代　号 10701　　学　号 0504421672

分类号 TP301；F832

西安电子科技大学

硕士学位论文

题（中、英文）目 基于贝叶斯网络的数据挖掘应用研究

Study on Application of Data Mining Based on Bayesian Networks

作 者 姓 名 李艳美　指导教师姓名、职务 张卓奎 副教授

学 科 门 类 理学　学科、专业 运筹学与控制论

提交论文日期 二〇〇八年一月

创新性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：李超美　　日期：2008.1.2

关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。同时本人保证，毕业离校后结合学位论文研究课题再撰写的文章一律署名单位为西安电子科技大学。（保密的论文在解密后遵守此规定）

本人签名：李超美　　日期：2008.1.2

导师签名：张华军　　日期：2008.1.5

摘 要

常用的数据挖掘方法有许多，贝叶斯网络(Bayesian Networks ,BN)方法在数据挖掘中的应用是当前研究的热点问题，具有广阔的应用前景。数据挖掘的主要任务就是对数据进行分析处理，从而获得其中隐含的、事先未知的而又有用的知识。它的最终目的就是发现隐藏在数据内部的规律和数据之间的特征，从而服务于管理和决策。贝叶斯网络作为在上个世纪末提出的一种崭新的数据处理工具，在进行不确定性推理和知识表示等方面已经表现出它的独到之处，特别是当它与统计方法结合使用时，显示出许多关于数据处理的优势。

本文致力于贝叶斯网络在数据挖掘中的应用研究，首先介绍了贝叶斯网络相关理论，贝叶斯网络的学习是数据挖掘中非常重要的一个环节，本文比较详细的讨论了网络图结构问题，为利用贝叶斯网络解决实际问题，建立样本数据结构和依赖关系奠定了基础。其次介绍了数据挖掘的相关问题以及主流的数据挖掘算法，并分析了各类算法的优缺点。针对目前还没有一种完整的在数据挖掘中构建贝叶斯网络的算法步骤，本文探讨性地提出了一种启发式的在数据挖掘中利用样本数据构建贝叶斯网络的算法思想，该算法较好的解决了在数据挖掘中利用样本数据设计贝叶斯网络问题。最后进行了实验分析，利用本文提出的算法，建立了大学生考研模型和农户信用等级评定模型，进行了较为详细的实验，并分别与决策树方法和传统的信用评分方法进行了比较,实验结果表明本文提出的算法设计简单、方法实用、应用有效，与其它算法相比还有精度较高的特点，同时也表现出了该算法在数据挖掘方面的优势，利于实际中的管理、分析、预测和决策等。

关键词:贝叶斯网络　数据挖掘　条件独立性　互信息　信用等级评定

ABSTRACT

Many methods have been used in data mining, Bayesian networks has become a focus currently. It has the broad application prospects. The main task of data mining is data analysis and processing, gains the implicit, prior unknown and useful knowledge. His ultimate purpose is to discover the characteristic concealing between internal law and data, thereby serving the management and decision-making. Bayesian networks is a new data modeling tool proposed at the end of last century. In the uncertainty inference carried on and the knowledge expressed, it has displayed its originality. When used in conjunction with statistical techniques, Bayesian networks has several advantages for data modeling.

This paper devotes to Bayesian networks in data mining applied research. Firstly studies and summarizes the Bayesian networks correlation theories. The study of Bayesian networks is an extremely important link in data mining. This paper discusses the question of network structure, for the use of Bayesian networks to solve actual problems, the establishment of the data structure and the foundation for dependence. Secondly discusses data mining related knowledge as well as the mainstream algorithms, and analyzes each kind of algorithms good and bad points. The discussion proposes to use the data in data mining to construct a heuristic algorithm thought of Bayesian networks. It solves the algorithmic question in data mining using the sample data well. Finally utilizes this method to establish the university student to take exams for postgraduate schools model and a farmer credit scoring model, and separately carries on the comparison with the decision tree method and the traditional credit marking method. The experimentation indicates the methods proposed practicality, availability and high precision, confirms the superiority in data mining, helps our management, analysis, forecast and decision-making and so on.

Keyword: Bayesian networks data mining conditional independency mutual information credit scoring

目 录

第一章 绪论

本章介绍了论文的写作背景，并对近几年备受研究者关注的贝叶斯网络以及数据挖掘的研究现状进行了介绍，最后给出了本文的主要工作。

1.1 引言

21世纪是数字化和信息化的时代，信息技术的发展极大地推动了社会的信息化进程。同时带来的是数据收集和数据存储技术的快速进步,也使得各组织机构积累了海量数据。如何有效地利用和处理大量的数据即通过分析数据对象之间关系提取隐含在数据中的知识，成为广大信息技术工作者所关注的焦点课题。通常，由于数据量太大，无法使用传统的数据分析工具和技术处理它们。有时，即使数据集相对较小，由于数据本身的非传统特点，也不能使用传统的方法处理。在另外一些情况下，需要回答的问题不能使用已有的数据分析技术来解决,在这种社会背景下,数据挖掘技术应运而生。

数据挖掘(DM)是一门新兴的交叉学科，涉及到数据库技术、人工智能、知识工程、统计学、机器学习、优化计算和专家系统等领域。它将传统的数据分析方法与处理大量数据的复杂算法相结合。从本质上来说，数据挖掘是智能信息处理的一种过程或技术，它在对大量数据实例全面而深刻认识的基础上，通过计算、归纳和推理等环节，从中抽取普遍的、一般的和本质的现象或特征。常用的方法有决策树、遗传算法、贝叶斯网络、粗糙集、神经网络等。其中决策树的优点是可理解性，很直观，主要用于分类和归纳挖掘，但在数据量较大和数据复杂的情况下，该算法则显得力不从心；遗传算法擅长于数据聚类，在组合优化问题上具有独特的优势。粗糙集在数据挖掘中具有重要的作用，常用于处理含糊性和不确定性的问题，以及特征归纳和相关分析，运用粗糙集进行数据预处理可以提高知识发现的效率。神经网络能够对复杂问题进行预测，它在商业界得到广泛的应用，对于信贷客户识别、股票预测和证券市场分析等方面具有良好的效果。贝叶斯网络是用来表示变量集合连接概率的图模型，它提供了一种表示因果信息的方法。贝叶斯网络能综合考虑先验信息和样本数据，充分地利用专家知识和经验，进行定性分析和定量分析。将主观和客观有机地结合起来，既避免了对数据的过度拟合，又避免了主观因素可能造成的偏见。将变量之间潜在的关联性用简洁的图解模型表达出来，表达的语义直观、清晰、推理的结果和结论可信度强，便于解释和易于理解；且具有分类、聚类、预测和因果分析等功能，预测效果较好，面对大规模数据时显示出它独特的优势。

本文的内容属于应用研究，主要针对工程领域与管理科学领域中存在的一些数据挖掘问题进行研究，利用本文提出的方法，阐述贝叶斯网络的具体应用以及它在数据挖掘中的优势。

下面将对贝叶斯网络的研究现状以及数据挖掘技术的研究现状作一些介绍，最后一节介绍本文的主要工作。

1.2 贝叶斯网络的研究历史与现状

1.2.1 国内对贝叶斯网络的研究历史与现状

国内对于贝叶斯网络的研究起步较晚，但是在近几年兴起了一股研究学习的热潮。清华大学陆玉昌、林士敏等对贝叶斯网络构建、学习以及应用开展了有益的探索。林士敏等从信息熵的角度讨论了无信息先验分布的Bayesian假设的合理性，分析了贝叶斯方法的计算学习机制，同时指出合理地指派先验分布对提高学习的效率和质量具有重要的意义[50]。林士敏等通过剖析Bayesian网络的结构和建造步骤，讨论了用Bayesian方法从先验信息和样本数据进行学习，从而确定网络结构和概率分布的基本方法[49]。刘大有教授等在2001年设计了结合数学期望的适应度函数，运用遗传算法进行结构学习，大大简化了学习的复杂度，并保证算法能够向好的结构不断进化，但不足的一点是期望统计因子的计算较为复杂。中国科学院计算技术研究所的宫秀军等在2002年将简单贝叶斯方法应用于增量分类中，提出一种增量贝叶斯分类模型，并给出了增量贝叶斯推理过程，何盈捷和刘惟一提出通过发现Markov网得到等价的Bayesian网的方法，基于依赖分析的边删除算法发现Markov网，然后根据表示的联合概率函数相等，得到与其等价的Bayesian网，但选择好的无环序是研究中有待于解决的问题[11]。羌磊等提出了一种新的基于最小描述长度理论的结构学习算法，将独立性测度与预测估计相结合[51]。云南大学的张忠玉等对Bayesian网的信息熵问题进行了探讨[52]。

此外，胡兆勇和屈梁生提出了一种贝叶斯网络的近似仿真算法。由随机数发生器产生随机数，并按节点的先验概率，由赌轮对网络各个节点状态赋值，得到一个采样样本序列。当样本序列的数量足够大时，边缘统计量和条件统计量与节点的边缘概率和条件概率接近，从而得到网络的近似推理结果。仿真结果表明，该算法与精确解接近，有较好的适应性[53]。田凤占等提出了一种贝叶斯网络增量学习方法－ILBN。ILBN 将EM算法和遗传算法引入到了贝叶斯网络的增量学习过程中,用EM算法从不完整数据计算充分统计量的期望,用遗传算法进化贝叶斯网络的结构,在一定程度上缓解了确定性搜索算法的局部极值问题. 通过定义新变异算

子和扩展传统的交叉算子, ILBN 能够增量学习包含隐变量的贝叶斯网络结构，同时ILBN 改进了Friedman 等人的增量学习过程。ILBN Friedman 等人的增量学习方法存储开销相当,但在相同条件下,学到的网络更精确[3]。冀俊忠等针对I-B&B-MDL算法的不足，提出了2点改进：一是仅利用0阶和部分1阶测试确定网络候选连接图，在有效限制搜索空间的同时，减少了独立性测试及对数据库的扫描次数；二是利用互信息的启发性知识作为候选父母节点排序，加大了B&B搜索树的截断，加速了搜索过程，算法的整体时间性能有了较大改进[4]。杨莉和孙华昕等针对多目标进化算法中存在的无效进化和计算浪费，提出了一个新的贝叶斯多目标优化算法。该算法结合个体的强度值和密度值完成非劣择优，利用具有局部结构BD度量机制进行网络度量，采用树形模型构建网络结构[1]。王双成等给出了一种有效实用的贝叶斯网络结构学习方法, 不需要结点有序, 并能避免打分- 搜索方法存在的指数复杂性, 以及现有依赖分析方法存在的大量高维条件概率计算等问题[2]。

在应用领域，冀俊忠等将贝叶斯网络应用于智能教学，提出并实现了以贝叶斯网为学生模型的智能教学系统，证明了用BBN建立学生解题模型的正确性、有效性和实用性[5]。李剑川等将贝叶斯网络应用在解决复杂设备诊断中存在的不确定性和关联性问题，并以某型SINS/GPS组合导航系统的故障诊断应用实例说明了该方法的可行性[44]。臧玉卫等将贝叶斯网络应用到股指期货风险预警中[10]。冀俊忠还针对网上智能中推荐系统的个性化问题，提出了一种新的基于贝叶斯网模型的商品推荐方法，该方法是一种有效的能为不同客户产生准确而个性化的商品推荐方法[7]。周忠宝等人针对故障树分析方法在可靠性分析中的局限性,研究了贝叶斯网络在可靠性分析中的应用,结果表明,基于贝叶斯网络的建模分析方法可以得到更丰富的信息[6]。田子德通过分析超媒体系统中的不确定性因素,讨论了自适应超媒体系统中贝叶斯网络构造过程[8]。叶跃祥等融合贝叶斯网络推理技术来求解不确定多属性决策问题, 采用此方法求解不确定环境下多属性决策问题时,降低了思考的复杂程度,适用于大规模的复杂问题求解[9]。总之，贝叶斯网络已经深入到工业、金融、水利水电、国防军事、医疗等各个方面，特别是在分类、预测、智能推理诊断和可靠性评估等方面更加显示出该技术的优越性。

1.2.2 国外对贝叶斯网络的研究历史与现状

国外对贝叶斯网络的研究起源于18世纪英国牧师Tomas Bayes一篇论文《论机会学说中的一个问题》。但由于在理论和应用中出现了许多问题，贝叶斯方法没有得到普遍接受。直到20世纪50年代开始，越来越多的统计学者推崇和研究贝叶斯的观点和思想，特别的，在1955年，美国统计学家Robbins提出经验贝叶斯方法，

受到统计学界的关注和重视。在社会科学和经济活动中，贝叶斯方法得到成功的应用。60年代初，人们把贝叶斯方法广泛地应用于解决具有不确定性的决策和评估问题，进一步开拓了应用统计决策的研究。基于主观贝叶斯方法，Dube等人1976年研制出用于地质勘探方面的PROSPECTOR专家系统。美国将贝叶斯方法应用在导弹发射的可靠性评估中，极大地节省了研制和实验费用。目前，贝叶斯方法在水利水电、土地资源评价、国防军事、金融保险等各个领域得到广泛应用。

