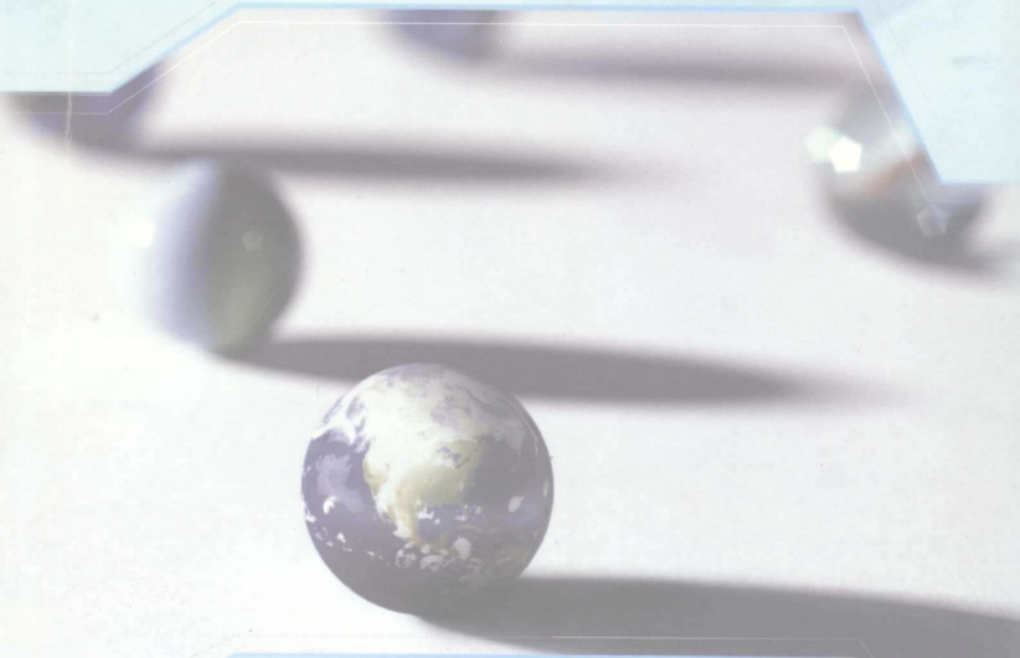
管理科学博士论丛

贝叶斯分类方法研究

石洪波 著



中国科学技术出版社

内容提要：贝叶斯网络是基于贝叶斯定理的概率统计方法，是表示和处理不确定知识的理想模型。作为分类知识发现的一种重要方法，贝叶斯分类是贝叶斯学习研究的关键问题之一。本书从限定贝叶斯网络结构规模的角度出发，对贝叶斯网络分类理论及其应用进行了深入研究，以期提高贝叶斯分类方法的分类性能，扩展其应用领域。

本书包含了作者的部分科研成果，对从事知识发现、机器学习、人工智能、计算机科学等研究的科技人员具有重要的参考价值。

ISBN 7-5046-4009-3

9 787504 640093 >

ISBN 7-5046-4009-3

O·93 定价：16.00元

管理科学博士论丛

贝叶斯分类方法研究

石洪波　著

中国科学技术出版社

·北　京·

图书在版编目（CIP）数据

贝叶斯分类方法研究 / 石洪波著. —北京：
中国科学技术出版社，2005.3
（管理科学博士论丛）
ISBN 7-5046-4009-3

Ⅰ.贝… Ⅱ.石… Ⅲ.贝叶斯推断—研究
Ⅳ.0212

中国版本图书馆 CIP 数据核字（2005）第 024600 号

责任编辑： 郑洪炜
封面设计： 李瑞峰
责任校对： 林 华
责任印刷： 王 沛

中国科学技术出版社出版
北京市海淀区中关村南大街 16 号 邮政编码：100081
电话：010－62103210 传真：010－62183872
http://www.kjpbooks.com.cn
科学普及出版社发行部发行
山西荣博印业有限责任公司印刷
*
开本：850 毫米×1168 毫米 1/32 印张：5.25 字数：145 千字
2005 年 3 月第 1 版 2005 年 3 月第 1 次印刷
印数：1－1000 册 定价：16.00 元

