南京大学管理学学术文库

面向大数据决策的概率图模型研究与应用

李小琳 何湘东 著

概率图模型是概率理论和图论的结合。它提供了一种自然的工具，用来处理贯穿于应用数学和工程中的两个问题——不确定性和复杂性。本书围绕面向智能决策的概率图模型，从关系学习以及传统机器学习两个方面展开研究。

南京大学出版社

南京大学管理学学术文库

面向大数据决策的概率图模型研究与应用

李小琳 何湘东 著

南京大学出版社

图书在版编目(CIP)数据

面向大数据决策的概率图模型研究与应用 / 李小琳，何湘东著. — 南京 ：南京大学出版社，2017.12
（南京大学管理学学术文库）
ISBN 978-7-305-18282-2

Ⅰ. ①面… Ⅱ. ①李… Ⅲ. ①贝叶斯理论—应用—数据采集 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 030448 号

出版发行　南京大学出版社
社　　址　南京市汉口路 22 号　　邮　编　210093
出 版 人　金鑫荣

丛 书 名　南京大学管理学学术文库
书　　名　面向大数据决策的概率图模型研究与应用
著　　者　李小琳　何湘东
责任编辑　唐甜甜　　编辑热线　025-83594087

照　　排　南京南琳图文制作有限公司
印　　刷　江苏凤凰数码印务有限公司
开　　本　850×1168　1/32　印张 6.25　字数 103 千
版　　次　2017 年 12 月第 1 版　2017 年 12 月第 1 次印刷
ISBN 978-7-305-18282-2
定　　价　69.80 元

网址：http://www.njupco.com
官方微博：http://weibo.com/njupco
官方微信号：njupress
销售咨询热线：(025) 83594756

本著作的出版得到了
国家自然科学基金（60803055，61773199）
资　助

前　言

随着信息技术的飞速发展，各种信息正源源不断被数字化。对数据资源的有效利用无疑将会优化整合社会资源，加深对社会现象的认识，促进科学的发展与进步，提高人们的生活质量，甚至对人们的生活方式产生革命性的影响。从知识的角度而言，对数据资源的有效利用包含两个方面：一是归纳整合已有知识，二是探索发现新知识。

概率图模型是概率理论和图论的结合。它提供了一种自然的工具，可以用来处理贯穿于应用数学和工程中的两个问题——不确定性和复杂性。而概率图模型中的贝叶斯网络以其独特的不确定性知识表达形式、丰富的概率表达能力、综合先验知识的增量学习特性，成为当前数据挖掘众多方法中最为引人注目的焦点之一。2012年，有“计算机界诺贝尔奖”之称的图灵奖颁发给 Judea

Pearl，以奖励他在图模型学习方法方面做出的重要贡献。

图模型结构学习是一个富有挑战性的问题，其主要困难在于如何从众多可能的结构中寻找到最适合的依赖结构。对于贝叶斯网络来说，找到具有最高打分的网络结构是 NP 问题。而在关系学习中，概率关系模型(probabilistic relational models，PRMs)学习的难度至少等同于贝叶斯网络学习。本书围绕面向智能决策的概率图模型的研究与应用，从关系学习以及传统机器学习两个方面，对若干问题进行了研究，并对其在真实世界中的应用进行讨论。

本书可以看作是作者多年来在图模型学习领域研究工作的总结，书中有关研究成果是在国家自然科学基金(60803055，61773199)的支持下完成。但科学发展极其迅速，笔者自认才疏学浅，书中错谬之处在所难免，欢迎诸位读者不吝告知，将不胜感激。

李小琳　何湘东

2017 年 11 月

图目录

图 1-1 KDD 处理过程示意图 …… 4
图 2-1 变量之间的依赖关系 …… 19
图 2-2 结点之间的 d-separation 情况 …… 22
图 2-3 贝叶斯网络小球模型 …… 23
图 2-4 变量之间的基本依赖关系 …… 26
图 3-1 通信系统框图 …… 40
图 3-2 Chest-Clinic 网结构图 …… 56
图 3-3 用 1 000 个样本的数据集定向后的 Chest-Clinic 网 …… 58
图 3-4 ALARM 网结构图 …… 59
图 4-1 进化过程中 *mean* 的变化趋势 …… 71
图 4-2 进化过程中 *nstd* 的变化趋势 …… 71
图 4-3 个体编码方案 …… 77
图 4-4 Niche-EP 与 GA 学习 ALARM 网的 MDL 打分结果 …… 81
图 4-5 Restart-EP 与 GA 学习 ALARM 网的 MDL 打分结果 …… 85

