Using Semantic Web and Relational Learning in the Context of Risk Management

Thomas Fischer

Department of Information Systems
University of Jena
Carl-Zeiss-Straße 3, 07743 Jena, Germany
fischer.thomas@uni-jena.de

Abstract. The semantic web increasingly offers information that can be useful for decisions in the context of risk management. The concept of this thesis is to perform research on relational learning techniques that are able to improve risk management through the incorporation of background information based on description logics of the semantic web.

Key words: Semantic Web, Relational Learning, Risk Management

1 Problem Statement

Financial risk management is one prominent domain of risk managament (RM), which ensures the functioning and stability of banking and insurance systems. Naturally, a financial RM analysis should adhere as much as possible relevant information to evaluate an investment or portfolio of investments. The complex net of relations between macroeconomic factors, sectors, companies, products, services, people, geographical locations, financial statements, news etc. make this a very difficult task. Two main problems typically arise in such a situation, if one wants to perform appropriate decisions. First, one needs appropriate data, and second, one needs a quantitative methodology that utilizes a suitable underlying representational framework for the data.

Since the advent of the semantic web (SW) an increasing amount of machine readable and "understandable" meta information can be automatically retrieved from a vide variety of sources. Standards, such as RDF, RDF-S and OWL, provide a formal logical way to specify shared vocabularies that can be used in statements about resources. It is therefore interesting to analyse, which kind of quantitative strategies are suitable to utilize the increasing amount of semantic information in RM. How can risks be quantified based on available relational information from the SW? Risk measures like value at risk (VaR) or expected shortfall (ES)[15] are an important instrument to quantify risks, which are in general based on a so called loss distribution, which is estimated for instance for an investment or a portfolio of financial positions. The value of such a portfolio at time t is denoted by V_t. The value V_t of such a portfolio is a function $V_t = f(t, Z_t)$

of the time and a vector of observable risk factors Z_t (i.e. stock prices). Equation 1 defines the loss at time $t+1$ of an investment or portfolio.

$$L_{t+1} := -(V_{t+1} - V_t) \tag{1}$$

One can define the conditional loss distribution based on the available information $\mathcal{F}_t$ that are observable at time t.

$$F_{L_{t+1}|\mathcal{F}_t}(l) := P(L_{t+1} \leqslant l|\mathcal{F}_t) = P(-(V_{t+1} - V_t) \leqslant l|\mathcal{F}_t) \tag{2}$$

While in a typical setting the sigma field $\mathcal{F}_t$ is based only on the historical risk factor changes [15] (i.e. changes of the stock prices), this thesis proposes to incorporate information from the semantic web into the quantification of the loss distribution via relational learning algorithms, because the SW provides an increasing free available amount of relational information and relational learning algorithms are able to utilize the rich relational structures of the underlying description logics of the SW. The measured risks should be more objective, if more information (relations between entities) in the respective domain are considered.

2 Related Work

Relational learning builds upon the solid theoretical foundations of machine learning and knowledge representation [17]. Relational indicates that such algorithms are able to adhere different entities and relationships among them. There a two main research directions: Inductive Logic Programming (ILP) and Statistical Relational Learning (SRL).

ILP is concerned about the development of relational data mining algorithms to perform (deterministic) inductive inference based on the observations of a first-order representation of the information [7]. The propositional data mining algorithms have been upgraded to its first-order variants [7], with several application scenarios [17]. A prominent example are relational association rules and relational decision and regression trees [6]. In general these algorithms are deterministic, but they can have a probabilistic interpretation, such as relational association rules.

SRL performs research to "*represent, reason and learn in domains with complex relational and rich probabilistic structure*"[11]. There are logic and frame based algorithms, with the logical ones fitting naturally to semantic web description logics. Neville and Jensen [16] point out that in relational data sets the evidence of autocorrelation provides opportunity to improve the performance of statistical relational models, because inferences about one object can inform inferences about related objects [16], which is called collective inference. Relational dependency networks (RDNs) [16] are a relational extension of dependency networks that can represent and reason with cyclic dependencies and exploit autocorrelations. RDNs has been successfully applied to fraud detection [16].

