would assign latent features $\varphi(x), \varphi(y)$ to x and y respectively. Given the latent features the blood types of both instances are independent:

$$P(B(y)|B(x), \varphi(y), F(x, y)) = P(B(y)|\varphi(y))$$

The latent features φ can be easily computed, for instance by maximum likelihood estimators based on the relational data. Since the data points are independent given the latent features, it is true that $P(\mathcal{D}|\varphi) = \prod_{d \in \mathcal{D}} P(d|\varphi)$.

Although vastly studied, most of the work on factorization models for relational learning focused on how the prediction function looks like, i.e. whether to consider three or two way interactions (Jenatton *et al.*, 2012), which kind of latent features to employ, e.g. whether to use feature vectors or matrices (Nickel *et al.*, 2011) or the usage of only non-negative features (Takeuchi *et al.*, 2013), whether to use link functions (London *et al.*, 2012) and so on. Other aspects of the relational learning problem were either not considered or not yet fully investigated. For instance, most of the available relational data come only with positive observations. Since usually machine learning models need both positive and negative examples for training, this issue needs to be closely examined. Another aspect not fully addressed is the fact that in most of the multi-relational learning tasks, predictions are to be made for multiple target relations. State-of-the-art models are optimized for a loss that is the (weighted) sum of the losses on each relation. How to carefully optimize each relation individually is still an open issue. This

thesis discusses how the state-of-the-art factorization models look like, identifies open problems in the field and approach them in a principled way.

1.1.2 Applications of Multi Relational Learning

One question that can be asked is: is there enough relational information available so that it is worth to develop models capable of exploiting such data in large scales? The answer is yes. Mining multi-relational data with noise, partial inconsistencies, ambiguities, or duplicate entities, has gained relevance in the last years and found applications in a number of tasks. There is a plethora of datasets containing relational information, especially on the Web. One prominent example is the Semantic Web Semantic Web's Linked Open Data (LOD) initiative where the data consists of triples containing a predicate relating a subject and an object. Example of large LOD bases are DBpedia[1] and YAGO (Suchanek *et al.*, 2007). The task of **LOD mining** can be useful for statistically querying such databases (Drumond *et al.*, 2012) and for predicting new triples (Drumond *et al.*, 2012; Nickel *et al.*, 2012).

Another broad application area for relational learning methods are **recommender systems** (Koren *et al.*, 2009). The task of recommender systems can be seen as the prediction of a relation between users and items. Often, additional relational side information about users is available such as friendship relationship between them and about items, such as, for instance, which movies share the same director. This additional information can be exploited by multi-relational models for improving the recommendation performance (Lippert *et al.*, 2008; Singh & Gordon, 2008b) or for alleviating cold-start problems (Krohn-Grimberghe *et al.*, 2012).

Natural language processing is another field where the datasets available contain lots of relational information. For instance, relationships between words like the subject and object of a verb can be predicted using multi-relational models (Jenatton *et al.*, 2012; McCray, 2003). Other example of tasks involving relational data are **protein-interaction prediction** (Lippert *et al.*, 2008), **mining of geopolitical information** (Rummel, 1999) and **entity linking** (Shen *et al.*, 2012).

[1] `http://dbpedia.org/`

1.2 Contribution

Although a number of relational models have been proposed in the last years, there are still gaps in the state-of-the-art which need to be investigated. The main goal of this thesis is to provide a cohesive view of the state-of-the-art, identify such gaps and propose solutions to close them. Specifically, our contributions are summarized as follows:

- **Formalize the relational learning problem and study the state-of-the-art under a single notational framework**. We propose a formalization for representing multi-relational data and the multi-relational learning problem. State-of-the-art models are described under a single notational framework which makes it possible to identify redundancies (similar or equivalent models) and open problems not yet properly addressed in the literature.

- **Study the problem of learning from positive only data in the context of multi-relational models**. We investigate the impact of explicitly considering the open-world semantics of many datasets in the loss function. We argue why the evaluation protocols usually used in the literature are not suitable for evaluating models on data with only positive observations and propose a more suitable evaluation procedure. We also adapt approaches from the item recommendation community to the multi-relational learning problem and evaluate them.

- **Propose a new approach for learning models for multiple target relations**. A new factorization approach that optimizes directly for a number of target relations is proposed. We argue that the models should be optimized for the best performance on each relation individually. We show how this approach can improve state-of-the-art performance.

- **Apply multi-relational factorization models to semi-supervised binary classification**. The semi-supervised classification problem is formalized as a multi-relational learning problem using our proposed notational framework. We propose a new semi-supervised classification approach, namely PNT-CMF, a factorization model that collectively factorizes the predictor, neighborhood and target relation and devise a learning algorithm for it.

- **Empirical evaluation and analysis**. The proposed methods are evaluated using both small and large publicly available data sets. The proposed methods are compared against state-of-the-art methods. We empirically show that, in most of the cases, the proposed methods can achieve better prediction performance than their competitors and scale to large problems.

1.3 Submitted and Published Work

The contributions of this thesis were published in international conferences. The list of publications is as follows:

- Lucas Drumond, Steffen Rendle and Lars Schmidt-Thieme (2012). *Predicting RDF triples in incomplete knowledge bases with tensor factorization.* In Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC 12, 326331, Riva Del Garda, Italy.