图 5－1 Binary PSO 与 GA 学习 ALARM 网的 MDL 打分结果 …… 106
图 5－2 基本免疫算法流程图 …… 113
图 5－3 IB-PSO、Binary PSO 与 GA 学习 ALARM 网的 MDL 打分结果 …… 119
图 6－1 School 模型结构 …… 134
图 6－2 不完备数据学习 MLTEC 实验结果分析 …… 137
图 6－3 MLTEC 和 FR 方法学习到的 movie 的结构 …… 138
图 6－4 MLTEC 和 FR 方法学习到的 financial 的结构 …… 140
图 7－1 提高人口科学文化素质决策的贝叶斯网络拓扑结构 …… 158
图 7－2 提高人口健康素质决策的贝叶斯网络拓扑结构 …… 160

目　录

第 1 章　绪　论 …… 1

1.1　知识发现 …… 2

1.2　图形模式的概念及发展概述 …… 6

1.3　贝叶斯网络的应用 …… 9

1.4　研究框架 …… 14

第 2 章　贝叶斯网络概述 …… 17

2.1　贝叶斯网络基础理论 …… 18

2.2　贝叶斯网络学习方法 …… 29

2.3　贝叶斯网络学习算法的准确性评价方法 …… 35

第 3 章　贝叶斯网络弧定向方法研究 …… 38

3.1　贝叶斯网络弧定向方法介绍 …… 39

3.2　信息论的基本概念 …… 39

3.3　CE-GA 算法 …… 47

3.4 实验分析 …… 55
3.5 本章小结 …… 60

第 4 章 基于进化计算的贝叶斯网络学习 …… 62
4.1 进化计算 …… 64
4.2 进化规划 …… 65
4.3 进化规划中早熟收敛原因分析及刻画早熟收敛的两个量 …… 67
4.4 进化规划中防治早熟收敛现象的两种方法 …… 72
4.5 基于 Niche-EP 的贝叶斯网络学习 …… 76
4.6 基于 Restart-EP 的贝叶斯网络学习 …… 82
4.7 Niche-EP 与 Restart-EP 的比较分析 …… 86
4.8 本章小结 …… 87

第 5 章 基于粒子群算法的贝叶斯网络学习 …… 89
5.1 基本粒子群算法 …… 90
5.2 粒子群算法与遗传算法的比较 …… 95
5.3 粒子群算法在实际中的应用 …… 99
5.4 离散粒子群算法 …… 101
5.5 基于离散粒子群算法的贝叶斯网络结构学习 …… 102
5.6 基于改进离散粒子群算法的贝叶斯网络结构

学习 …… 107
5.7 Niche-EP、Restart-EP 及离散 PSO 学习贝叶斯网络结构比较 …… 120
5.8 本章小结 …… 121
第 6 章 一种从不完备关系数据中学习 PRM 的方法 …… 123
6.1 背景知识 …… 125
6.2 MLTEC 算法 …… 129
6.3 实验测试 …… 134
6.4 本章小结 …… 141
第 7 章 贝叶斯网络在人口预测与决策中的应用 …… 142
7.1 人口决策分析的基本过程和方法 …… 143
7.2 人口决策系统综合分析 …… 146
7.3 基于贝叶斯网络的人口决策系统指标体系 …… 148
7.4 基于贝叶斯网络的人口决策系统预测模型 …… 156
7.5 本章小结 …… 162
参考文献 …… 164

第1章 绪 论

计算机与信息技术经历了半个世纪的发展，给人类社会带来了巨大的变化与影响。在支配人类社会三大要素（能源、材料、信息）中，信息愈来愈显示出其重要性和支配力，它将人类社会由工业化时代推向信息化时代。随着人类活动范围的扩展，生活节奏的加快，以及技术的进步，人们能以更快速、更容易、更廉价的方式获取和存储数据，这就使得数据及其信息量以指数方式增长。然而，人类的各项活动都是基于人类的智慧和知识，即对外部世界的观察和了解，做出正确的判断和决策以及采取正确的行动，而数据仅仅是人们用各种工具和手段观察外部世界所得到的原始材料，它本身没有太大意义。从数据到知识再到智慧，需要经过分析、加工、处理、精炼的过程。因此，如何对数据与信息快速有效地进行分析、加工、提炼以获取所需知识，就成为计算机及信息技术领域

的重要研究课题。所有这些促进了知识发现(knowledge discovery)这一领域的形成和发展[1,2,3,4,5]。贝叶斯网络是用来表示变量间连接概率的图形模式,它提供了一种自然的表示因果信息的方法,用来发现数据间的潜在关系。本书研究基于贝叶斯网络的知识表示、知识获取和推理,为解决实际问题提供理论依据、方法和算法,并以提出的算法为工具,对长江地区人口数据库进行数据挖掘和知识发现,为人口问题及可持续发展提供决策支持。

1.1 知识发现

知识发现是从数据集中抽取和精化新的模式。知识发现的范围非常广泛,可以是经济、工业、农业、军事、社会、商业、科学的数据或卫星观测得到的数据。数据的形态有数字、符号、图形、图像、声音等。数据组织方式也各不相同,可以是有结构、半结构或非结构的。知识发现的结果可以表示成各种形式,包括规则、法则、科学规律、方程或概念网等。