前　言

随着因特网的普及以及计算机存储技术的迅猛发展，计算机存储的数据越来越多，如何从大量的数据中挖掘知识是当今信息科学领域的一个研究热点。贝叶斯方法是基于贝叶斯定理的概率统计方法，是表示和处理不确定知识的理想模型。作为分类知识发现的一种重要方法，贝叶斯分类是贝叶斯学习研究的关键问题之一。但是，贝叶斯网络构造算法的时间复杂度和空间复杂度很高。本书从限定贝叶斯网络结构规模的角度出发，对贝叶斯网络分类理论及其应用进行了深入研究，以期提高贝叶斯分类方法的分类性能，扩展其应用领域。

本书共分为 7 章，在论述贝叶斯网络及其学习理论基础上，重点介绍了作者的研究工作，全书组织安排如下：

第 1 章介绍了贝叶斯网络的研究现状，以及本书的主要研究内容和研究结果。

第 2 章论述贝叶斯网络的学习理论。主要包括贝叶斯学习的理论基础、贝叶斯网络的表示与学习以及贝叶斯推理技术。

第 3 章简单介绍了贝叶斯最优分类器，详细论述了朴素贝叶斯分类模型、朴素贝叶斯的最优性条件及可学习性，最后介绍了贝叶斯网络分类模型。

第 4 章研究限定性的贝叶斯网络分类器及其学习算法。分析了朴素贝叶斯、半朴素贝叶斯、TAN 等现有的限定性贝叶斯分类模型，提出了一种新的基于贝叶斯定理的限定性模型 DLBAN，并给出了该模型的学习算法，最后，将 DLBAN 分类器与朴素贝叶斯和 TAN 分类器进行了实验比较，实验结果表明，在大多数数据集上，DLBAN 分类方法具有较高的分类正确率。

第 5 章讨论贝叶斯网络的稳定性。本章从拓扑结构和分类性

能两个方面对贝叶斯网络的稳定性进行分析，并采用基于“一致”的方法和基于方差的方法，将贝叶斯网络分类器与决策树和朴素贝叶斯两种分类方法分类性能的稳定性进行实验比较，对限定和无限定贝叶斯网络的稳定性进行了研究。

第 6 章研究集成的贝叶斯网络分类器。首先介绍分类器集成方法，重点介绍两种流行的集成技术 Bagging 和 Boosting，然后，选择 TAN 模型为基分类模型，提出一种构造 TAN 的新算法 GTAN，并将由 GTAN 生成的多个 TAN 分类器用两种不同集成方法进行组合，最后将其与典型的 TAN 分类器进行了实验比较，实验结果表明，集成的 TAN 分类器显示出较高的分类精度。

第 7 章研究将贝叶斯网络分类方法用于词义排歧。形式化描述了词义排歧问题，提出了基于限定性贝叶斯多网的词义排歧方法，借助于词典资源，将多义词的上下文词语用语义代码进行表示，最后实验验证了贝叶斯网络词义排歧方法的有效性。

最后对前几章的研究工作做了总结，并对下一步研究工作进行了展望。

目　录

第1章 绪 论

1.1 研究背景

随着因特网的普及以及计算机存储技术的迅猛发展，收集数据的速度愈来愈快，存储数据的容量愈来愈大。“数据丰富，但信息贫乏”是当前信息社会面临的主要问题，如何从堆积如山的数据中获得实际领域中可利用的、有价值的信息和知识，提高商务管理、生产控制、市场分析和科学研究等的科学性和效率，成为计算机研究人员面临的具有挑战性的任务。