Pearl在20世纪80年代提出了基于概率论和图论的贝叶斯网络，并成功地应用于专家系统[54],由于贝叶斯网络具备严密的推理过程、清晰的语义表达和灵活的学习机制等特点，引起众多学者的极大兴趣，成为一个非常活跃的研究领域，并在学习理论、开发应用和算法研究等几个方面取得了显著的成果。基础理论方面，Pearl对贝叶斯网络推理、信息传播、网络构建进行了早期的研究[54]。美国Microsoft公司对贝叶斯网络的研究较为深入，并开发出基于Windows平台的相关软件，用于贝叶斯网络的建造和推理。在推理学习方面，研究热点主要是贝叶斯网络的推理机制、对不确定性知识的表达以及从数据中进行贝叶斯网络学习的方法。

贝叶斯网络结构学习最早的研究开始于对树－最简单的图类研究。Chow & Liu介绍了一种根据数据集恢复简单树型结构的方法[55]。如果数据集产生于一个树型结构的分布，只要给定足够的原始数据，该方法可以准确地恢复原始的树型结构。Geiger等提出了一种从数据集中发现最小边I-map的方法，其数据集的基本分布具有多树结构[56]。Wermuth & Lauritzen提出了一种构建有向图的方法，根据变量的输入顺序，对分布进行独立性检验从而构建网络，即最小I-map[57]。Spirtes等提出了不需要变量顺序即可发现有向无环图(DAG)的算法，这些算法要求学习对象的基本分布是DAG同构[23]。分布与DAG同构指DAG表达出其所有的依赖关系和独立关系。基于CI检验方法的固有问题是每个变量都需要确定n-2个变量顺序的独立关系。而且这些检验对于复杂图是不适用的，除非数据量足够大。

现阶段，贝叶斯网络的学习方法大体上分为两类，一类是基于概率统计理论;另一类是基于信息论。其中基于概率统计学的贝叶斯方法包括贝叶斯平均和最大后验概率（MAP）准则。Cooper&Herskovits最早提出将贝叶斯最大后验概率方法用于多联接的结构[58]。该方法运用贝叶斯评分寻找最大可能的网络，即运用给定网络结构数据的似然函数与结构的先验概率的乘积为评分准则。与其它贝叶斯方法一样，该方法必须假设结构空间的先验分布。然而，由于采取的先验分布是一致的，这样使得方法更加类似于ML估计。也就是都通过选定相同的先验，得到一个更加准确的网络而不受结构复杂性的影响。

为了避免结构先验限制的可行方法就是应用最小描述长度（MDL）准则。最小描述长度（Minimal description length,MDL）准则最早由Rissanen作为统计模

型的一个新准则正式提出的。运用MDL评估时，假定网络结构的先验值用结构的描述长度来代替，其最主要的原因是描述长度可以进行计算。Bouckaert等进行了MDL准则在贝叶斯网络学习中的应用研究[59]。后来运用MDL评估进行结构学习的方法得到进一步推广[60]。近年来，许多学者尝试了对不完备数据和隐藏变量的结构学习[20][26]。其中，Friedman提出了一种方法从参数学习的EM算法扩展到结构学习，称为结构化EM (Structural EM)算法[27]，大体上说，结构化EM 在结构与参数的联合空间中进行搜索。然而，由于结构空间的不连续，造成联合空间的不连续，使得EM算法可能达不到预期的效果。另外，Chickering将结构搜索空间转化为等价类空间来实现网络结构评估[33]; Gamez& Puerta将蚁群行为应用在网络结构的搜索优化[32]; Campos & Huete提出一种基于独立性准则的方法[19];许多学者尝试了用不同的方法进行结构学习。

贝叶斯网络在应用方面，也越来越广泛。如美国通用公司开发出基于贝叶斯网络的故障诊断系统等等，在医学诊断与治疗、金融投资与市场分析、智能决策与管理、故障诊断、电力系统和水资源开发等领域均取得显著的应用效果。此外，在智能推理和诊断方面,Mittel-stadt D等在1995年将其应用到集成电路的检测中。Nikovski D讨论了贝叶斯网络在医学诊断中的应用。

1.3 数据挖掘的研究历史与现状

数据挖掘理论研究出现于20世纪80年代后期。从数据库中发现知识KDD（Knowledge Discovery in Database）一词首次出现1989年8月在美国底特律举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的KDD国际研讨会已经召开了多次，规模由原来的专题讨论会发展到国际学术大会。近几年，从事数据挖掘研发的人员遍布世界80多个国家，数据挖掘的研究重点也逐渐从算法研究转向具体应用过渡，从实验室原型走向商品化阶段。注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1999年，国际上从事数据挖掘产品研发的软件公司已从1989年的几个公司，猛增为上百家公司，每年都有若干软件产品推出。IEEE的Knowledge and Data Engineering会刊率先在1993年出版了KDD技术专刊。进入21世纪后，数据挖掘技术向纵深发展，与多种学科交叉、多种技术结合的数据挖掘理论与技术开始涌现。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论，甚至到了脍炙人口的程度。

国内对数据挖掘的研究稍晚，没有形成整体力量。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北

京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究，华中理工大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及Web数据挖掘。1999年，亚太地区在北京召开的第三届PAKDD会议收到158篇论文，空前热烈。国内这两年也有相当多的数据挖掘方面的研究成果，许多学术会议上都设有专题进行学术交流。

随着DMKD研究逐步走向深入，数据挖掘和知识发现的研究已经形成了三根强大的技术支柱：数据库、人工智能和数理统计。因此，KDD大会程序委员会曾经由这三个学科的权威人物同时来任主席。目前DMKD的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现以及网上数据挖掘等。最近，Gartner Group的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。根据最近Gartner的HPC研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。”

此外，世界上比较有影响的典型数据挖掘系统有：SAS 公司的 Enterprise Miner、IBM 公司的 Intelligent Miner、SGI 公司的 SetMiner、SPSS 公司的 Clementine、Sybase 公司的 Warehouse Studio、RuleQuest Research 公司的 See5、还有 CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Quest 等。其中比较有名的网站如：http://www.datamininglab.com，该网站提供了许多数据挖掘系统和工具的性能测试报告。

目前，集中典型的数据挖掘研究是关联规则、分类（包括决策树算法、贝叶斯方法、贝叶斯网络方法、神经网络算法、遗传算法、粗糙集和模糊集算法等等）、聚类（包括基于模型的方法、基于密度的方法、基于划分的方法、基于层次的方法等等）、预测、Web 挖掘等。

数据挖掘技术是在应用需求下诞生的，也是在应用需求的推动下不断发展的。同时，数据挖掘技术的不断发展也启发和促使新的应用需求产生。数据挖掘技术在这种循环前进中得到了不断的发展和完善。现在，数据挖掘技术已经形成了较为完善的技术体系和设计模型，比如Usama M.Fayyad、Gregory Piatetsky Shapiro 等人定义的数据挖掘处理的九步模型。同时有不少的大型数据挖掘软件出现，并获得了成功的商业应用。它们中有：IBM开发的Intelligent Miner系统、Silicon Graphics Inc.(SGI)开发的MineSet系统以及DBMiner Techology公司开发的

DBMiner系统。这些商业软件所获得的成功应用，极大的推动了学术界对数据挖掘技术的研究热情和数据挖掘技术在产业界的推广。同时，微软公司于2000年推出了OLB DB For DM技术体系，着重描述了数据挖掘过程中的抽象概念，其形式类似于关系数据库中的SQL语言。它的出现在朝着将数据挖掘语言标准化迈进的重要一步，并可能成为未来数据挖掘描述语言的工业标准。

数据挖掘技术虽然已获得了长足发展，但仍然处于早期阶段，还有很多的研究难题，如数据的巨量性、动态性、噪声性、缺值和稀疏性，发现模式的可理解性、兴趣或价值性，应用系统的集成，用户的交互操作，知识的更新管理，复杂数据库的处理等等都有待人们去进行更深入的研究，这将不断的推动数据挖掘技术得到更深入的发展和广泛的应用，创造出更多的社会和经济价值。

1.4 本文的主要工作

本文致力于贝叶斯网络方法在数据挖掘中具体应用的研究，本文的主要工作有以下几点:

1. 学习和研究贝叶斯网络相关理论、网络结构学习的方法以及参数学习的方法。贝叶斯网络的学习是数据挖掘中非常重要的一个环节。贝叶斯网络学习分为结构学习和参数学习，其中结构学习是贝叶斯网络学习核心内容。结构学习和参数学习分为完备数据和不完备数据两种，针对不同情况可以采用不同的学习算法，从数据中学习网络的结构和条件概率表。

2. 提出一种在数据挖掘中构建贝叶斯网络的算法思想。首先学习了数据挖掘的相关问题以及主流的数据挖掘算法，并分析了各类算法的优缺点。针对目前还没有一种完整地在数据挖掘中构建贝叶斯网络的算法步骤，探讨性的提出一种启发式的在数据挖掘中利用样本数据构建贝叶斯网络的算法思想,该算法较好的解决了在数据挖掘中利用样本数据设计贝叶斯网络问题。

3. 研究贝叶斯网络在数据挖掘具体案例中的应用，运用此方法建立了大学生考研模型和农户信用等级评定模型并推广到个人信用等级的评定，并分别与决策树方法和传统的信用评分方法进行了比较。实验结果表明本文提出的算法设计简单，方法实用，应用有效，与其它算法相比还有精度较高的特点, 同时也表现出了该算法在数据挖掘方面的优势，利于我们实际中的管理、分析、预测和决策等。

第二章 贝叶斯网络的相关理论

本章对贝叶斯网络的相关理论进行了系统的论述与分析，并用一个简单的疾病诊断模型对贝叶斯网络的定义以及网络构成进行了介绍。结合信息论的有关知识，讨论了贝叶斯网络中重要的条件独立性研究，并学习和研究了贝叶斯网络在完备数据和不完备数据两种情况下的结构学习和参数学习方法。结构学习是利用训练样本集，尽可能结合先验知识，确定最合适的贝叶斯网络的拓扑结构;参数学习是在给定网络结构的情况下，确定贝叶斯网络中各变量的条件概率表。其中结构学习是贝叶斯网络学习的核心，有效的结构学习方法是构建最优贝叶斯网络结构的前提。

2.1 预备知识

贝叶斯网络（Bayesian Networks）是一种关于变量集合中概率性联系的图解模型，接近于概率和统计，它的理论依据是概率统计，并以图论的形式来表达和描述数据实例中的关联和因果关系。首先介绍一些相关的概念和公式。

1. 条件概率

条件概率是概率论中的一个重要而实用的概念。所考虑的是事件 A 已发生的条件下事件 B 发生的概率。

定义2.1 设 A 、 B 是两个事件，且 $P(A)>0$,称

$$P(B|A)=\frac{P(AB)}{P(A)} \qquad \text{式（2－1）}$$

为在事件 A 发生的条件下事件 B 发生的条件概率。

显然，条件概率符合概率定义中的三个条件，即

(1) 非负性：对于每一事件 B ,有 $P(B|A)\geq 0$

(2) 规范性：对于必然事件 S ，有 $P(S|A)=1$

(3) 可列可加性：设 $B_1,\ldots,B_n$ 是两两互不相容的事件,则有

$$P(\bigcup_{i=1}^{\infty}B_i|A)=\sum_{i=1}^{\infty}P(B_i|A)$$

2. 乘法定理

由条件概率的定义2.1，立即得出下述定理：

定理2.1 设A、B为两个事件，且$P(A)>0$，则有

$$P(AB)=P(B|A)P(A) \qquad \text{式（2—2）}$$

若有$P(B)>0$,则也可以定义$P(A|B)$,这时有

$$P(AB)=P(B|A)P(A)=P(A|B)P(B) \qquad \text{式（2—3）}$$

（2—2）式可以推广到多个事件的积事件的情况。一般的，设$A_1,A_2,\cdots,A_n$为n个事件，$n\geq 2$,且$P(A_1A_2\cdots A_{n-1})>0$，则有

$$P(A_1A_2\cdots A_n)=P(A_n|A_1A_2\cdots A_{n-1})P(A_{n-1}|A_1A_2\cdots A_{n-2})\cdots\times P(A_2|A_1)P(A_1) \qquad \text{式（2—4）}$$

设随机变量集合$X=\{X_1,X_2,\cdots,X_n\}$，用x_i表示X_i的取值，在实际的应用中，联合概率满足一定的条件独立性，由乘法定理可以得到联合概率的表达式：

$$P(X_1,X_2,\cdots X_n)=\prod_{i=1}^{n}P(X_i|X_1,\cdots,X_{i-1}) \qquad \text{式（2—5）}$$