目前,关系型数据库应用广泛,并且具有统一的组织结构,一体化的查询语言,关系之间及属性之间具有平等性等优点。因此,数据库知识发现(knowledge discovery

in databases, KDD)的研究非常活跃。该术语于1996年出现,Fayyad定义为"KDD是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的非平凡过程"[6]。在上面的定义中,涉及几个需要进一步解释的概念:"数据集""模式""过程""有效性""新颖性""潜在有用的"和"最终可理解性"。数据集是一组事实F(如关系数据库中的记录)。模式是一个用语言L来表示的一个表达式E,它可用来描述数据集F的某个子集F_E,E作为一个模式要求它比对数据子集F_E的枚举要简单(所用的描述信息量要少)。过程在KDD中通常指多阶段的处理,涉及数据准备、模式搜索、知识评价以及反复的修改求精;该过程要求是非平凡的,意思是要有一定程度的智能性、自动性(仅仅给出所有数据的总和不能算作一个发现过程)。有效性是指发现的模式对于新的数据仍保持有一定的可信度。新颖性要求发现的模式应该是新的。潜在有用性是指发现的知识将来有实际效用,如用于决策支持系统里可提高经济效益。最终可理解性要求发现的模式能被用户理解,目前它主要是体现在简洁性上。有效性、新颖性、潜在有用性和最终可理解性综合在一起被称为兴趣性。

整个KDD过程是由若干步骤组成,而数据挖掘(data mining, DM)仅是其中的一个主要步骤。整个

KDD的过程可粗略地理解为三部曲：数据准备(data preparation)、数据挖掘以及结果解释和评价(interpretation and evaluation)(见图1-1)。

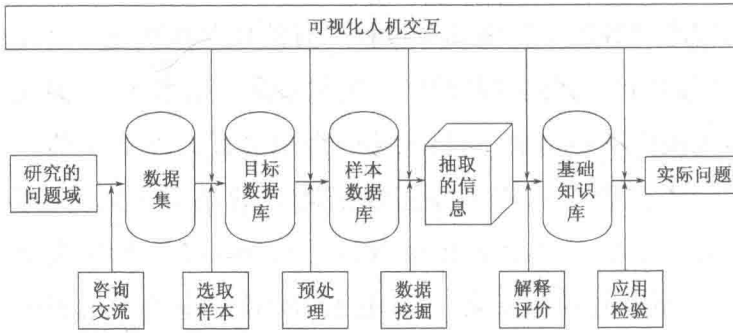

图1-1 KDD处理过程示意图

1. 数据准备

数据准备又可分为三个子步骤：数据选取(data selection)、数据预处理(data preprocessing)和数据变换(data transformation)。数据选取的目的是确定发现任务的操作对象，即目标数据(target data)，它是根据用户的需要从原始数据库中抽取的一组数据。数据预处理一般可能包括消除噪声、推导计算缺省数据、消除重复记录、完成数据类型转换(如把连续型的数据转换为离散型的数据，以便于符号归纳，或是把离散型的转换为连续型

的，以便于神经网络归纳）等。当数据开采的对象是数据仓库时，一般来说，数据预处理已经在生成数据仓库时完成了。数据变换的主要目的是消减数据维数或降维（dimension reduction），即从初始特征中找出真正有用的特征以减少数据开采时要考虑的特征或变量个数。

2. 数据挖掘

数据挖掘阶段首先要确定开采的任务或目的是什么，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了开采任务后，就要决定使用什么样的开采算法。同样的任务可以用不同的算法来实现，选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来开采；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的（descriptive）、容易理解的知识（采用规则表示的开采方法显然要好于神经网络之类的方法），而有的用户或系统的目的是获取预测准确度尽可能高的预测型（predictive）知识。

完成上述准备工作后，就可以运用选定的算法，从数据中提取出用户所需要的知识了。

3. 结果解释和评价

数据挖掘阶段发现出来的模式，经过用户或机器的评价，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要返回到前面处理步骤中的某些步骤，如重新选取数据、采用新的数据变换方法、设定新的数据挖掘参数值，甚至换一种采掘算法（如当发现任务是分类时，有多种分类方法，不同的方法对不同的数据有不同的效果）。另外，KDD 由于最终是面向人类用户的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的另一种表示。

1.2 图形模式的概念及发展概述

图形模式是概率理论和图论的结合。它提供了一种自然的工具，可以用来处理贯穿于应用数学和工程中的两个问题——不确定性和复杂性，尤其在机器学习算法的设计和分析方面扮演着越来越重要的角色。模块化的概念对于图形模式是很重要的——一个复杂系统是由每个简单部分组成的。概率理论提供了各个部分联合起来的黏合剂，保证系统作为整体是一致的，并提供模型到数