

摘 要

统计机器学习方法假设所有数据都是具有相同结构的实体，数据之间是独立且同分布的。然而在现实世界中存在着大量的半结构化关系数据，如超文本、Web 网页(网站)、Web 图像、数字图书、教育资源等，这些数据集合由不同类型的数据对象组成，数据对象本身具有复杂的内部结构，同时不同数据对象之间通过（超）链接、引用等联系起来构成关系数据集合。传统的统计学习方法忽略了数据对象间的关系结构，而这些语义信息却有助于使学习算法具有更好的性能。因此本论文研究的中心内容就是如何在统计机器学习中明确地利用数据实例间的关系信息来构建健壮的学习模型。

本论文采用的主要方法论是上下文建模与分析。在研究中，上下文被定义为使得目标对象具有唯一的和可理解的语义的关联对象和其他影响因素的集合；相应地，上下文依赖关系则是传达了明确语义相关的“关系”。论文在对上下文分析和建模、统计关系学习等两方面的研究现状进行全面综述的基础上，以不同的应用问题为背景，开展了基于上下文分析的统计关系学习方法的研究。创新和研究成果如下：

第一，提出了基于多粒度语义模型的 Web 站点挖掘方法。Web 站点可以看作是一种具有复杂结构的超文本文档。论文用多粒度树来作为站点的描述模型，同时提出四种上下文模型来刻画树中结点间的主题相关关系。在此基础上，论文采用隐 Markov 树作为树结构的统计模型，研究了两阶段分类和多粒度分类等两个 Web 站点分类算法，以期通过利用结点间的上下文依赖关系来优化分类性能。同时，还利用两阶段文本去噪程序和基于熵的页面树动态剪枝策略来减少网页下载开销并进一步提高分类准确率。实验结果表明，多粒度语义模型能有效地刻画复杂对象内部的上下文依赖关系，而相应的分类算法能在较少的时间开销内达到较高的站点分类准确率。

第二，通过扩展依赖网络模型，提出了一种上下文依赖网络模型(CDN)来刻画链接结构中的上下文主题依赖关系。在各种现实的链接关系数据（如 Web）中，噪声链接或不相关“关系”是普遍存在的。为刻画这种复杂的链接规律性，CDN 模型用链接特征和互信息来定量刻画链接对象间的上下文依赖关系，并利用一个简单但有效的上下文优化方法来优化对象的关系近邻，从而有效地减少噪声链接信息对分类过程的影响。CDN 模型具有对链接特征的选择能力，易于适应不同的内容模型，并比传统的 DN 具有更简单的参数估计。实验结果表明，CDN 模型在噪声数据集上具有较好的健壮性，并能为链接对象的属性提供较好的预测。

第三，提出了链接语义核来刻画链接对象之间的语义关系。特别地，将链接图中的语义相关关系看作一种扩散过程，提出了一种“语义扩散核”，并在核空间利用特征分解来获得潜在链接语义核。在此基础上描述了两类基于链接语义核的算法，即核化上下文依赖网络(KCDN)来进行协作分类，以及基于链接语义核的相关页发现算法。论文在

WebKB 和 CORA 上执行协作分类实验，以及在 WT10G 上执行相关页发现实验，从而验证了链接语义核的表达能力。为更有效地计算在大数据量下的链接语义核，我们还提出了一种基于块的链接语义核计算方法 BlockKernel。实验表明，BlockKernel 算法能在大数据量下具有良好的可扩展性。

第四，提出了在线社会网络的影响力模型及其增量学习算法。此模型用隐 Markov 模型(HMM)来建模交互用户的状态序列及其相应的行为，并基于影响模型(IM)理论来建模用户之间在线群体交互行为的交互动力学。为满足应用问题中增量模型学习的需要，还提出基于梯度的方法来进行模型参数的增量训练。在线社会网络的影响力模型研究可以在协作过滤、信息推荐、群体决策、在线病毒式行销等方面都有广泛的应用。

第五，基于视觉、文本、链接信息，研究并实现了基于多上下文模型的 Web 图像的语义分类系统 ConWic。在 ConWic 中，图像的相关文本建模为图像的多模态上下文，而与目标图像相链接的相关图像则建模为其链接上下文。在此基础上 ConWic 系统利用跨模态相关分析来刻画不同模态特征空间的语义相关模式，利用链接相关模型来刻画 Web 图像因链接关系而具有的语义相关关系。实验结果表明，当利用单一模态的特征信息时，Web 图像的分类效果往往不能达到较理想的要求，而综合利用视觉、文本和链接信息则有助于改进 Web 图像的分类性能。