式(2—5)（特别地$i=1$时，定义$X_0=X_1$）称为链式法则（chain rule）。

3. 全概率公式

定理2.2 假设试验E的样本空间为S，A为E的事件，$B_1,B_2,\cdots,B_n$为S的一个划分，即满足（i）$B_iB_j=\varnothing,i\neq j,j=1,2,\cdots,n$;(ii)$B_1\cup B_2\cup\cdots\cup B_n=S$，且$P(B_i)>0$（$i=1,2,\cdots,n$），则有

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n) \qquad \text{式（2—6）}$$

在很多实际问题中$P(A)$不容易求得，但却容易找到S的一个划分$B_1,B_2,\cdots,B_n$，且$P(B_i)$和$P(A|B_i)$或为已知，或容易求得，那么就可以根据（2—6）式求得$P(A)$。

4. 先验概率

定义2.2 设$B_1,B_2,\cdots,B_n$为样本空间S中的事件，$P(B_i)$可根据以前的数据分析得到，或根据先验知识估计获取，则称$P(B_i)$为先验概率。

注意：$P(B_i)$的值以过去的实践经验和认识为依据，在实验之前得到或已经确定。

5. 后验概率

定义2.3 设$B_1,B_2,\cdots,B_n$为样本空间S中的事件，则事件A发生的情况下，B_i发生的概率$P(B_i|A)$，可根据先验概率$P(B_i)$和观测信息重新修正和调整后得到，通

常将$P(B_i|A)$称为后验概率。

随着样本信息的不断变化，后验概率也不断地更新。前一次的后验概率将作为再次调整时的先验概率使用，从而得到新的后验概率，这是一个不断更新、反复调整的过程。

6. 贝叶斯公式

定理2.3 假设试验E的样本空间为S，A为E的事件，$B_1,B_2,\cdots,B_n$为S的一个划分，且$P(A)>0$，$P(B_i)>0$（$i=1,2,\cdots,n$），则

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)},i=1,2,\cdots,n \qquad \text{式（2—7）}$$

（2—7）式称为贝叶斯（Bayes）公式。

7. 贝叶斯概率

简单地说，贝叶斯概率是观测者对某一事件发生的信任程度(一般称为主观概率，相对而言，传统概率称为客观概率或物理概率)。观测者根据先验知识和现有的统计数据，用概率的方法来预测未知事件发生的可能性。贝叶斯概率不同于事件的客观概率。客观概率是多次重复试验，然后统计事件发生的频率。而贝叶斯概率则是利用现有的知识对未知事件的预测。记$D=\{X_1=x_1,X_2=x_2,\cdots,X_m=x_m\}$为重复m次试验所得的观测样本。其中$X$为事件变量，$x$为变量值或状态。记参数$\theta$为事件$X=x$发生的客观概率或先验概率$(\theta=p(x|\xi))$，$p(\theta|\xi)$为它的概率密度函数，其中$\xi$为观测者的先验知识。现在贝叶斯概率的计算问题可以陈述如下:已知先验概率$p(\theta|\xi)$和样本D，求第$m+1$次试验中的事件$X_{m+1}=x_{m+1}$发生的概率$p(X_{m+1}=x_{m+1}|D,\xi)$。

由全概率公式得：

$$\begin{aligned}p(X_{m+1}=x_{m+1}|D,\xi)&=\int p(X_{m+1}=x_{m+1}|\theta,D,\xi)p(\theta|D,\xi)d\theta\\&=\int\theta p(\theta|D,\xi)d\theta=E_{p(\theta|D,\xi)}(\theta)\end{aligned} \qquad \text{式（2—8）}$$

这表明，事件$X_{m+1}=x_{m+1}$发生的贝叶斯概率即为先验概率θ相对于后验概率分布$p(\theta|D,\xi)$的期望值。根据贝叶斯规则，由先验概率$p(\theta|\xi)$计算后验概率$p(\theta|D,\xi)$的公式为

$$p(\theta|D,\xi)=\frac{p(\theta|\xi)p(D|\theta,\xi)}{p(D|\xi)}=\frac{p(\theta|\xi)p(D|\theta,\xi)}{\int p(\theta|\xi)p(D|\theta,\xi)d\theta} \qquad \text{式（2—9）}$$

在先验概率θ已知的条件下，样本D中的各事件$X=x$条件独立。如果事件变

量 X 为二元分布，即事件只有发生或不发生两种情况，则

$$p(D|\theta,\xi)=\theta^h(1-\theta)^t \qquad \text{式（2－10）}$$

其中 h 为样本 D 中事件发生的次数，$h+t=m$。现设先验概率为Beta分布，即 β 分布的期望值是已知的，即：

$$\int\theta Bata(\theta|\alpha_h,\alpha_t)d\theta=\frac{\alpha_h}{\alpha} \qquad \text{式（2－11）}$$

于是，预测事件的贝叶斯概率为：

$$\begin{aligned}p(X_{m+1}=x_{m+1}|D,\xi)&=\int\theta p(\theta|D,\xi)d\theta\\&=\int\theta Beta(\theta|\alpha_h+h,\alpha_t+t)d\theta=\frac{\alpha_h+h}{\alpha+m}\end{aligned} \qquad \text{式（2－12）}$$

现在讨论事件变量 X 是多元分布的情况，即 X 有 $x^1,x^2,\cdots x^r$ 共r个可能的状态，则 θ_k 为：

$$\theta_k=p(X=x^k|\theta,\xi),k=1,2,\cdots,r \qquad \text{式（2－13）}$$

其中 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$ 为参数矢量，统计数 N_i 为样本 D 中事件 $X=x^i$ 发生的次数，$i=1,2,\cdots,r$。先假设先验概率为Dirichlet分布，即：

$$p(\theta|\xi)=Dir(\theta|\alpha_1,\alpha_2,\cdots\alpha_r)=\frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_{k-1}} \qquad \text{式（2－14）}$$

其中 $\alpha=\sum_{i=1}^{r}\alpha_k,\alpha_k>0$，$k=1,2,\cdots,r$；

其后验概率分布为：$p(\theta|D,\xi)=Dir(\theta|\alpha_1+N_1,\alpha_2+N_2,\cdots,\alpha_r+N_r)$。

给定共轭先验概率和数据集合 D，下一个预测的概率分布为：

$$\begin{aligned}p(X_{m+1}=x^k|D,\xi)&=\int\theta_k Dir(\theta|\alpha_1+N_1,\cdots,\alpha_r+N_r)d\theta\\&=\frac{\alpha_k+N_k}{\alpha+N}\end{aligned} \qquad \text{式（2－15）}$$

2.2 贝叶斯网络的描述

贝叶斯网络（BN）是描述变量之间概率关系的图形模式。它是一个有向图，其中每个节点都标注了定量的概率信息，其完整的详细描述为：

①一个随机变量组成网络节点，变量可以是离散的或是连续的；

②一个连接节点对的有向边或箭头集合，如果存在从节点 X 指向节点 Y 的有向边，则看成 X 是 Y 的父节点；

③每个节点 X_i 都有一个条件概率分布 $P(X_i \mid parents(X_i))$，量化其父节点对该节点的影响；

④图是一个有向无环图，缩写为DAG。

贝叶斯网络是概率信息的载体，是联合概率分布的图形表现形式。一个贝叶斯网络通常有两部分组成：第一部分是有向无环图，其每一个节点代表一个随机变量,而每条有向边代表一个概率依赖；第二部分是每个属性一个条件概率表（CPT）。下面图2.1[21]所示是贝叶斯网络的一个简单例子，对心脏病或心口痛患者建模。假设图中每个变量都是二值的。心脏病节点（HD）的父节点对应于影响该疾病的危险因素，例如锻炼（E）和饮食（D）等；心脏病节点的子节点对应于该病的症状，如胸痛（CP）和高血压（BP）等。如图所示，心口痛（HB）可能源于不健康的饮食，同时又可能导致胸痛。

影响疾病的危险因素对应的结点只包含先验概率，而心脏病、心口痛及它们的相应症状所对应的节点都包含条件概率。为了节省空间，图 2.1 中省略了一些概率。我们注意到 $P(X=\bar{x})=1-P(X=x)$，$P(X=\bar{x}\mid Y)=1-P(X=x\mid Y)$，其中 $\bar{x}$ 表示和 x 相反的结果。因此，省略的概率可以很容易求得。例如：条件概率 $P(HD=No\mid E=No, D=\text{健康})=1-P(HD=Yes\mid E=No, D=\text{健康})=1-0.55=0.45$。

假设我们对使用图2.1中的贝叶斯网络来诊断一个人是否患有心脏病感兴趣。下面阐释在不同情况下如何做出诊断：

情况一：没有先验信息的情况下，可以通过计算先验概率 $P(HD=Yes)$ 和 $P(HD=No)$ 来确定一个人是否可能患心脏病。为了表述方便，设 $\alpha\in\{Yes, No\}$ 来表示锻炼的两个值，$\beta\in\{\text{健康，不健康}\}$ 表示饮食的两个值。

$$
\begin{aligned}
P(HD=Yes) &= \sum_{\alpha}\sum_{\beta} P(HD=Yes\mid E=\alpha, D=\beta)P(E=\alpha, D=\beta)\\
&= \sum_{\alpha}\sum_{\beta} P(HD=Yes\mid E=\alpha, D=\beta)P(E=\alpha)P(D=\beta)\\
&= 0.25\times0.7\times0.25+0.45\times0.7\times0.75+0.55\times0.3\times0.25+0.75\times0.3\times0.75\\
&= 0.49
\end{aligned}
$$

因为 $P(HD=No)=1-P(HD=Yes)=0.51$，所以，此人不得心脏病的机率略微大一点。

情况二：高血压。如果一个人有高血压，可以通过比较后验概率 $P(HD=Yes\mid BP=\text{高})$ 和 $P(HD=No\mid BP=\text{高})$ 来诊断他是否患有心脏病。为此，我们

必须首先计算 $P(BP=\text{高})$：$P(BP=\text{高})=\sum_{\gamma}P(BP=\text{高}|HD=\gamma)P(HD=\gamma)$ ＝ $0.85\times0.49+0.2\times0.51=0.5185$, 其中 $\gamma\in\{Yes,No\}$。因此，我们可以得到此人患心脏病的后验概率是：

$$P(HD=Yes|BP=\text{高})=P(BP=\text{高}|HD=Yes)P(HD=Yes)/P(BP=\text{高})$$

$$=0.85\times0.49/0.5185$$

$$=0.8033\text{。}$$

同理，得到：$P(HD=No|BP=\text{高})=1-0.8033=0.1967$。因此，当一个人有高血压时，他患心脏病的危险就会增加。

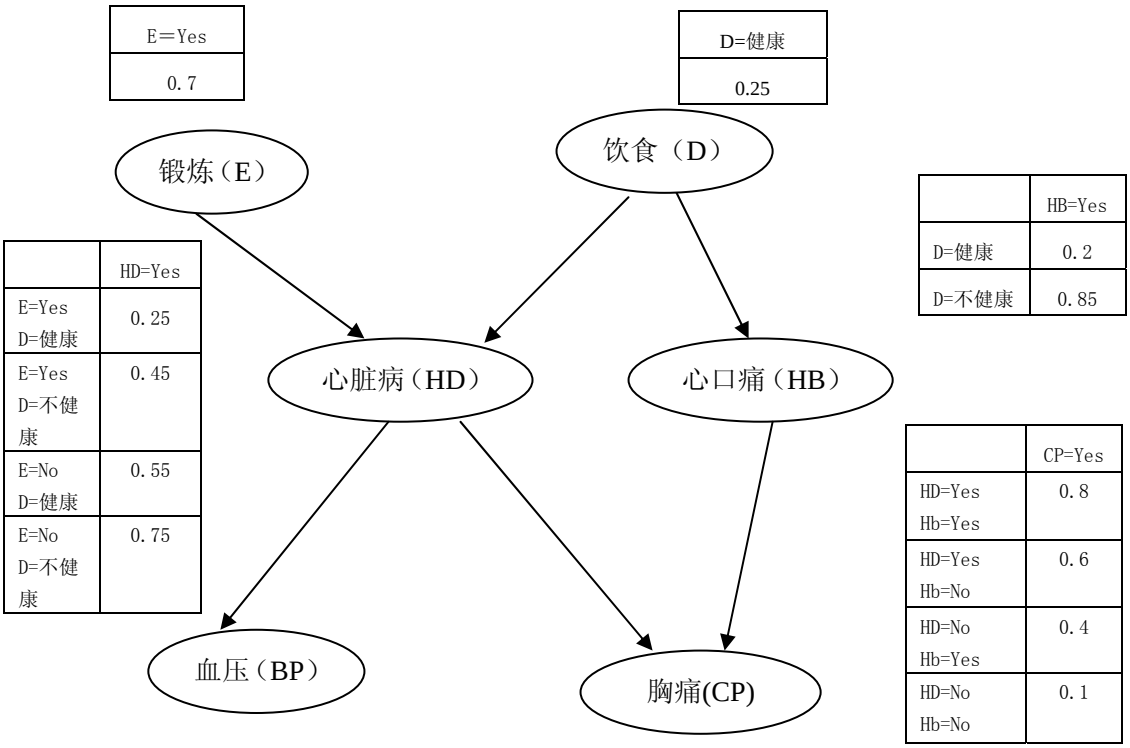


图2.1 发现心脏病和心口痛病的贝叶斯网络

2.3 贝叶斯网络中的条件独立性关系

所谓的条件独立性是指在一定条件的约束下，一个事件的发生不受另一个事件的影响。在解决实际问题中，如果已知一定的先验知识，即给出了一定的基本条件，我们就可以排除某些因素与结果的相关性，有效地节省时间和精力，大大提高决策的效率。而在贝叶斯网络中，条件独立性关系具有重要的地位和作用。

