AI Communications 00 (2005) 1–3
IOS Press

Context-based statistical relational learning

Yonghong Tian
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
E-mail: yhtian@ict.ac.cn

Abstract. The relational structure is an important source of information, which is often ignored by the traditional statistical learning methods. Thus this thesis focuses on how to explicitly exploit such relational information in statistical learning tasks so as to build more effective and more robust models. The main methodology used in the thesis is derived from context-based modeling and analysis. Several models and algorithms are investigated from different viewpoints of context, thereby demonstrating the general applicability of context-based statistical relational learning.

Keywords: Statistical relational learning, context modeling, contextual dependency networks, linkage semantic kernels

1. Introduction

Typically, statistical machine learning methods assume that data instances are independent and identically distributed (i.i.d.). However, many data sets are innately relational and heterogeneous, in which instances are related to each other via different types of relations. The relational structure is an important source of information, which is often ignored by the traditional statistical learning methods. Therefore, it is important to exploit the dependencies between related instances to improve the predictive performance of the learned models.

Interest in statistical relational learning has grown rapidly in recent years. Some works have successfully demonstrated the feasibility of a number of probabilistic models for relational data, such as probabilistic relational models [7], relational Markov networks [8] and relational dependency networks [9]. However, the real-world data sets are complex. Often, many objects have complex internal structure; links can be from an object to another object of the same topic, or they can point at objects of different topics (See Fig. 1). The latter are sometimes considered as legitimate and some other times referred to as “noise” when they do not provide useful and predictive information. Ideally, we would like the statistical relational models to be both predictive and robust, so that not only clean data sets can be reasoned about, but also the ones that consist of objects with complex internal structure or demonstrate complex link regularity. Towards this end, this thesis focuses on context-based statistical relational learning, in which not only the relational structure can be exploited explicitly to improve the predictive performance of probabilistic models, but also the context modeling and analysis approach is used to improve the robustness of the learned models on the real-world data sets.

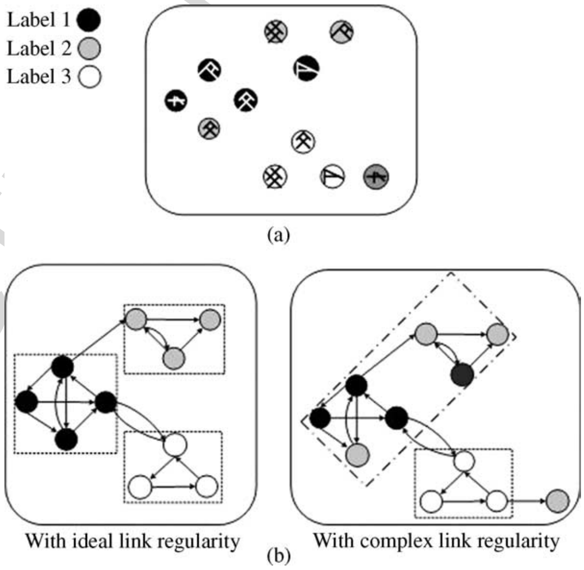

Fig. 1. Two motivated examples: (a) A data set that consists of internally structured instances; (b) Link data sets with different link regularities: ideal vs. complex. In the figure, the linked instances in a dotted rectangle have the same class label, and the linked instances in a dash-dotted rectangle may have semantically related class labels.

2. Main contributions

From different viewpoints of context modeling and analysis, several statistical relational models and learning algorithms are investigated in this thesis.

2.1. Learning with internal context modeling

Many objects may have complex internal structure. In general, the contextual dependencies among internal nodes inside each complex object can be exploited to improve the performance of the classification of that object. Consider web sites as an example of such complex objects. The thesis uses a multiscale tree as the representation model of web sites, and proposes four kinds of context models to characterize the topical correlation among nodes in the tree. Using this model, the thesis presents a two-phase classification algorithm [3] and a multiscale classification algorithm for web sites [6], both of which employ the hidden Markov tree model as the statistical model of tree-based data structure. For further improving performance while reducing the classification overheads, a two-stage denoising procedure is adopted to remove the noise information within sites, and an entropy-based strategy is introduced to dynamically prune the page trees. We performed several experiments on web site classification and spotting tasks, showing that the proposed approaches are able to offer high accuracy and efficient processing performance.

2.2. Learning with link context modeling

Links among objects contain rich semantics that can be very helpful in inference over the objects. Thus we need to investigate how to reveal such semantic information from the link structure, which is then exploited to improve the predictive performance of relational models. Two models are proposed in the thesis.

On the one hand, the thesis extends the dependency network (DN) [10] to the relational domain, and proposes a contextual dependency network model (CDN) for relational data with some noisy and irrelevant links [2]. The CDN model makes use of a dependency function that characterizes the contextual dependencies among linked objects. As a form of DNs, CDNs first approximates the full joint distribution for an entire collection of related objects with a set of conditional probability distributions (CPDs). Then each CPD can be further modeled as a linear combination of an intrinsic CPD and a set of relational CPDs with the weights represented by dependency functions. In this way, CDNs can differentiate the impacts of related objects on the classification and consequently reduce the effect of the irrelevant links on the classification. We use a self-mapping algorithm to learn the CDN model effectively, and use Gibbs inference over the learned model for collective classification of multiple linked objects. The experiments on two publicly available data sets (i.e., Cora and WebKB) show that the CDN model demonstrates relatively high robustness on data sets containing irrelevant links.