关键词：统计关系学习、上下文模型、多粒度挖掘、上下文依赖网络、链接语义核、影响模型

Research on Context-Based Statistical Relational Learning

Tian Yonghong (Computer Application)

Supervised By Gao Wen

The vast majority of work in statistical machine learning methods has focused on “flat” data – data consisting of identically-structured entities, typically assumed to be independent and identically distributed (IID). However, many real-world datasets are innately relational: hypertext, web pages or sites, web images, scientific papers, e-books, educational resources and more. Such semi-structured relational data consist of entities of different types, where each entity is characterized by a different set of attributes and generally has complex internal structure. Entities are related to each other via different types of relations. The relational structure is an important source of semantic information, which is often ignored by the traditional statistical learning methods. Thus the paper focuses mainly 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 this paper stems from the context-based modeling and analysis. Here the context is defined as a collection of relevant objects and surrounding influences that make the semantics of an object unique and comprehensible. Accordingly, the contextual dependency can be regarded as a special relationship among related objects that conveys explicit semantic correlation. Starting with an in-depth discussion of the related work on context analysis methods and statistical relational learning, the paper investigates several statistical contextual learning methods on different application domains. The creativities and contributions are discussed in detail as follows:

First, the paper proposes a novel web site representation and mining algorithm using multiscale semantic models. In general, a web site can be regarded as a hypertext document with complex internal structure. The paper 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 multiscale site tree. Using this model, the paper presents an HMT-based two-phase classification algorithm and a multiscale classification algorithm for web sites, both of which employ the hidden Markov tree model as the statistical model of tree-based data structure, and explicitly exploit the contextual topical correlation among nodes to improve the classification accuracy of web sites. 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. The experiments demonstrate that the proposed approach is able to offer high accuracy and efficient processing performance.

Second, the paper extends the dependency network model (DN) to the relational domain, and proposes a *contextual dependency network* model (CDN). Links among objects contain rich semantics that can be very helpful in classifying the objects. However, many irrelevant links can be found in real-world link data such as Web pages. Often, these noisy and irrelevant links do not provide useful and predictive information for categorization. It is thus important to automatically identify which links are most relevant for categorization. In this paper, we present a CDN model for categorization in the presence of noisy and irrelevant links. The CDN model makes use of a dependency function that characterizes the contextual dependencies among objects and attempts to differentiate the impacts of the related objects on the classification. Using this model, it is possible to identify a context for a given object as its most relevant neighbors in a link graph, with which the semantic meaning of that object can be determined. We show how to learn the CDN model effectively, and how to use the Gibbs inference framework over the learned model for collective classification of multiple linked objects. The experiments show that the CDN model demonstrates relatively high robustness on datasets containing irrelevant links.

Third, the paper proposes the *linkage semantic kernels* to capture the latent semantic relations among linked objects that are induced by the local and global structure of the link graph. Specifically, the paper assumes that higher order correlation between indirectly connected objects can affect their semantic relations as a diffusion process on the link graph, and then 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. Based on the linkage semantic kernels, the paper also presents a *kernelized contextual dependency network* model (KCDN) to exploit the dependencies in a network of objects for collective classification, and describes a relevant page finding algorithm, *KernelRank*. The experimental results on several collective classification and relevant page finding tasks indicate that linkage semantic kernels have the ability to capture the complex regularity in the link data. For the computational efficiency on large datasets, we also develop a block-based algorithm, called *BlockKernel*, for LLSK kernels by exploiting the block structure of link data. We evaluate the BlockKernel algorithm on the whole Cora dataset, showing that this algorithm can scale well with varying sizes of the problem.

Forth, the paper proposes the influence model of online social networks and its incremental learning algorithm. 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. To incrementally learn the model from time-series interaction data, a

gradient-based algorithm is also induced. The influence model of online social networks can be explored in a wide variety of application domains, such as collaborative information filtering and recommendation, collective decision-making, viral marketing plan, and so on.

Lastly, the paper implements a context-based web image classification system, *ConWic*. The objective of *ConWic* system is to exploit the visual, textual and link information to aid classification of web images. Therefore, it models the relevant textual information of a given web image as its *multi-modal context*, and regards 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 relation among different modals of features and the topical correlation among images induced by the link structure. The experimental results show that for web images, the classification models using single modal features often perform poorly, but the combination of three kinds of features can facilitate better semantic classification of web images.

Keywords: Statistical relational learning, context models, multiscale mining, contextual dependency network models, linkage semantic kernels, influence models.