贝叶斯网络结构是对问题领域的定性描述，网络结构中的每个节点（变量），在已知其父节点的条件下独立于所有其余的非子代节点。根据条件独立性，贝叶斯网络将联合概率分解成若干个条件概率的乘积，有效地节省了参数的存储空间，同时，使得概率推理更加直观和便捷，简化了知识获取和领域建模的过程。在定量推理的过程中，条件独立性的利用可以减少先验概率的数目，降低推理过程中的计算复杂度，提高学习和推理的效率。

定义 2.4 如果 X 、Y 和 Z 是具有联合分布 $P(X,Y,Z)$ 的随机变量，如果存在：

$$p(x|y,z)=p(x|z) \qquad \text{其中 } p(y,z)>0 \qquad \text{式（2－16）}$$

则称在 P 分布中，给定 Z 的条件下， X 与 Y 条件独立。通常 P 是不变的，可以从表达式中省略，因此可以表示为 $(X \perp Y|Z)$ 。

特别地，若 $P(X|Y,\varnothing)=P(X)$，也就是 Z 为空集时， X 和 Y 条件独立。

从定义 2.4 可知，通过计算数据集中变量的概率参数，结合条件概率、联合概率等概念以及计算方法，可以判断变量之间的条件独立性。

在信息论中，有两个很重要的概念，一个是互信息，一个是熵。他们在统计学习中是非常重要的。通过计算两个随机变量之间的互信息为变量之间的条件独立性判断提供了一条很好的途径。

定义2.5 对给定的两个离散事件集 $\{X,q(x_k)\}$ 和 $\{Y,w(y_j)\}$ 。事件 $y_i \in Y$ 的出现给出关于事件 $x_k \in X$ 的信息量 $I(x_k;y_j)$ 定义为：

$$I(x_k;y_j)=\log_a p(x_k|y_j)/q(x_k) \qquad \text{式（2－17）}$$

其中 X 和 Y 分别表示两个离散事件集，其中 $X=\{x_k,k=1,2,\cdots,K\}$，对应每个事件， $x_k \in X$ ，相应概率为 $q(x_k)$ 。类似地有 $Y=\{y_j,j=1,2,\cdots,J\}$，对每个事件 $y_j \in Y$ ，相应概率为 $w(y_j)$ 。式（2－17）中对数的底 $a(>1)$ 可任意地选择，不同的底决定不同的信息量单位。最常用的底是“2”和“e”，以2为底时信息的单位称作比特(bit)，即二进制单位；以“e”为底时信息的单位称作奈特（nat）。不同单位之间可通过换算式计算，即 $\log_a x=\log_a b\cdot\log_b x$ ，可得出：1bit=0.698nat，1nat=1.443bit。

类似于定义 $I(x_k;y_j)$ ，可以定义事件 $x_k \in X$ 出现给出的关于事件 $y_i \in Y$ 的信息量为：

$$I(y_j;x_k)=\log_a p(y_j|x_k)/w(y_j) \qquad \text{式（2－18）}$$

有 $p(x_k y_j)=q(x_k)p(y_j|x_k)=w(y_j)p(x_k|y_j)$ 得出这两个信息量是相等的，这就是说事件 x_k 出现给出的关于事件 y_j 的信息量等于事件 y_j 出现给出的关于事件 x_k 的信息

量。我们称 $I(x_k;y_j)$ 为事件 x_k 与事件 y_j 之间的互信息量。之所以有互信息是因为两个事件 x_k 和 y_j 之间统计相关。

若两个事件 x_k 和 y_j 彼此独立，即 $p(x_k y_j)=q(x_k)w(y_j)$，则互信息 $I(x_k;y_j)$ = log1=0；若事件 y_j 的出现有助于肯定事件 x_k 的出现，则互信息大于0；反之，事件 y_j 的出现告诉我们肯定事件 x_k 出现的可能性减小了，则互信息小于0。因此，利用两个事件的互信息的大小，我们可以判断他们之间的独立相关性，同时可以检验变量之间的有向因果关系。

同样的，我们可以将两个概率空间中事件之间的互信息推广到三个概率空间事件之间的互信息。

定义2.6 对于三个离散事件集的联合概率空间 $\{U_1U_2U_3, p(u_1u_2u_3)\}$，在给定事件：$u_3 \in U_3$ 条件下。事件 $u_1 \in U_1$ 和事件 $u_2 \in U_2$ 之间的条件互信息量定义为：

$$I(u_1;u_2|u_3)=\log p(u_1|u_2u_3)/p(u_1|u_3)=\log p(u_1u_2|u_3)/p(u_1|u_3)\,p(u_2|u_3) \quad \text{式（2—19）}$$

对于离散变量的非平均自信息量，有

定义 2.7 给定集合 $\{X,q(x_k)\}$，事件 $x_k \in X$ 的自信息量定义为：

$$I(x_k)=\log 1/q(x_k)=-\log q(x_k) \quad \text{式（2—20）}$$

任意两个事件之间的互信息量不可能大于其中任一事件的自信息量。若将其推广到二事件集的联合空间中，有

定义 2.8 联合空间 $\{XY, p(xy)\}$ 中任一事件 xy，$x \in X$ 和 $y \in Y$ 的联合自信息量定义为：

$$I(xy)=-\log p(xy) \quad \text{式（2—21）}$$

而对于离散变量集的平均自信息量即熵我们定义如下：

定义 2.9 集 $\{X,q(x)\}$ 上定义的随机变量 $I(x)$ 的数学期望：

$$H(X)=MI(x)=\sum_{x\in X} q(x)I(x)=-\sum_{x\in X} q(x)\log q(x) \quad \text{式（2—22）}$$

称为集 X 的平均自信息量，又称作是集 X 的信息熵，简称为熵。

集 X 的平均自信息量表示集 X 中事件出现的平均不确定性，即为了确定集 X 中出现一个事件平均所需的信息量（观测之前），或集 X 中每出现一个事件平均给出的信息量（观测之后）。

定义 2.10 定义在集 $\{Y,w(y)\}$ 上的随机变量 $H(x|y)$ 的数学期望：

$$H(X|Y)=MH(X|y)=\sum_{y} w(y)H(X|y)=-\sum_{x}\sum_{y} p(xy)\log p(x|y) \quad \text{式（2—23）}$$

称作是集 X 相对于集 Y 的条件熵。

当集 X 和集 Y 统计独立时有： $H(X|Y)=H(X)$ 式（2－24）

定义 2.11 定义在集 $\{XY, p(xy)\}$ 上的随机变量 $I(xy)=-\log p(xy)$ 的数学期望

$$H(XY)=MI(xy)=\sum_x\sum_y p(xy)I(xy)=-\sum_x\sum_y p(xy)\log p(xy) \qquad \text{式（2－25）}$$

称作是集 X 和集 Y 的联合熵。

前面我们定义了联合空间 XY 中一对事件 $x\in X$ 和 $y\in Y$ 之间的互信息 $I(x;y)$，它是定义在空间 XY 中的随机变量，对空间 X 上的概率分布进行统计平均得

$$I(X;y)=M_X I(x;y)=\sum_x p(x|y)\log\frac{p(x|y)}{q(x)} \qquad w(y)>0 \qquad \text{式（2－26）}$$

它是特定事件 $y\in Y$ 出现时所给出的关于集合 X 中各事件的平均信息量。

定义 2.12 集合 $\{XY, p(xy)\}$ 上的随机变量 $I(x;y)$ 的数学期望

$$I(X;Y)=M_{XY}I(x;y)=\sum_x\sum_y p(xy)\log\frac{p(x|y)}{q(x)} \qquad \text{式（2－27）}$$

定义为集 $\{X, q(x)\}$ 和集 $\{Y, w(y)\}$ 之间的平均互信息量。

平均互信息量描述了两个集合之间，一个集合中事件出现后所给出的关于另一集合中事件出现的信息量的平均值。

定义 2.13 集 $\{XYZ, p(xyz)\}$ 上定义的随机变量 $I(x;y|z)$ 的数学期望

$$I(X;Y|Z)=MI(x;y|z)=\sum_x\sum_y\sum_z p(xyz)\log\frac{p(x|yz)}{p(x|z)} \qquad \text{式（2－28）}$$

定义为集 X 和 Y 在集 Z 给定条件下的平均条件互信息量。

而对于连续随机变量的互信息和相对熵，有如下定义：

定义 2.14 连续联合集 $\{XY, p(xy)\}$ 中事件 $x\in X$ 和 $y\in Y$ 之间的互信息 $I(x;y)$ 定义为：

$$I(x;y)=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\log\frac{p_{XY}(x|y)\Delta x\Delta y}{p_X(x)\Delta x p_Y(y)\Delta y}=\log\frac{p_{XY}(xy)}{p_X(x)p_Y(y)} \qquad \text{式（2－29）}$$

虽然互信息的概念易于推广到连续情况，但离散变量的熵推广到连续变量时会遇到很大的困难。这是因为连续变量用离散变量逼近时 x 和 y 均取特定值的概率虽趋于零，但定义式中分子和分母的比值常可保持为一个非零的有限值，且随着划分的精细，互信息值会愈来愈大。通常我们称

$$H_c(X) = -\int_{-\infty}^{\infty} p_X(x)\log p_X(x)dx \qquad \text{式（2-30）}$$

为连续随机变量 X 的相对熵，或称为微分熵,简称为熵。同样的，相对熵的概念可推广到联合事件集情况。对联合集$\{XY, p(xy)\}$，我们定义

$$H_c(XY) = -\int\int_{-\infty}^{\infty} p(xy)\log p(xy)dxdy \qquad \text{式（2-31）}$$

为联合事件集 XY 的相对熵。

我们把互信息和相对熵应用到贝叶斯网络结构的构造中，通过计算每对变量的互信息，我们就比较容易的确定存在直接因果关系的变量，对贝叶斯网络的结构进行初步优化。研究贝叶斯网络中的条件独立性,目的是对网络结构进行优化，删除无效边，产生更优的网络结构。在构造贝叶斯网络的方法中，有一个普遍的拓扑规范被称为 d －分离(d-separation),用于在给定第三个节点集 Z 的情况下，判定一个节点集 X 与另一个节点集 Y 是否独立。但是这个规范相当复杂[Russel 和 Norvig,1995]。

由贝叶斯网络的定义可知它实际是一个有向无环图，贝叶斯网络通过有向无环图变量之间的条件独立性，来表达变量之间的依赖关系，而连接边的方向有利于推导和判别变量之间的时序性或因果性。条件独立性的识别和判断对于贝叶斯网络的研究具有重要意义，是贝叶斯网络的关键和核心部分，基于互信息的条件独立性检验理论便于我们更好的从数据中学习网络结构模型，检验节点之间的连接关系，获得更有效的网络结构模型。

2.4 贝叶斯网络的学习

贝叶斯网络的学习是数据挖掘中非常重要的一个环节。而构建一个指定领域的贝叶斯网络包括三个方面:①能够描述该领域的变量及其取值范围;②结构的学习，学习变量间的依赖关系，并以图形化的方式表示出来;③参数的学习，学习变量间的分布函数，获得局部条件概率分布表。这三个任务一般是顺序进行的，然而在构造过程中一般需要在以下两个方面进行折衷:一方面为了达到足够的精度，需要构建一个足够大的、丰富的网络模型;另一方面，考虑到构建、维护模型的费用和概率推理的复杂性，构建的模型应尽可能简单。实际上，建立一个贝叶斯网络往往是上述三个过程迭代地、反复地交互进行。第一个任务主要在领域专家的指导下来选取适宜的变量，同时也需要一定的策略从专家提供的变量中选择重要的因子。后面的两个任务是目前比较活跃的研究领域，即:贝叶斯网络的结构学习和参数学习。其中结构学习是贝叶斯网络学习核心内容。学习分为完整数据和不完整数据学习两种，针对不同情况可以采用不同的学习算法。

现在，学习方法主要针对离散变量，对连续变量一般要经过一定的离散化处

理。为了简化计算，一般情况下，针对样本集D都基于以下假设:①样本中的数据是完备的；②各实例之间是相互独立的；③各实例服从统一的概率分布。对参数θ都基于以下假设:① 参数的局部独立性，即：$p(\theta_i|S)=\prod_j p(\theta_{ij}|S)$；②参数的全局独立性,即：$p(\theta|S)=\prod_i p(\theta_i|S)$ 。