On the other hand, the thesis also proposes the linkage semantic kernels to capture the latent semantic relations among related objects from the link structure [1]. With the assumption that higher order correlation between related objects can affect their semantic relations as a diffusion process on the link graph, the thesis proposes a semantic diffusion kernel. Moreover, the eigen decomposition is directly performed in the kernel-induced space so as to obtain the kernels corresponding to the latent semantic space. For the computational efficiency on large data sets, the thesis also develops a block-based algorithm, called BlockKernel, to exploit the block structure of link data for calculating linkage semantic kernels. We performed several experiments on collective classification and relevant page finding tasks, demonstrating that linkage semantic kernels have the ability to capture the complex regularity in the link data.

2.3. Learning with multimodal context modeling

Each web image can be represented by using the visual, textual and relational features. Thus it is important to exploit the correlation among different modals of features for classifying and retrieving images. To do this, the thesis proposes a semantic image classification approach using multi-context analysis [4]. For a given image, we model the relevant textual information as its multimodal context, and regard the related images connected by hyperlinks as its link context. Two kinds of context analysis models, i.e., cross-modal correlation analysis and link-based correlation model, are used to capture the correlation among different modals of features and the topical dependency among images in the link structure. We also propose a new collective classification model called relational support vector classifier (RSVC) based on the well-known support vector machines (SVMs) and the link-based correlation model [4]. Experiments show that the web image classification models using textual/visual features often perform poorly, but with combining the three kinds of features, the proposed approach can significantly improve classification accuracy.

2.4. Learning with dynamic context modeling

User context is a type of often-used dynamic context. Online social networks, which are webs of relationships growing from computer-mediated interaction, can be viewed as a special context for users. It is crucial to understand how the networks dynamically affect the users' behaviors. Towards this end, the thesis proposes an influence model of online social networks [5]. In this model, the sequential states of each actor and their corresponding observable behaviors can be modeled as a Hidden Markov Model (HMM), and the dynamical inter-influence relationship among them can be characterized with the Influence Model [11]. To incrementally learn the model from time-series interaction data, we also present a gradient-based algorithm.

The influence model of online social networks can be explored in a wide variety of application domains, such as collaborative information filtering, collective decision-making, viral marketing plan, and so on. We also performed several experiments on collective information seeking and online viral marketing tasks. The experimental results show the influence model of online social networks can effectively capture the changing inter-influence relationship among actors (e.g., users or customers) during interactions, thereby driving the networks toward one which guides actors to more effective information seeking, or which has a higher profit potential.

3. Conclusion

The thesis addresses the problem of learning robust statistical models for the complex data sets that contain internally structured instances or externally related instances with complex link regularity. The main methodology used in the thesis is derived from context-based modeling and analysis. Based on that, several models and algorithms are presented. Although built on some specific application problems, they address different aspects of context-based statistical relational learning and can be easily extended to other settings. Moreover, the demonstrated possibilities of context-based statistical relational learning might spark a large interest into the research field since many real-world data sets are both relational and full of noise.

4. Availability

The full thesis (in Chinese) is accessible at: http://www.jdl.ac.cn/user/yhtian/publication.files/Dissertation-Tian.pdf, with the title *Research on Context-Based Statistical Relational Learning*. The degree granting institution is *Graduate School of Chinese Academy of Sciences*, and the date of defense is Jun 4, 2005.

Acknowledgements

Many thanks to my supervisors Prof. Wen Gao and Dr. Tiejun Huang for their valuable guidance and support.

References

[1] Y.H. Tian, T.J. Huang and W. Gao, Latent linkage semantic kernels for collective classification of link data, *Journal of Intelligent Information Systems* (2005) (accept for publication).

[2] Y.H. Tian, Q. Yang, T.J. Huang, C. Ling and W. Gao, Learning contextual dependency network models for link-based classification, *IEEE Transaction on Knowledge and Data Engineering* (2005) (submitted).

[3] Y.H. Tian, T.J. Huang and W. Gao, Two-phase web site classification based on hidden Markov tree models, *International Journal: Web Intelligence and Agent System* **4**(2) (2004), 249–264.

[4] Y.H. Tian, T.J. Huang and W. Gao, Exploiting multi-context analysis in semantic image classification, *Journal of Zhejiang University SCIENCE* **6A**(11) (2005), 1268–1283.

[5] Y.H. Tian, T.J. Huang and W. Gao, The influence model of online social interactions and its learning algorithms, *Chinese Journal of Computers* **28**(7) (2003), 848–858.

[6] Y.H. Tian, T.J. Huang and W. Gao, A web site representation and mining algorithm using the multiscale tree model, *Chinese Journal of Software* **15**(9) (2004), 1393–1404 (in Chinese).

[7] L. Getoor, N. Friedman, D. Koller and B. Taskar, Learning probabilistic models of relational structure, *Journal of Machine Learning Research* **3** (2002), 679–707.

[8] B. Taskar, P. Abbeel and D. Koller, Discriminative probabilistic models for relational classification, in: *Proceedings of Uncertainty on Artificial Intelligence*, 2001, pp. 485–492.

[9] J. Neville and D. Jensen, Collective classification with relational dependency networks, in: *Proceedings of 2nd Multi- Relational Data Mining Workshop in KDD2003*, 2003, pp. 77–91.

[10] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite and C. Kadie, Dependency networks for inference, collaborative filtering, and data visualization, *Journal of Machine Learning Research* **1** (2001), 49–75.

[11] C. Asavathiratham, The Influence Model: a tractable representation for the dynamics of networked Markov chains, PhD thesis, Dept. of EECS, Massachusetts Institute of Technology, Cambridge, US, 2000.