SRL approaches such as probabilistic relational models (PRMs) [9], Bayesian logic programs (BLPs) [17] attempt to model a probability distribution over a

set of relational interpretations. PRMs and BLPs extend Bayesian networks with expressive relational representations. However, as discussed by Braz et al.[5], these solutions still perform inference mostly on propositional level, because they instantiate propositional graphical models based on a given query. Braz et al. outline that this propositional grounding can be computational expensive and therefore motivate first-order probabilistic inference, which is one of the current important research topics. Since risk management is a complex domain with large numbers of data, a simple reuse of existing tools is not suitable. Structure and parameter learning has to be adapted to the present context. Furthermore, most tools operate on first-order logic and not on description logics, which are in focus of this thesis.

In the context of the semantic web, relational learning has also gained interest. Initially the approaches had a focus on learning to create the semantic web, but there is also increasing research on learning from the semantic web [19, 20]. Research on ILP techniques in the SW is outlined in Lisi and Esposito [13]. Rettinger et al. [18] give insight into statistical relational learning in the semantic web. Several approaches also analyse clustering of data with background information [14, 12, 8].

In the context of risk management, there are a vide variety of (probabilistic) quantitative methodologies. However, we concentrate here on the Bayesian approaches in RM, because of their ability to incorporate quantitative and qualitative data as well as their ability to provide useful estimates of model parameters even when data is sparse. Bayesian networks, which are closely related to PLPs and PRMs, are often used in risk management [10, 2], in particular they have been also used to estimate the loss distribution of operational risks [1, 4, 3]. Despite their success, Bayesian networks are not able to model complex relational domains appropriately [17]. Therefore it seems to be desirable to utilize statistical relational learning to improve financial risk management, which to the best of the authors knowledge has been not reported so far.

3 Proposed Approach and Methodology

The idea is to use a kind of statistical relational model to estimate the loss distribution of an investment. This means that the different information (entities and relations), which are represented i.e. in RDF-S or OWL, are used to learn the structure and the parameters of this model, which is itself used to estimate the distribution of risk factor changes such as stock price movements. Several characteristics are particularly desirable to utilize a relational dependency network as the model. On the one hand, autocorrelation between stocks etc. in financial markets is present. RDNs have the ability to efficiently represent cycles, which facilitates reasoning with autocorrelation. On the other hand, the large amount of data in this domain requires also an efficient approach to learning. RDNs are an approximate representation of the joint distribution, which leads to significant efficiency gains.

There are different criteria of success for the proposed approach. First, it is important to utilize a suitable framework that combines probabilistic and logic in the complex domain of financial risk management. Much research has been done in other application scenarios, which resulted in different approaches such as PRMs, BLPs, RDNs etc. Suitability of such a framework depends on the type of patterns that can be covered, computational effectiveness of structure and parameter learning as well as inference. Due to this, the author will compare different statistical relational models in this domain according to the criteria.

Second, it will be important to gather appropriate data from different sources to evaluate the approach. The author will employ a quantitative experimentation of the approach, based on different real world as well as artificial data sets. In particular, artificial data sets are important, because the distribution of the random variables is known in advance, and therefore the prototype algorithms can be evaluated according to the found patterns. Furthermore, complexity of the artificial data set can be increased in a stepwise process. Real world data will be based on data from the open semantic web [1], such as geographical data, company and product information as well as other sources such as stock market data, news and financial statements [2] that can be transformed through schemas into needed logical formalisms. The thesis should demonstrate the feasibility as well as performance of the approach in comparison to the existing RM approaches on real world data.