知识发现(Knowledge Discovery in Database, KDD)正是适应这一任务需求而提出来的，是当前数据库与人工智能领域研究的热点课题[158]，其目标是在现实世界中，针对具有量的、质的复杂形态的海量、不完全、不确定的信息源，挖掘先前未知的、具有潜在应用价值的、最终可被用户所理解的模式的非平凡提取过程。它利用人工智能、机器学习、数据库、模式识别、统计学、自然语言理解等的理论与方法，并在此基础上发展，形成了自身的一整套体系结构，产生了许多新的算法，其应用已在许多领域取得了可喜的成果。然而，尽管取得了不少研究成果，但由于数据对象的巨量性、动态性、噪声性、不完整性和稀疏性等特点，知识发现仍然面临着大量亟待解决的问题，例如，从海量数据中获取知识时，现有学习算法的有效性和可扩充性；为了从动态、海量数据中获取知识，对增量学习算法的需求；发现模式的可理解性、兴趣或价值性等。

目前，知识发现方法主要包括机器学习方法、统计学方法、

基于数学的方法等等。其中，机器学习方法可分为决策树方法、规则归纳方法、基于事例的方法、神经网络方法、遗传方法等；统计学方法包括回归分析、判别分析、关联分析、贝叶斯方法、时间序列分析等；基于数学的方法有粗糙集方法、模糊逻辑方法等。

贝叶斯方法是基于概率统计的知识发现方法，是表示和处理不确定知识的理想模型，它的最大特点是用概率表示所有形式的不确定性。与决策树、神经网络、基于事例等方法相比，贝叶斯方法具有以下特点：

（1）具有坚实的数学基础

贝叶斯理论是贝叶斯概率和经典的统计学理论相结合的结果，它给出了信任函数在数学上的计算方法，刻画了信任度与样本数据的一致性以及信任度随数据而变化的增量学习特性，长期的理论研究和实践应用，证明了其有效性和正确性。

（2）能够充分利用领域知识和样本数据信息

贝叶斯理论将先验知识和样本信息巧妙地结合在一起，既避免了只使用先验信息可能带来的主观偏见，也可避免只使用数据样本信息带来的噪音的影响，在样本数据难得或者代价昂贵时特别有用。

（3）能够方便地处理不完整数据

贝叶斯网络可以处理不完整和带噪音的数据集，它用概率测度的权重来描述数据间的相关性，从而解决了数据间的不一致性，甚至是相互独立的问题。

（4）能够描述变量间的因果关系

贝叶斯网络能够用图形的方法描述变量间的相互关系，语义

清晰、可理解性强，这有助于利用数据间的因果关系进行预测分析。

人工智能的发展，尤其是机器学习、数据挖掘等的兴起，为贝叶斯理论的发展和应用提供了更为广阔的空间，贝叶斯理论的内涵也比以前有了很大的变化。对贝叶斯方法的深入研究，无论对贝叶斯方法的发展，还是对贝叶斯方法在数据挖掘中的实际应用，都具有特别重要的意义。

1.2 贝叶斯网络的研究现状

贝叶斯方法起源于英国学者 Reverend Thomas Bayes 的论文《An essay toward solving a problem in the doctrine of chances》[4]。该文给出了著名的贝叶斯公式和一种归纳推理方法。20 世纪 30 年代形成了贝叶斯学派，50～60 年代，Robbins 等学者提出了经验贝叶斯方法和经典方法相结合的观点，发展成很有影响的贝叶斯统计学派。80 年代，随着人工智能的发展，贝叶斯网络成功地应用于专家系统，成为不确定专家知识和推理的流行方法[67]。90 年代以后，随着机器学习、数据挖掘技术的兴起，以及贝叶斯独特的不确定性知识表示能力、综合先验知识的能力、抗噪音能力等特性，贝叶斯方法成为数据挖掘和机器学习中一个重要的研究方向[69, 157]。

贝叶斯网络[21, 110]是一个带有概率注释的有向无环图，它能够表示随机变量之间的因果关系和概率关系，可以直接进行概率推理，作出最优决策。贝叶斯网络最早是由 R.Howard 和 J.Matheson 于 1981 年提出来的，J.Pearl 的《Probabilistic reasoning in intelligent systems: Networks of plausible inference》较好地描述了贝叶斯网络的理论和方法，是有关贝叶斯网络的经典著作，此后，许多有关贝叶斯网络的专著和论文相继刊出。Artificial Intelligence 和