本文采用$U=\{X_1,X_2,\cdots,X_m\}$表示领域变量，$r_1,r_2,\cdots,r_m$为相应变量取值情况，$D=\{D_1,D_2,\cdots,D_n\}$为训练样本集，$\prod_{X_i}$为网络结构S中节点X_i的父节点集，定义$q_i=\prod_{X_i\in\prod_{X_i}} r_i$，$\theta_{ijk}=p(X_i=x_{ik}|\prod_{X_i}=\prod_{x_i}^{j})$表示$X_i$取其第k个取值，而$\prod_{X_i}$取其第j个取值时的条件概率。$\theta_{ij}=(\theta_{ij1},\theta_{ij2},\cdots,\theta_{ijr_i})$，显然有：$0\le\theta_{ijk}\le 1,\sum_k\theta_{ijk}=1$。

2.5 贝叶斯网络的结构学习

通常的网络结构学习的方法有两种。一种是利用变量之间的独立性和条件独立性进行学习，另一种是利用贝叶斯方法对所有备选的网络结构计算后验分布。贝叶斯网络结构学习就是尽可能结合先验知识，找到和样本数据拟合得最好的网络拓扑结构。

2.5.1 完备数据集下网络结构的学习

当数据集完备时，贝叶斯网络的结构学习方法分成两类:基于评分搜索的方法和基于条件独立性测试(Conditional In dependence Test, CIT)的方法。基于评分搜索的方法把贝叶斯网络看成是表示属性之间联合概率分布的结构，学习的目的是得到评分最优的网络结构，该方法一般首先选择网络结构的评分函数，然后搜索评分最优的网络结构。基于条件独立性测试(Conditional In dependence Test, CIT)的结构学习方法把贝叶斯网络结构看作是编码了变量之间条件独立关系的结构，通过学习变量之间独立性关系来确定网络结构。由于需要的条件独立性检验的次数和备选的网络结构的数量都非常庞大，因此在通常情形下，计算量都非常大。

1. 基于评分搜索的网络结构学习方法

基于评分搜索的结构学习方法主要有两部分组成:评分函数和相应的搜索算法。评分函数是一种评价网络拓扑结构与样本集吻合程度的测度，常用的评分函数有:最大似然函数评分(MLE)、贝叶斯评分、最小描述长度(MDL)评分和

Kullback-Leiber等。每种方法都具有各自的特点和不足。

① 最大似然评分(MLE)法

给定样本集D，对任意贝叶斯网络拓扑结构S，可通过最大似然估计得到参数的最大似然估计$\hat{\theta}$。根据拓扑结构S和参数$\hat{\theta}$组成的有向图模型，训练样本集D出现的似然测度为$P(D|S,\hat{\theta})$。基于最大似然评分的结构学习方法的核心是寻求使训练样本集D出现的似然测度最大的拓扑结构S。

为了便于计算，最大似然评分通常取其对数，即:

$$\begin{aligned} Score_L(S|D) &= \ln P(D|S,\hat{\theta}) = \ln \prod_{l=1}^{n} P(D_l|S,\hat{\theta}) \\ &= \ln \prod_{l=1}^{n}\prod_{i=1}^{m} P(x_i(l)|\Pi_{X_i}(l)|S,\hat{\theta}_i) \\ &= \ln \prod_{i=1}^{m}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} (\hat{\theta}_{ijk})^{N_{ijk}} \end{aligned} \qquad \text{式（2—32）}$$

最大似然评分随着网络结构中边的增加而增加，因此，搜索算法总是试图加入更多的边以获得更高的评分，结果，使得似然评分最大的网络结构往往是一个完全图，同时随着网络中节点数目的增加，搜索的复杂度将成倍增长。为了解决这个问题，通常对最大似然附加一个网络结构复杂度的惩罚函数，从而得到如下的评分:

$$Score_L(S|D) = \ln \prod_{i=1}^{m}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} (\hat{\theta}_{ijk})^{N_{ijk}} - f(n)\times \dim(S) \qquad \text{式（2—33）}$$

其中$\dim(S)$表示网络拓扑结构S的复杂度，定义为拓扑结构S中参数的数目，$\dim(S) = \sum_{i=1}^{n} q_i(r_i - 1)$为非负函数，有多种不同的定义，例如AIC准则中定义$f(n)=1$；BIC准则中$f(n)=\frac{1}{2}\ln n$，采用BIC准则的最大似然评分又称为BIC评分，本文记为$Score_{BIC}$。

② 贝叶斯评分法

这里简单介绍一下因果结构贝叶斯后验的计算方法。$p(S|D)$表示在数据D的条件下结构S的后验分布，由贝叶斯定理有

$$p(S|D) = \frac{p(S)p(D|S)}{p(D)} \propto p(S)p(D|S) \qquad \text{式（2—34）}$$

其中，$p(S)$为S的先验，而$p(D|S)$为数据的似然函数，并且

$$p(D|S)=\prod_{i=1}^{m}\prod_{j=1}^{q_i}\frac{\Gamma(a_{ij})}{\Gamma(a_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(a_{ijk}+N_{ijk})}{\Gamma(a_{ijk})} \qquad \text{式（2－35）}$$

其中 $q_i=\prod_{X_i\in\prod_{X_i}} r_i$，$N_{ijk}$ 表示训练样本集中满足条件 $X_i=x_{ik};\prod_{X_i}=\prod_{X_i}^{j}$ 的实例数目，满足 $a_{ij}=\sum_k a_{ijk}, N_{ij}=\sum_{k=1}^{r_i} N_{ijk}$。

对备选网络结构中的每一个结构都计算其后验概率，我们可以选择后验最大的网络结构。贝叶斯最优特性的代价很高，对于真实的学习问题，假设空间通常很大甚至是无限的。

③ 最小描述长度评分(MDL)法

MDL评分是常用的评分函数之一,它综合考虑了似然函数和计算复杂度两方面的因素。该评分函数包括模型自身的编码长度和训练样本数据集的编码长度两个部分。前者代表了网络结构的复杂度，后者则衡量了候选模型与真正模型的相似程度。它的基本思想是:如果描述模型自身的编码长度与利用该模型描述训练样本集的编码长度之和最小，则该模型就是描述训练样本集的最佳模型.我们试图尝试找到最大似然假设（MAP），可最终的结果是找到一个全连接网络，这是因为在给一个节点添加更多的父节点不会减少似然程度。因此我们必须以某种方式对模型的复杂度加以惩罚。MDL方法只是在比较不同的结构之前简单地根据结构的似然度（在参数调整之后）减去惩罚值。

对MDL评分可以表示为：

$$Score_{MDL}=Score_{data}+Score_{network} \qquad \text{式（2－36）}$$

$Score_{data}$ 表示数据的编码长度评分，$Score_{network}$ 表示模型自身网络结构编码长度评分，MDL准则就是使得二者之和达到最小对应的网络结构才是我们所需要的。当然，MDL测度还可以用数据的似然函数值 L_{data} 和网络模型的复杂度 $L_{network}$ 来表示：

$$Score_{MDL}=L_{data}-L_{network} \qquad \text{式（2－37）}$$

其中 L_{data} 采用Kullback-Leibler的距离熵计算得到 $L_{data}=\sum_{i=1}^{m}(p_i-q_i)^2$，$p_i$ 为通过样本网络获得的第 i 个数据集的分布，q_i 为通过参数估计获得的第 i 个数据集的分布。

对模型的结构有了初步的猜测后，然后使用爬山法或者模拟退火搜索、抽样法等进行修正，修正方法可以包括反转、添加或者删除弧。在这个过程中我们必须避免引入环，所以很多算法假定变量是有顺序的，而一个节点的父节点必须是

在排序中先出现的节点。最基本的搜索算法是启发式局部搜索算法(例如贪心算法)，这种方法从给定的初始网络结构(可以是空网络结构、随机指定的网络结构、先验网络结构等)开始，通过增加、删除和转向操作使得局部最大化，再逐渐扩展到整个网络。

基于评分搜索的网络结构学习方法运算复杂程度和结构搜索空间大小随变量增加指数增长(完全搜索是N-P困难问题)，一般要求结点有顺序，或增加一些约束条件来减小搜索范围，并结合启发式搜索方法进行结构学习，适合于具有比较复杂结构且结点较少的贝叶斯网络结构学习。

2. 基于条件独立性测试（CIT)的结构学习方法

基于条件独立性测试（CIT）的结构学习方法是把贝叶斯网络结构看作是编码了变量之间条件独立关系的结构，通过学习变量之间独立性关系来确定网络结构。核心思想是:首先对训练样本集进行统计测试，尤其是条件独立性测试(Conditional Independence Test)，确定出不同结点集之间的一致条件独立性；然后，利用结点集之间的条件独立特性，构造一个有向无环图，以尽可能多地涵盖这些条件独立性。此方法比较容易理解，这与贝叶斯网络的定义紧密相连，它将独立的概念从结构构造中分离出来。结构学习算法可分为启发式搜索和完全搜索,如：SGS算法、CL算法、三阶段算法等。

Spirtes等提出的SGS算法是典型的以条件独立性测试确定拓扑结构的算法，该算法从无向完全图出发，如果相邻结点间存在无向分隔割集，则删除它们间的边，然后通过统计测试来确定剩余边的方向。为寻找结点间的无向分隔割集，需要对其它结点所组成集合的所有子集进行统计测试[22][23]。此外，Acid等证明了有向图模型是否为单连接结构对分类问题的效果影响不大，并提出有向图模型构造EP[24]算法。

Cheng Jie等将信息论用到独立性测试中，使用互信息代替了条件独立测试。经Drafting、Thicking、Thinning三个步骤，通过计算相互信息量(Mutual Information)来确定结点间的条件独立性，从而构造多连接有向图模型。该方法要求事先给定变量的顺序[25]。

2.5.2 不完备数据集下网络结构的学习

数据完备时，贝叶斯网络结构学习算法比较成熟。如何有效地从不完备数据中学习结构是一个难题。数据值缺失是由于数据搜集的不完整或没有相关数据而产生的，如在医疗诊断记录中，没有进行相关的检查就没有相应的数据记录。此外存在隐含或潜在的变量（是指从来也没观测到的变量）时，这样就更复杂了。如果学习算法不知道某个隐变量包含在模型当中，则有两种选择：要么假装数据

集实际是完全的，要么创建新的隐变量以便简化模型。后一个方案可以通过把新的修正选择包含在结构搜索中实现，除了修正连接关系，该算法能够添加和删除隐变量或者改变它的数量。当然，这个算法并不知道它创建的新变量就是要找的隐变量，专家可以根据新变量的局部条件分布推断出它的具体含义。

类似于完备数据的情况，纯粹的最大似然结构学习会产生一个全连接的网络（而且，是没有隐变量的网络），所以需要某种形式的复杂性惩罚手段。我们还可以应用MCMC（马尔可夫蒙特卡洛）方法来近似贝叶斯学习。

(1) Friedman[27]借鉴参数学习的EM 算法, 提出模型选择－期望最大(model selection EM) 的结构学习方法。该算法是将不完备数据下的结构学习问题转化为较容易解决的完备数据下的结构学习问题, 在EM 过程的内部利用启发式搜索算法搜索最佳的网络结构。在1999年提出SEM算法，尽管Friedman简要证明了算法的收敛性，而且目前的实验显示近似过程是收敛的，但是还需要严格的理论证明，而且需要大量计算。

(2) 除Friedman 外, 其他一些学者也对EM 算法如何应用于结构学习进行了探索。基于EM算法框架进行具有丢失数据的贝叶斯网结构学习, 使用期望充分统计因子代替不存在的充分统计因子, 在一些假设下, 可使打分函数具有可分解形式(可进行局部搜索) , 并且在每次迭代中, 结构都有所改进, 使结构序列收敛。该方法能够在一定程度上提高学习效率,但一般是收敛到局部最优结构；而且期望统计因子的计算量非常巨大，是应用结构EM类算法解决大型问题的主要瓶颈。

(3) Pedro[28]将遗传算法引入到结构学习,为这个领域的研究拓展了思路, 补充了新的方法。W. Myers等人改进了Pedro 的工作。他们采用遗传操作把不完备数据转化成完备数据, 把遗漏的数据编码成基因, 同时进化网络结构和遗漏的变量值。这种做法的缺点在于指数级的扩大了搜索空间(遗漏的变量值×网络结构)。国内吉林大学刘大有、王飞等人提出了HGA 算法, 适用于大型数据样本或遗漏的数据量较大时的结构学习。吉林大学王双成、范森淼提出了BN－GS (Bayesian network & Gibbs sampling)。该方法使用Gibbs抽样修复丢失的数据, 基于依赖分析方法进行贝叶斯网结构学习和调整，提高了结构学习的效率。

(4) Myers提出了基于随机搜索思想的学习方法,此类方法主要基于MCMC方法中的MHS (Metropolis-Hastings Sampler)抽样方法的基本思想，此方法可以避免陷入局部极值。