4 Conclusion

This thesis proposes an original approach to quantitative risk management that should overcome the methodical limitations of widely used propositional learning approaches through the application and enhancement of statistical relational learning. The approach utilizes a representational framework for information based on semantic web standards, because an increasing amount of relevant information in this domain is exposed to the semantic web. The semantic web standards utilize description logic a subset of first-order logic and are therefore a suitable framework for relational learning techniques.

References

1. V. Aquaro, M. Bardoscia, R. Bellotti, A. Consiglio, F. De Carlo, and G. Ferri. A Bayesian Networks approach to Operational Risk. *Physica A: Statistical Mechanics and its Applications*, 389(15):1721–1728, 2010.
2. C.E. Bonafede and P. Giudici. Bayesian Networks for enterprise risk assessment. *Physica A: Statistical Mechanics and its Applications*, 382(1):22–28, 2007.
3. Chiara Cornalba and Paolo Giudici. Statistical models for operational risk management. *Physica A: Statistical Mechanics and its Applications*, 338(1-2):166–172, 2004.

[1] `http://linkeddata.org/`
[2] `http://www2.reuters.com/productinfo/`

4. L. Dalla Valle and P. Giudici. A bayesian approach to estimate the marginal loss distributions in operational risk management. *Computational Statistics & Data Analysis*, 52(6):3107–3127, 2008.
5. Rodrigo de Salvo Braz, Eyal Amir, and Dan Roth. Lifted First-Order Probabilistic Inference. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 11 2007.
6. Sašo Džeroski. Inductive Logic Programming in a Nutshell. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 11 2007.
7. Sašo Džeroski and Nada Lavrac, editors. *Relational Data Mining*. Springer, 1 edition, October 2001.
8. Nicola Fanizzi, Claudia d'Amato, and Floriana Esposito. Conceptual Clustering and Its Application to Concept Drift and Novelty Detection. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *ESWC 2008*, volume 5021 of *Lecture Notes in Computer Science*, pages 318–332. Springer, June 1–5, 2008.
9. Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1309. Morgan Kaufmann, 1999.
10. Jozef Gemela. Learning bayesian networks using various datasources and applications to financial analysis. *Soft Computing*, 7(5):297–303, 2003.
11. Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 11 2007.
12. Gunnar Aastrand Grimnes, Peter Edwards, and Alun D. Preece. Instance Based Clustering of Semantic Web Resources. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *ESWC 2008*, volume 5021 of *Lecture Notes in Computer Science*, pages 303–317. Springer, June 1–5, 2008.
13. Francesca A. Lisi and Floriana Esposito. An ILP Perspective on the Semantic Web. In *SWAP*, volume 166 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.
14. Alexander Maedche and Valentin Zacharias. Clustering Ontology-Based Metadata in the Semantic Web. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *PKDD*, volume 2431 of *Lecture Notes in Computer Science*, pages 348–360. Springer, August 19–23, 2002.
15. Alexander J. McNeil, Rudiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, 2005.
16. Jenniver Neville and David Jensen. Relational Dependency Networks. In Lise Getoor and Ben Taskar, editors, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 11 2007.
17. Luc De Raedt. *Logical and Relational Learning (Cognitive Technologies)*. Springer, 1 edition, October 2008.
18. Achim Rettinger, Matthias Nickles, and Volker Tresp. Statistical Relational Learning with Formal Ontologies. In *ECML/PKDD*, volume 5782 of *Lecture Notes in Computer Science*, pages 286–301. Springer, 2009.
19. Gerd Stumme, Andreas Hotho, and Bettina Berendt. Semantic Web Mining - State of the Art and Future Directions. *Journal of Web Semantics*, 4(2):124–143, 2006.
20. Volker Tresp, Markus Bundschus, Achim Rettinger, and Yi Huang. Towards Machine Learning on the Semantic Web. In *Uncertainty Reasoning for the Semantic Web I*, volume 5327 of *Lecture Notes in Computer Science*, pages 282–314. Springer, 2008.