Machine Learning 等重要杂志每年发表若干关于贝叶斯网络的论文。1995 年的 Communication of the ACM 和 1998 年的 Artificial Intelligence 分别发表了关于贝叶斯网络的专集。目前，人工智能、数据挖掘以及机器学习等方面的许多重要国际会议都开设专题对其进行讨论。

下面从贝叶斯网络的学习方法（包括参数学习和结构学习）、贝叶斯推理、贝叶斯分类以及贝叶斯网络的表示能力等方面综述贝叶斯网络的研究现状。

1.2.1 贝叶斯网络的学习

在早期的贝叶斯网络学习中，常常是通过领域专家的知识确定贝叶斯网络的结构，指定网络的分布参数。但是，当应用领域的随机变量较多时，利用领域知识，由专家人工指定网络结构和分布参数往往是非常困难的，也是不准确的。与此同时，随着计算机存储数据能力的提高，大量有潜在价值的领域数据存储在各个应用领域。因此，人们开始研究从应用领域的数据中学习贝叶斯网络模型，并且用数据更新已有领域知识所确定的局部概率分布（即先验分布），从而得到后验参数分布，然后再利用后验参数分布进行概率推理和预测。

利用领域数据，由先验信息和样本数据得到后验信息的过程，称为贝叶斯网络学习。贝叶斯网络学习分为参数学习和结构学习。

（1）贝叶斯网络参数的学习

根据领域数据是否存在丢失值，可将数据集分为两种类型：完整数据的数据集（即：没有丢失的数据）和不完整数据的数据集（即：有些变量存在数据丢失）。因此，从领域数据学习网络参数，也分完整数据和不完整数据两种情况研究。

假定已知正确的贝叶斯网络结构，从完整数据中学习网络参

数时，无论网络节点上的概率分布是离散概率表还是连续值的高斯分布，由于存在闭型解，可以采用最大似然估计技术，很容易地以与样本数据规模成比例的时间学习到[16, 129]。这是参数学习最简单的情形。

对于从不完整数据中学习网络参数的问题，由于存在隐藏变量（其值从未在数据中出现的变量）和缺值变量，一般不能简单地采用最大似然估计技术，往往需要借助近似求解方法，例如，采用梯度下降法[6, 118, 133]，从不完整数据中学习参数；将 EM 算法应用于贝叶斯网络[92]；用 Gibbs 抽样技术学习贝叶斯网络的参数[57, 58]等。尽管这些技术可以从不完整数据中学习到网络参数，但是计算开销往往比较大。

（2）贝叶斯网络结构的学习

目前，对于完整数据集，贝叶斯网络结构的学习算法分为两类：一类是使用启发式搜索方法构造一个模型，然后用计分函数评估该模型，搜索和评估过程一直进行到新模型的计分值不是明显地比前一个模型的计分值更好为止。不同的计分标准应用于这些算法，例如，贝叶斯计分方法[32, 66]，基于熵的方法[70]，最小描述长度方法[132]。第二类算法是通过分析属性变量之间的相关性来构造贝叶斯网络结构。属性之间的相关性可用某种条件独立性(CI)测试来衡量[22, 79, 130, 131]。这两类算法各有优缺点，第一类方法的时间复杂度较高[25]，并且它的启发式特性使其可能无法找出它的最优解；第二类算法通常渐近地正确，但是，如果训练数据量不是足够大，CI 测试的结果极有可能是不可靠的。

不完整数据中网络结构的学习是目前最困难的一个问题，特别是存在隐藏变量的情况下，研究人员提出许多方法和技术解决这一问题。Friedman 提出了 SEM 算法[55]，但是由于需要事先知道网络中隐藏变量的个数，实际应用不方便。Ramoni 提出学习贝叶斯网络的 BKD 算法[117]，但不能发现隐藏变量。Connolly 使用