不完备数据的结构学习算法还不成熟，限制了贝叶斯网络实际应用，目前发展的趋势是采用混合算法。

2.6 贝叶斯网络的参数学习

关于贝叶斯网络的参数学习，人们已经做了大量的研究工作。贝叶斯网络参数学习就是已知网络结构，利用先验知识，确定贝叶斯网络模型各节点的条件概率分布表（CPT）。相对于训练样本数据，根据数据的观测情况，可分为完备数据集和不完备数据集，参数学习算法各不相同。

2.6.1 完备数据集下网络参数的学习

完备数据集是指实例具有完整的观测数据，而且由于贝叶斯网络处理的是离散变量，对于连续变量首先要进行离散化处理。常用的参数学习方法有:最大似然参数学习法(MLE)、贝叶斯参数学习法和基于梯度下降的参数学习法(APN)。

1. 最大似然参数学习法(MLE)

最大似然估计基于传统的统计分析思想，依据样本与参数的似然程度来评判样本与模型的拟合程度。似然函数的一般形式为：

$$L(\theta:D)=\prod_{l=1}^{n}P(D_l|\theta) \qquad \text{式（2－38）}$$

最大似然估计MLE的基本思想就是寻找使得似然函数取得极大值的参数θ^*作为对参数的估计值。因此似然性是判断具体θ优劣的一种标准。似然性越大，具体的θ就越好。则有：

$$L(\theta:D)=\prod_{l=1}^{n}P(D_l|\theta) \quad =\prod_{l=1}^{n}\prod_{i=1}^{m}p(x_i(l)\left|\prod_{x_i}(l),\theta_i\right.)$$

$$=\prod_{i=1}^{m}L_i(\theta_i:D) \qquad \text{式（2-39）}$$

其中局部似然函数还可进一步分解，即：

$$L_i(\theta_i:D)=\prod_{l=1}^{n}P(x_i(l)\left|\prod_{x_i}(l),\theta_i\right.)$$

$$=\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}(\theta_{ijk})^{N_{ijk}}$$

其中$q_i=\prod_{X_i\in\prod_{X_i}}r_i$，$N_{ijk}$表示训练样本集中满足条件$X_i=x_{ik};\prod_{X_i}=\prod_{X_i}^{j}$的实例数目，

定义$N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$，则MLE方法计算出$\theta_{ijk}^{*}=\dfrac{N_{ijk}}{N_{ij}}$ 式（2－40）

利用该公式，可以很容易的计算出各节点的条件概率分布表。通过以上工作我们总结出进行最大似然参数学习的一种标准方法：

① 写出数据的似然表达式，它是待学习参数的一个函数；

② 对每个参数的对数似然进行求导；

③ 找到满足导数为0的对应参数值。

尤为注意的是，当数据集很小时，以致有些事件没有被观测到，最大似然假设会对这些事件赋予0概率，即 $N_{ij}=0$，尽管采用了很多技巧避免这个问题，但是对数据集很小的实例，此方法得到的结果不是很可靠，因此一般要求数据规模很大，而且计算速度相对较慢，缺乏处理先验知识的能力。

2. 贝叶斯参数学习法

贝叶斯方法是利用贝叶斯网络表示数据取样过程中的不确定性，与传统的统计方法最大的差别在于两者对不确定性的看法上，前者认为不确定性是人们对事物的一种认知程度，这种认知程度是由原来的主观知识和观察到的现象共同决定的，而后者把概率简单的看作是频率的无限逼近。此方法的基本思想是：给定一个含有未知参数的分布以及一个完整的实例数据集合 D。将未知参数看成是一个随机变量，根据以往对参数 θ 的知识，确定先验分布 $p(\theta)$，或者认为 $p(\theta)$ 是一个均匀分布，利用先验知识 ξ，然后根据公式（2－9）计算后验概率 $p(\theta|D)$，作为参数估计的依据。

Raiffa和Schaifeer提出先验分布应选取共扼分布。通常先验分布 $p(\theta)$ 取Dirichlet分布。即 $p(\theta|S)=Dir(\theta|\alpha_1,\alpha_2,\cdots,\alpha_r)$，则有

$$p(\theta_{ij}|S)=\frac{\Gamma(\alpha_{ij})}{\prod_{k=1}^{r}\Gamma(\alpha_{ijk})}\prod_{k=1}^{r}\theta_{ijk}^{\alpha_{ijk}} \qquad \text{式（2－41）}$$

其中 $\alpha_{ij}=\sum_{k=1}^{r}\alpha_{ijk}$，$\alpha_{ijk}$ 大于0，$i=1,2,\cdots,n;\ j=1,2,\cdots,q_i;k=1,2,\cdots r_i$，$\alpha_{ijk}$ 为超参数，$\Gamma(\cdot)$ 为Game函数，即 $\Gamma(x)=\int_0^{\infty}t^{x-1}e^{-t}dt$，满足 $\begin{cases}\Gamma(x+1)=x\Gamma(x)\\ \Gamma(1)=1\end{cases}$，那么参数的后验概率也服从Dirichlet分布，即：

$$p(\theta_{ij}|D,S)=\frac{p(\theta_{ij}|S)p(D|S,\theta_{ij})}{p(D|S)}=\frac{\Gamma(\alpha_{ij}+N_{ij})}{\prod_{k=1}^{r}\Gamma(\alpha_{ijk}+N_{ijk})}\prod_{k=1}^{r}(\theta_{ijk})^{\alpha_{ijk}+N_{ijk}}$$

$$=Dir(\alpha_{ij1}+N_{ij1},\alpha_{ij2}+N_{ij2},\cdots,\alpha_{ijr}+N_{ijr}) \qquad \text{式（2－42）}$$

其中 N_{ijk} 表示训练样本集中满足条件 $X_i=x_{ik};\prod_{X_i}=\prod_{X_i}^{j}$ 的实例数目，定义

$N_{ij}=\sum_{k=1}^{r_i}N_{ijk}$，此时参数$\theta$的最大后验估计为：$\theta_{ijk}^{*}=\frac{\alpha_{ijk}+N_{ijk}}{\alpha_{ij}+N_{ij}}$ 式（2－43）

由于贝叶斯方法可以综合先验信息和后验信息，即可避免只使用先验信息带来的主观偏见，也可避免只使用后验信息带来的噪音的影响，克服了MLE方法的缺点。

2.6.2 不完备数据集下网络参数的学习

不完备数据集是指对某个实例的观测有部分缺值或观测异常的情况。此时似然函数的计算将变得很复杂，精确计算极大值几乎是不可能的，因此对不完备数据的学习，一般要借助于近似的方法，求出似然函数的极大值，并将该点的参数作为估计值。如Monte-Carlo方法，Gaussian逼近，以及EM（期望一极大化)算法求ML(极大似然)或MAP(极大后验)等。尽管有成熟的算法，其计算开销也是比较大的。

1. Monte-Carlo 方法

Gibbs取样方法是最为流行的Monte-Carlo方法之一。Gibbs把含有不完备数据集D的每一个缺项当作待估计参数，通过对未知参数后验分布的一系列随机抽样过程，计算参数的后验均值的经验估计。Gibbs抽样法[29]的基本思想是以Gibbs取样估算变量集X上联合概率分布$p(X)$的数学期望$f(X)$。修复过程是:①以某种方式初始化不完备的数据集D，得到一完整的数据集D_c.②在初始样本D中选取一个没有被观察到的变量X_{il}(第l个事例的第i个变量)，按照下面的概率分布给X_{il}的状态随机地赋值：$p(x_{il}'|D_c\setminus x_{il},S)=\frac{p(x_{il}'|D_c\setminus x_{il},S)}{\sum_{x_{il}''}p(x_{il}''|D_c\setminus x_{il},S)}$,其中$D_c\setminus x_{il}$表示$D_c$去掉观察$X_{il}$；③按照②逐个为$D$中每个事例中未观察到地变量赋值，直至产生一个新的随机完备数据集D_c'；④计算$p(\theta_s|D,S)$;⑤重复①－④足够多次，取众多$p(\theta_s|D,S)$的平均值作为最终的参数估计值。

Monte-Carlo方法在其他方法不适用时，可以应用，非常灵活，样本越大，运行时间越长，结果越精确，但计算复杂度是事例数目n的指数幂。很多学者也提出了不同的类似的版本，如Zhang和Poole提出了基于剪枝的重要性抽样，使用重要

性抽样方法估计出近似的后验概率，按照它抽取出样本等等。

2. EM 方法

对不完备数据集时下有一种非常流行的极大似然估计方法Expectation-Maximization算法，通常简称为EM算法。它不是直接对复杂的后验分布进行极大化或模拟，而是在观察数据的基础上添加一些“潜在数据”，从而简化计算并完成一系列简单的极大化或模拟。它之所以被称为EM算法是因为算法的每一次迭代由一个期望步(E-step)和极大步(M-step)构成[30]，EM算法的特点是简单和稳定，特别是每一次迭代能保证观察数据对数后验似然是单调不减的。

EM算法是一种一般的从不完备数据集中求解模型参数的极大似然估计的方法。它的基本思想是:可以观察到的数据集是D，完备数据集$D_C=(D,Z)$，Z是缺失数据集，θ是模型参数。θ关于D的后验分布$p(\theta|D)$很复杂，难以进行各种不同的统计计算。假如缺失数据集Z已知，则可能得到一个关于θ的简单后验分布$p(\theta|D,Z)$,利用$p(\theta|D,Z)$可以进行各种统计计算。然后，又可以对Z的假定作检查和改进，如此将一个复杂的极大化或抽样问题转化为一系列简单的极大化或抽样问题。EM参数学习方法包含两个步骤:①E-step :利用当前网络结构和参数对缺失数据计算它的期望值;②M-step :基于E-step参数的期望值，计算出新的最大可能的参数分布θ'，用θ'替换原有的θ;重复过程①和②，直到所估计的参数达到指定的迭代次数或达到局部最优。Dempster[31]证明，EM算法存在局部极值问题，但是很多人提出了不同的改进的EM算法，如Neal和Hinton提出的采用“求部分期望”的增量EM算法，EM算法仍然是很流行的参数学习方法。

3. Gaussian 逼近方法

Gaussian逼近方法是一种比Monte-Carlo方法计算复杂度低，在处理大型样本也可以得到较准确结果的逼近方法。它的基本思想是：对大规模数据而言，用多元高斯分布近似模拟$p(\theta_s|D,S)\propto p(D|\theta_s,S)p(\theta_s|S)$,要完成高斯近似，必须找到$\theta_s'$，使得$p(\theta_s|D,S)$取得最大。Rubin等人讨论了计算$\theta_s'$的技巧，Raftery则介绍了如何利用很多统计包计算出似然比率来近似计算Horace行列式。Thiesson介绍了计算无限制的多项分布的Bayesian网推理。

Monte-Carlo方法、EM方法等与缺失数据的比例密切相关，比例越高，Monte-Carlo方法越不精确，达到收敛时间越长，而EM算法出现局部极值的可能性就越大；此外二者都需要很大的资源，收敛速度较慢，算法的执行事件和缺失数据的数目密切相关，而且要求满足MAR假设。从算法的计算复杂性来看，Gibbs抽样算法的效率是最低的，且只能对条件概率表进行批量更新。

第三章 基于贝叶斯网络的数据挖掘方法

本章首先介绍了数据挖掘的相关问题以及目前主流的一些算法，贝叶斯网络(Bayesian Networks, BN)方法在数据挖掘中的应用是当前研究的热点。贝叶斯网络是一种进行不确定性推理和知识表示的有力工具，当与统计方法结合使用时，显示出许多关于数据处理的优势。针对目前还没有一种完整的在数据挖掘中构建贝叶斯网络的算法步骤，探讨性的提出了一种启发式的在数据挖掘中利用样本数据构建贝叶斯网络的算法思想。该算法较好的解决了在数据挖掘中利用样本数据设计贝叶斯网络问题，得到了将贝叶斯网络应用于数据挖掘当中，充分挖掘数据的隐含信息和内在本质,具备良好的预测能力等优点。

3.1 数据挖掘的定义

随着计算机硬件和软件的飞速发展，尤其是数据库技术与应用的日益普及，人类正面临着“数据海洋”的困惑。如何有效地利用和处理大量的数据即通过分析数据对象之间关系提取隐含在数据中的知识，成为广大信息技术工作者所关注的焦点。正是在这种背景下，数据挖掘技术应运而生。

数据挖掘(Data Mining, DM)是一门新兴的交叉学科，涉及到数据库技术、人工智能、知识工程、统计学、机器学习、优化计算和专家系统等领域。从本质上来说，数据挖掘是智能信息处理的一种过程或技术，它在对大量数据实例全面而深刻认识的基础上，通过计算、归纳和推理等环节，从中抽取普遍的、一般的和本质的现象或特征。

数据挖掘有如下特点：

第一、数据挖掘的数据量常常是巨大的。因此，如何高效率地存取数据，如何根据一定应用领域找出数据关系即高效率算法以及是使用全部数据还是使用部分随机或有目的地选择出的数据子集，都成为数据挖掘工作者要考虑的问题。