  The content of this paper is mostly covered in Chapter 3.

- Lucas Drumond, Lars Schmidt-Thieme, Christoph Freudenthaler and Artus Krohn- Grimberghe (2014). *Collective Matrix Factorization of Predictors, Neighbor- hood and Targets for Semi-Supervised Classification.* In Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2014, 286297, Tainan, Taiwan.

  The content of this paper is covered in Chapter 5.

Also the following paper, which covers the content of Chapter 4, is under the revision process for publication:

- Lucas Drumond, Lars Schmidt-Thieme, Ernesto Diaz-Aviles. Optimizing Multi-Relational Factorization Models for Multiple Target Relations. *submitted.*

During the time of my doctoral studies I co-authored further publications that, although not covered in this thesis, are related to or have influenced the work presented here.

- Josif Grabocka, Lucas Drumond, Lars Schmidt-Thieme (2013): Supervised Dimensionality Reduction Via Nonlinear Target Estimation, in Proceedings of the 15th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2013.

- Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Lars Schmidt-Thieme (2012): Using factorization machines for student modeling, in Workshop and Poster Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization, Montreal, Canada.

- Ernesto Diaz-Aviles, Lucas Drumond, Zeno Gantner, Lars Schmidt-Thieme, Wolfgang Nejdl (2012): What is Happening Right Now ... That Interests Me? Online Topic Discovery and Recommendation in Twitter , Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012).

- Ernesto Diaz-Aviles, Lucas Drumond, Lars Schmidt-Thieme, Wolfgang Nejdl (2012): Real-Time Top-N Recommendation within Social Streams , Proceedings of the 6th ACM International Conference on Recommender Systems (RecSys'12).

- Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Lars Schmidt-Thieme (2012): Personalized Ranking for Non-Uniformly Sampled Items, Journal of Machine Learning Research Workshop and Conference Proceedings.

- Artus Krohn-Grimberghe, Lucas Drumond, Christoph Freudenthaler, Lars Schmidt-Thieme (2012): Multi-Relational Matrix Factorization using Bayesian Personalized Ranking for Social Network Data , Proceedings of the Fifth ACM International Conference on Web Search and Data Mining.

- Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Lars Schmidt-Thieme (2011): Bayesian Personalized Ranking for Non-Uniformly Sampled Items, in KDD Cup Workshop 2011, San Diego, USA.

- Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Lars Schmidt-Thieme (2011): Multi-Relational Factorization Models for Predicting Student Performance, in KDD 2011 Workshop on Knowledge Discovery in Educational Data (KDDinED 2011). Held as part of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

- Timo Reuter, Philipp Cimiano, Lucas Drumond, Krisztian Buza, Lars Schmidt-Thieme (2011): Scalable event-based clustering of social media via record linkage techniques, in Fifth International AAAI Conference on Weblogs and Social Media.

- Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Artus Krohn-Grimberghe, Alexandros Nanopoulos, Lars Schmidt-Thieme (2011): Factorization Techniques for Predicting Student Performance, to appear in Educational Recommender Systems and Technologies: Practices and Challenges (ERSAT 2011): Santos, O. C. and Boticario, J. G. (Eds.), IGI Global.

- Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Alexandros Nanopoulos, Lars Schmidt-Thieme (2011): Matrix and Tensor Factorization for Predicting Student Performance, in Proceedings of the 3rd International Conference on Computer Supported Education (CSEDU 2011). Best Student Paper Award.

- Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, Lars Schmidt-Thieme (2010): Learning Attribute-to-Feature Mappings for Cold-Start Recommendations, in Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010), Sydney, Australia.

- Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe, Lars Schmidt-Thieme (2010): Recommender System for Predicting Student Performance, in Proceedings of ACM RecSys 2010 Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010), Elsevier Computer Science Procedia, pp. 2811-2819.

1.4 Chapter Overview

The thesis is organized as follows:

- In **Chapter 2** a formalization for the relational learning problem is proposed and the state-of-the-art is discussed and rewritten under the proposed formalization. This allowed to identify similarities between various models and gaps in the current technology.

- **Chapter 3** investigates the impact of considering positive only observations on the loss function. It builds on previous work from the recommender systems literature on learning from positive only instances (Rendle *et al.*, 2009a) and further investigates this issue on LOD datasets.

- A new framework for multi-relational learning is proposed in **Chapter 4**. This chapter investigates the problem of making predictions for multiple

relations and proposes to employ a different set of parameters in the prediction function per target relation, so that the model can be optimized for the best performance on each relation individually instead of the best average performance over the target relations.

- **Chapter 5** presents an application of multi-relational factorization models to a standard machine learning problem, namely semi-supervised classification. The problem is formulated as an instance of a relational learning problem and a new semi-supervised classification model is proposed, which is based on a factorization model. Experiments on real world datasets show that the model outperforms state-of-the-art semi-supervised classifiers.

- Finally, **Chapter 6** puts all the proposed methods into context for comparison and conclusion. We also give an outlook in this area and raise some works for the future.