聚类技术发现隐藏变量[29]，但只能学习树状的网络结构。Larranaga 利用进化算法学习网络结构[88]，该方法可以有效地避免陷入局部极值，但学习效果不理想。到目前为止，从不完整数据中学习网络结构仍然是一个尚未解决的问题。

1.2.2 贝叶斯网络推理

贝叶斯网络是随机变量间的概率关系的图形表示，在给定其他变量的观察值时，可以用贝叶斯网络推理出某些目标变量。贝叶斯网络早期最主要的应用就是不确定专家知识的表示以及不确定性推理，例如，医疗诊断[5, 128]、故障诊断[65]、金融市场分析[1]等。

贝叶斯网络的推理可以分为确切推理和近似推理。确切推理主要包括基于消除的方法[36]和基于连接树的方法[73, 91, 97, 123]。但是，确切推理的时间复杂度很高，Cooper 已经证明，对任意贝叶斯网络的概率的确切推理是一个 NP 难题[30]。针对确切推理较高时间复杂度的问题，许多研究人员开始研究贝叶斯网络的近似推理，例如，Monte Carlo 方法[114]通过对未观察到的变量进行随机抽样，得到近似推理。尽管理论上贝叶斯网络的近似推理也可能是 NP 难题[35]，但是，近似推理在许多实际问题中，被证明是非常有效的。

1.2.3 贝叶斯分类方法

朴素贝叶斯[46]是最早用于分类任务的贝叶斯模型，由于不现实的属性独立性假设，朴素贝叶斯分类方法起初并没有引起机器学习研究人员的重视，只是作为比较复杂分类算法的参照对象。从 20 世纪 80 年代末开始，研究人员惊奇地发现朴素贝叶分类器具有人们没有意料到的优良性能，研究人员将朴素贝叶斯与决策

树、k-最近邻、神经网络以及基于规则等方法实验比较[20, 27, 86, 108]，发现它在某些领域中表现出很好的性能。为了探究朴素贝叶斯产生较好性能的原因，Domingos 等人深入研究了朴素贝叶斯的分类机制[44]，结果发现，如果类后验概率估计值的顺序与真正类后验概率值的顺序一致，就能获得正确的分类，而与后验概率估计值的具体数值无关。

但是，当属性独立性假设改变了真正类后验概率值的排列顺序时，朴素贝叶斯的分类性能将会降低，这种情况在实际应用中并不少见。为此，许多方法和技术用于改进朴素贝叶斯分类器的性能，主要思路是如何减少属性独立性假设的负面影响，一个改进方向是选择部分属性参与分类模型的学习，例如，Langley 和 Sage 提出的选择性贝叶斯分类器[87, 126]，Pazzani 提出的采用属性联合与选择改进分类器[108]，Kohavi 和 John 提出的 Wrapper[82]，这种方法只有对包含冗余属性的数据集，才能取得较好的结果。另一个改进方向是放松朴素贝叶斯的属性独立性假设条件，例如，半朴素贝叶斯分类器[84]，TAN 分类器[54]，k-依赖关系贝叶斯分类器[119]等。TAN 分类器是目前公认的朴素贝叶斯性能改进最好的分类器之一。

采用适合的方式和有效的机制来表示和操纵属性独立性问题，是提高朴素贝叶斯分类性能最直观的解决方法。贝叶斯网络[110]恰恰提供了一种自然的表示属性之间依赖关系的方式。尽管从理论上讲，贝叶斯网络分类器应该比朴素贝叶斯分类器具有更好的分类性能，然而，如果选择了不可靠的依赖关系集，贝叶斯网络分类器的分类性能将严重受损[54]。此外，贝叶斯网络分类器的时间复杂度、空间复杂度都很高，因此，需要研究适用于高维属性以及特殊任务的贝叶斯网络分类方法。再者，在现有分类模型的分类精度难以提高的情况下，能否采用其他方法和技术提高分类性能，也值得进一步研究。

贝叶斯分类器适合处理非数值型数据，数值属性的传统处理