第二、数据挖掘面临的数据常常是为其他目的而收集好的数据。这就为数据挖掘提出了一个问题，即收集数据时，可能有一个或几个重要的变量未被收集，而这些变量在后来做数据挖掘时被证明是有用的，甚至是至关重要的。也有是说，未知性和不完全性将始终伴随数据挖掘的过程。

第三、数据挖掘工作者常常不愿把先验知识预先嵌入算法内，因为这样就等于做“假设检验”（但这不排除把统计中的假设检验作为其中间一步来做）。

3.2 数据挖掘的任务

数据挖掘所涉及的学科领域和方法很多，主要的任务简单的说就是从海量的信息或数据中挖掘出有用的信息。数据总结其目的是对数据进行浓缩，给出它的紧凑描述。数据挖掘主要关心从数据泛化的角度来讨论数据总结。数据泛化是一种把数据库中的有关数据从低层次抽象到高层次上的过程。

分类其目的是学会一个分类函数或分类模型(也称作分类器)，该模型能把数据库的数据项映射到给定类别中的某一个。

聚类是把一组个体按照相似性归类，即"物以类聚"。它的目的是使得属于同一类别的个体之间的距离尽可能地小，而不同类别的个体间的距离尽可能地大。

关联规则是形式如下的一种规则，"在购买面包和黄油的顾客中，有90%的人同时也买了牛奶"(面包+黄油+牛奶)。关联规则发现的思路还可以用于序列模式发现。用户在购买物品时，除了具有上述关联规律，还有时间或序列上的规律。此外，偏差分析或称孤立点分析也是数据挖掘的主要任务之一。偏差分析的基本思想是寻找观察结果与参照量之间的有意义的差别。数据挖掘比较常见的成功运用包括：解释性数据的分析、描述性建模、预测性建模、知识发现、模式和规则的识别、文字图像信息的搜寻等。

3.3 数据挖掘的过程

数据挖掘是一个需要经过反复的多次处理过程。数据挖掘的体系框架为数据挖掘提供了宏观指导和工程方法。合理的体系框架将各个处理阶段有机地结合在一起，指导人们更好地开发及使用数据挖掘系统，为此，建立了如图3.1所示的数据挖掘体系框架。其中选择数据为数据准备阶段，准备阶段做的好，数据质量高，数据挖掘更加快捷，得到的知识和信息更有效；数据预处理主要包括数据清洗、数据转换、数据归并和数据集成等形式，可以保证数据的质量，利于我们进行挖掘；根据要实现的目标，进行算法设计和实现，建立模型，并进行分析，获取所需的知识与信息，最终实现数据挖掘的目的。

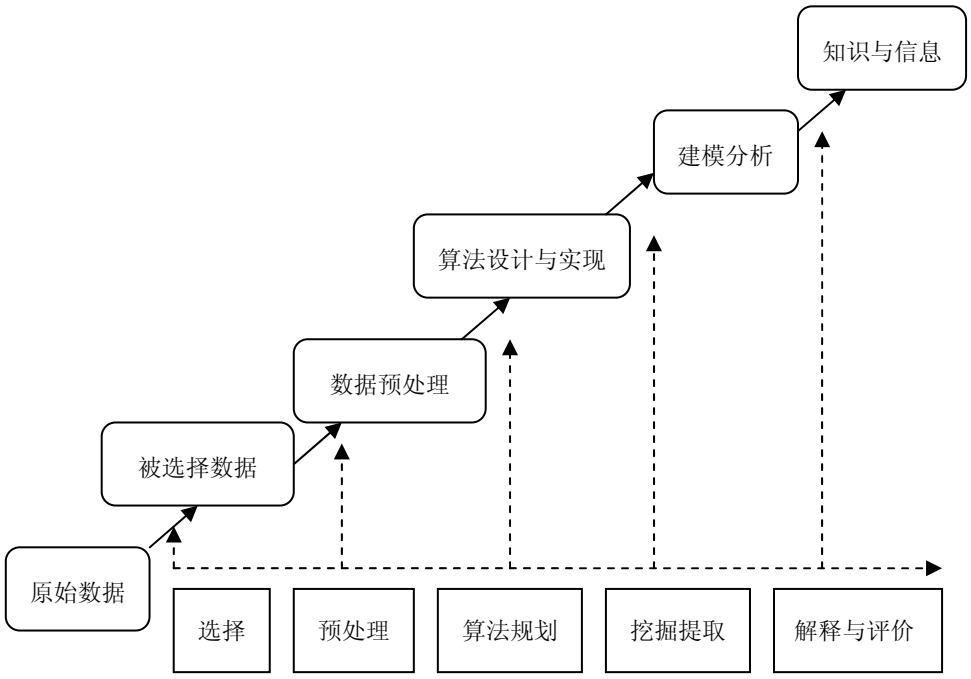


图3.1 数据挖掘体系框架

数据挖掘与知识发现处理过程的共同点都要经过准备、预处理、算法设计、数据挖掘和后处理等共同的阶段。处理的流程图如图 3.2 所示。

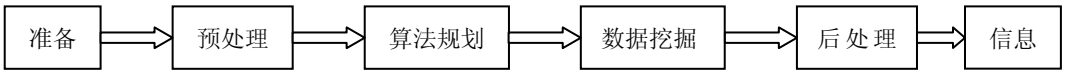


图 3.2 数据挖掘处理的流程图

图 3.2 中的准备阶段主要指了解相关领域的有关情况，熟悉背景知识，弄清用户要求，根据要求从数据库中提取相关的数据；数据预处理主要对前一阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行处理，对丢失的数据进行填补；算法设计就是要根据目标选择合适的算法；数据挖掘就是运用选定算法，从数据中提取出用户所需要的知识，这些知识可以用一种特定的方式表示或使用一些常用的表示方式；在后处理阶段，主要是使用统计度量或假设检验，删除虚假的数据挖掘结果，根据需要对知识发现过程中的某些处理阶段进行优化，直到满足要求，获得所需信息。

3.4 数据挖掘的算法

数据挖掘技术包含很多算法，常用的主流方法有决策树、遗传算法、贝叶斯网络方法、粗糙集、神经网络等。每种算法都有自身的功能和优势，其中决策树

的优点是可理解性，很直观，主要用于分类和归纳挖掘，但在数据量较大和数据复杂的情况下，该算法则显得力不从心；遗传算法擅长于数据聚类，在组合优化问题上具有独特的优势；粗糙集在数据挖掘中具有重要的作用，常用于处理含糊性和不确定性的问题，以及特征归纳和相关分析，运用粗糙集进行数据预处理可以提高知识发现的效率；神经网络能够对复杂问题进行预测，它在商业界得到广泛的应用，对于信贷客户识别、股票预测和证券市场分析等方面效果良好；而贝叶斯网络具有分类、聚类、预测和因果分析等功能，易于理解，预测效果较好，面对大规模数据时显示出它独特的优势，具体的如2006年梁循提出的图3.3所示。

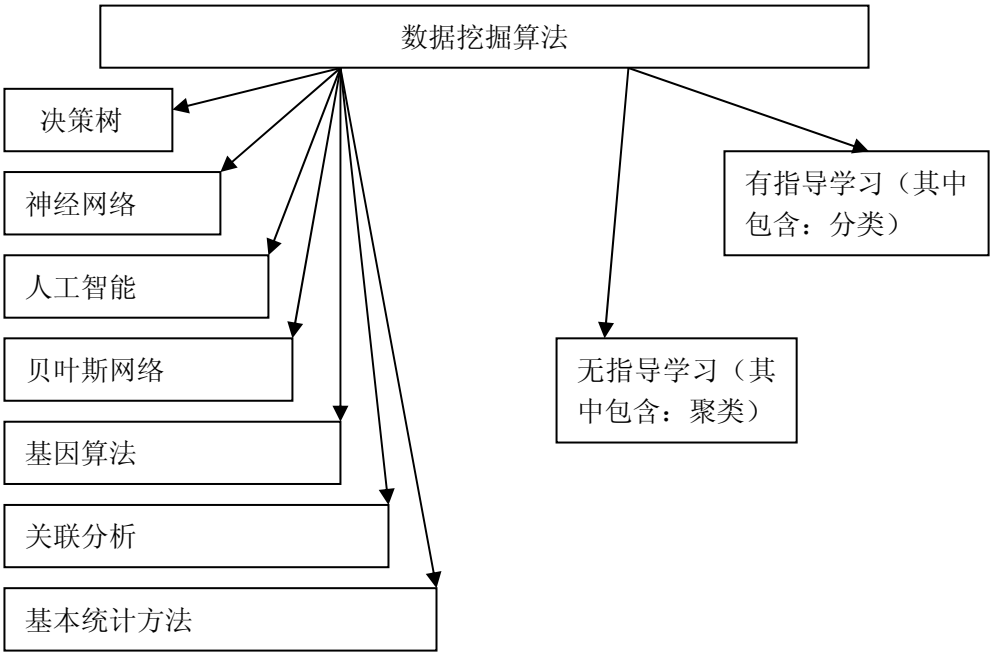


图 3.3 数据挖掘主要算法框架

3.4.1 决策树算法

决策树算法是一种常用的数据挖掘算法，它是从机器学习领域中逐渐发展起来的一种分类函数逼近方法。所谓决策树就是一个类似流程图的树型结构,其中树的每个节点对应一个非类别属性,每条边对应这个属性的每种可能值,而树的每个叶结点代表一个类别。

决策树学习的基本算法是贪心算法，采用自顶向下的递归方式构造决策树。生成决策树的一个著名的算法是Quinlan 提出的ID3算法。ID3算法的基本思想是自顶向下地使用贪心算法搜索训练样本集，在每个节点处测试每一个属性，从而构建决策树。为了选择训练样本地最优分支属性，ID3使用信息增益作为分支指标。ID3 算法存在着许多不足，如：不能够处理连续值属性，计算信息增益时偏向于

选择取值较多的属性，对噪声较为敏感等。后来Quinlan在1993年提出了ID3的改进算法C4.5算法，算法采用深度优先策略，C4.5的分支指标采用增益比例，克服了信息增益度量的缺点。常用的决策树算法还有Breiman等人在1984年提出的CART算法，也采用深度优先策略，它的一大优点是它将模型的验证和最优通用树的发现嵌在了算法中。SLIQ 算法，采用的是宽度优先策略，后人也研究出许多改进的决策树算法，性能得到进一步改善。

决策树主要用于分类和归纳挖掘，优点是可理解性，很直观，构建速度比较快，但缺点是在数据量较大和数据复杂的情况下，该算法则显得力不从心；对于具有连续值的属性预测较困难；当类别太多时，误差较大，而且无法处理缺失数据等。

3.4.2 遗传算法

遗传算法（Genetic Algorithm, GA）最早是由美国Holland教授在20世纪70年代提出的，是一种由生物进化过程激发灵感而产生的算法。它们提供一种自适应、鲁棒、并行和随机化的搜索技术，其中解的群体一代代顺序进化而形成总体最优解，是一种高效的全局并行搜索优化算法。遗传算法应用于数据挖掘，其基本原理是：类比生物进化过程，每一代同时存在许多不同的种群个体。这些个体的适应性以适应度函数 $f(x)$ 表征，个体的保留与淘汰取决于它们对环境的适应能力，优胜劣汰。适应度函数 $f(x)$ 是整个遗传算法极为关键的一部分，其构成与目标函数密切相关，往往是目标函数的变种。基本遗传算法中使用选择、交叉和变异这3种遗传算子。规则群体通过交叉和变异操作“进化”，直到群体中所有的规则都满足指定的阈值。遗传算法可起到产生优良后代的作用，经过若干代遗传,将会得到满足要求的后代(问题的解)。遗传算法具有启发式策略以便脱离局部最小值，最适合以较长的计算时间解决复杂的、了解很少的问题，此算法易于并行，并且业已用于分类和其他优先问题。

在数据挖掘中,遗传算法擅长于数据聚类，在组合优化问题上具有独特的优势。但仍然存在缺点，如：（1）算法较复杂；（2）收敛于局部极小的过早收敛等难题尚未解决；（3）通常计算时间较长等。

3.4.3 粗糙集算法

粗糙集理论是波兰科学家Pawlak在1982年提出的，借鉴了逻辑学和哲学中对不精确、模糊的各种定义，针对知识库，提出不精确范畴等概念，并在此基础上形成了完整的理论体系－粗糙集理论。粗糙集理论在知识表达系统的基础上定义

了约简与核这两个重要的概念，进而提供了分析多余属性的方法，对属性进行约简。而属性约简是数据挖掘研究的一个重要内容。粗糙集算法是基于给定训练数据内部的等价类的建立,发现不准确数据或噪声数据内在的结构联系，它用于离散值属性。用粗糙集来处理不确定性问题的最大优点在于，它不需要关于数据的预先或附加的信息，而且容易掌握和使用。粗糙集算法对知识不完全的处理是有效的，但由于这个理论包含处理不精确或不确定原始数据的机制，单纯地使用这个理论不一定能有效地描述不精确或不确定的实际问题，需要其他方法补充。如：粗糙集理论和模糊集理论相结合，它们既相互独立，又相互补充。

粗糙集在数据挖掘中具有重要的作用，常用于处理含糊性和不确定性的问题，以及特征归纳和相关分析，运用粗糙集进行数据预处理可以提高知识发现的效率。

3.4.4 神经网络算法

人工神经网络（neural network, NN）是基于对人脑思维的探索和模仿而发展起来的一门学科。神经网络学习方法对于逼近实数值、离散值或者向量值的目标函数提供了一种健壮性很强的方法。神经网络应用于数据挖掘领域，获得了卓有成效的成果。神经网络的研究已经获得许多方面的进展，提出了大量的网络模型，发现了许多学习算法。神经网络可以分成四种类型:前向型、反馈型、随机型和自组织竞争型。其中，前向型神经网络是数据挖掘中广为应用的一种网络，其原理或算法也是其他一些网络的基础。神经网络的性质主要取决于以下两个因素:一个是网络的拓扑结构；另一个是网络的权值、工作规则。二者结合起来就可以构成一个网络的主要特征。随着网络结构和功能的不同，网络的权值的学习算法也不同。神经网络的学习问题归根结底就是网络的权值调整问题。权值的确定一般有两种方式:一种是通过设计计算确定;另一种是网络按一定的规则通过学习而得到的。

神经网络能够对复杂问题进行预测，它在商业界得到广泛的应用，对于信贷客户识别、股票预测和证券市场分析等方面具有良好的效果。尤其在金融市场预测领域里显示了较好的应用前景。基于人工神经网络的分类法相对于决策树方法通常分类出错率比较低，但它的缺点是需要很长的学习训练时间。

3.5 贝叶斯网络方法应用于数据挖掘

贝叶斯网络(Bayesian networks)，是伴随着影响图(influence diagram) 而发展起来的一类决策分析工具,它提供了不确定性环境下的知识表示、推理、学习手段,可以完成决策、诊断、预测、分类等任务，贝叶斯网络具有分类、聚类、预

测和因果分析等功能，易于理解，预测效果较好，已广泛应用于语音识别、工业控制、经济预测、医疗诊断等。近几年来，在数据挖掘中的应用则为贝叶斯网络开辟了一个新的研究空间。

贝叶斯网是概率信息的载体，是联合概率分布的图形表现形式。一个贝叶斯网络通常有两部分组成：第一部分是有向无环图，其每一个节点代表一个随机变量,而每条弧代表一个概率依赖；第二部分是每个属性一个条件概率表（CPT）. 图3.4给出了6个布尔变量的简单贝叶斯网络，弧表示因果知识，例如，得肺癌受吸烟的影响，也受家族肺癌史的影响。此外，该弧还表明，给定其双亲家族肺癌史（FH）和吸烟(S)，变量肺癌（LC）条件独立于肺气肿（E）,这意味着，一旦家族肺癌史（FH）和吸烟（S）的值已知，变量肺气肿（E）并不提供关于肺癌（LC）的附加信息。表3.1则给出了肺癌（LC）的CPT，对于其双亲节点家族肺癌史（FH）和吸烟（S）的每个可能组合，表中给出了LC的每个值的条件概率，如表3.1所示。如: $P(LC="no"|FH="no",S="yes")=0.5$ 。

对应于属性或变量 $X_1,X_2,\cdots,X_n$ 的任意元组（ $X_1,X_2,\cdots,X_n$ ）的联合概率由下式计算 $P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|parents(X_i))$ 其中 $P(X_i|parents(X_i))$ 的值对应于 X_i 的CPT中的值。

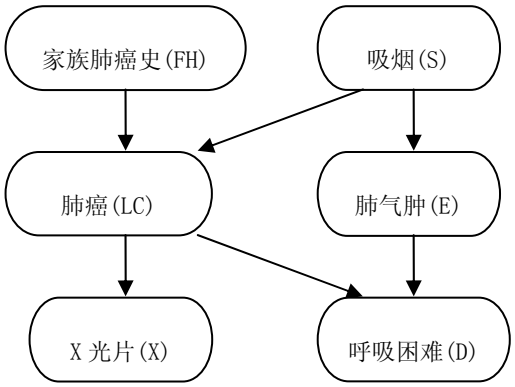


图3.4 一个贝叶斯网络的结构示意图

表 3.1 有关肺癌（LC）值的条件概率表（CPT）

	FH, S	FH,～S	～FH,S	～FH,～S
L C	0.7	0.4	0.5	0.2
～LC	0.3	0.6	0.5	0.8

3.5.1 贝叶斯网络方法优势

贝叶斯网络作为一种图形化的建模工具，应用于数据挖掘，具有一系列的优势：

① 贝叶斯网络具备高效灵活的学习机制，对于问题域的建模，当条件或信息等发生变化时，不用对模型进行修正，就可以发现潜在有用的模式或关系，实现对数据实例的分类、聚类和预测，还可以有效的处理缺失和带有噪声的数据集。

② 贝叶斯网络最大的优点就是能够处理各种不确定性信息，用图形的方法描述数据间的相互关系,语义清晰,易于理解和接受。

③ 由于贝叶斯网络中节点之间是相互影响的，任何节点观测值的获得或者节点信息的任何改变，都会影响其他的节点，因此，具有良好的推理和预测能力。

④ 贝叶斯网络是图论和概率论的有机结合，不但具有直观的知识表示形式，而且具有因果和概率性语义,可以有机地结合人类的先验知识和样本数据，将主观和客观有机地结合起来，更加全面客观地反映数据对象内在的联系与本质，克服了神经网络等方法不够直观的缺点。

⑤ 贝叶斯网络可以有效的避免对数据的过度拟合。

由于贝叶斯网络具有以上优点，在数据挖掘等领域得到了广泛的研究和应用。

3.5.2 基于贝叶斯网络的数据挖掘算法思想

由于贝叶斯网络能综合考虑先验信息和样本数据，充分地利用专家知识和经验，将主观和客观有机地结合起来等众多优于其他方法的特点，针对目前还没有一种完整的在数据挖掘中构建贝叶斯网络的算法步骤，探讨性的提出一种启发式的在数据挖掘中利用样本数据构建贝叶斯网络的算法思想：

（1）首先根据数据挖掘的目标和任务，进行数据分析和变量的选取，确定需要哪些变量描述该领域，以及每个变量的确切含义。

（2）假设任意两个变量之间都存在依赖关系，用连接边表示变量之间的关联性，构成一个全连接图，共有 $n(n-1)/2$ 条边。

（3）利用基于互信息测度和条件独立性检验(CI)的边删除算法，结合先验信息和专家知识学习出一个最小无向图。

互信息是一个随机变量包含另一个随机变量的信息量测量，它表明一个随机变量由于获得另一个变量的信息而减少不确定性。互信息的定义如下：

① 两个离散随机变量 X 和 Y 具有联合概率函数 $p(x,y)$ 和边缘概率函数 $p(x)$

和 $p(y)$，其互信息 $I(X;Y)$ 定义为：$I(X;Y)=\sum_{x\in Q_x}\sum_{y\in Q_y}p(x,y)\frac{p(x,y)}{p(x)p(y)}$

② 给定 Z 的条件下，随机变量 X 和 Y 的条件互信息定义为

$$I(X;Y|Z)=\sum_{x\in Q_x}\sum_{y\in Q_y}\sum_{z\in Q_Z}P(X,Y,Z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}$$

$$=H(X|Z)-H(X|Y,Z)=E_{p(x,y,z)}\left[\log\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right]$$

计算任意两个节点 V_i 和 V_j 之间的互信息 $I(V_i,V_j)$，按降序进行排列，构成节点对 (V_i,V_j) 的集合K。应用边删除算法，设定一个为较小的门槛值 $\varepsilon>0$，如果互信息 $I(V_i,V_j)<\varepsilon$，则删除 V_i 和 V_j 之间的连接，否则保留节点之间的连接。然后对节点进行条件独立性（CI）检验，进一步删除多余的连接，直到得到一个最小无向图。

（4）确定连接边的方向性。利用先验信息和专家知识，可以排除大量的不合理的结构，再同时运用最大后验概率（MAP）和最小描述长度(MDL)的准则，调整节点间的连接边，同时确定图中节点之间连接边的方向性。

把一个变量 X 初始化为一个值集 $\{x_1,x_2,...,x_n\}$，如果一个值 x 对应于信源的概率 $p(x)=p(X=x)$，值 x 的信息量定义为：$I(X=x)=I(x)=-\log_2 p(x)$。对该信息源的最优描述语言应该是变量 X 的最小描述长度，因此描述长度 $L(x)$ 与信息量 $I(x)$ 相等。基于这种关系，选取最优结构 S_0 的最大后验概率（MAP）和最小描述长度（MDL）的准则为：$S_0=\arg\max_{S\in s}\left[P(D|S)P(S)\right]=\arg\min_{S\in s}\left[L(D|S)+L(S)\right]$

（5）通过相关的专家知识和规则，利用样本数据,对有向无环图进行修正和优化，调整贝叶斯网络的网络结构,经过计算与分析，学习出与样本数据达到最佳匹配的网络结构。

（6）利用参数学习算法确定节点的条件概率分布函数表（CPT）。当数据集完整时，概率参数学习一般有两种方法：一是经典的样本统计法，二是贝叶斯统计法；当数据集不完备时，一般采用近似计算的方法，近似求出似然函数的极大值，并将该点的参数作为估计值。通过由专家确定的网络结构中每个变量的条件概率分布函数，量化变量之间的依赖关系,利用先验信息和专家知识，得到最优的贝叶斯网络模型。

第四章 实验结果与分析

基于贝叶斯网络的数据挖掘方法思想，运用到具体的实例中，首先建立了大学毕业生考研情况的贝叶斯网络模型，并与决策树方法进行了比较，得到将贝叶斯网络应用于数据挖掘当中,充分挖掘数据的隐含信息和内在本质,具备良好地预测能力等优点；同时就当前国内研究农户信用等级评定主要是运用传统的信用评分方法，主观因素比重大，严重影响了等级评定的准确性。基于此，建立了基于贝叶斯网络的农户信用等级评定模型。并与传统的信用评分法进行比较，找出了影响农户信用等级的决定因素，此方法利于信用等级的评定，提高了信用评级的准确性。实验结果表明本文提出的算法设计简单，方法实用，应用有效，与其它算法相比还有精度较高的特点，同时也表现出了该算法在数据挖掘方面的优势，利于实际中的管理、分析、预测和决策等。

4.1 大学毕业生考研情况网络模型

如今，就业形势非常严峻，大学生大学毕业后都面临着是上研究生还是走向社会工作的两种选择，下面通过对某一地区大学毕业生考研情况进行调查，利用起决定性的几个因素，对10000名大学生的情况进行统计，构造贝叶斯网络，验证算法的可行性和有效性，以及贝叶斯网络在数据挖掘中的优势。

4.1.1 案例实验结果

通过对某地区的大学毕业生考研情况进行调查, 找出以下几个变量因素对他们是否继续深造产生影响:

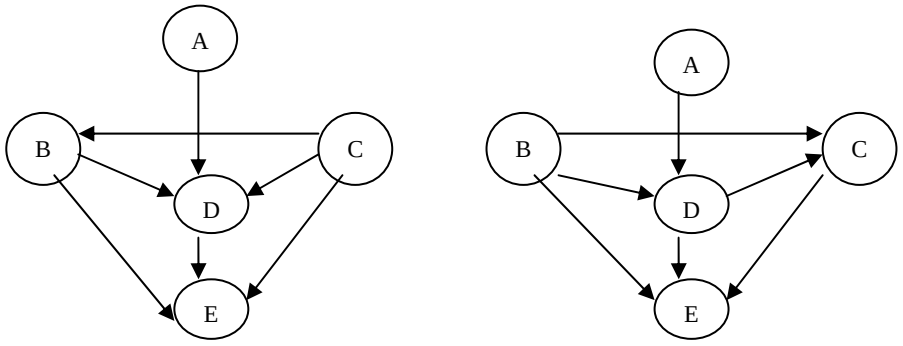
- 性　　别(A) : 男、女
- 学业成绩(B) : 低、中等偏下、中等偏上、高
- 家庭经济(C) : 差、中等偏下、中等偏上、优
- 就业形势(D) : 坏、好
- 是否打算上研究生(E) : 是、否

表4.1是对10000名大学生的统计结果，共计128个数据。 表2中的第1行第1格数据表示A = 男 , B = 低 , C = 差 , D = 坏 , E = 是 的个数为3, 第1行第2格数据表示A = 男 , B = 低 , C = 差 , D = 坏, E = 否 的个数为331,. . . , 依此类推, 在表2 的下半部分(即5～8 行),A的取值都为女。

表4.1 对10000名大学生不同状况的人数统计结果

性别	依据上述原则对大学生不同状况的统计人数															
男	3	331	13	65	11	105	31	62	12	119	38	64	10	67	35	41
	1	241	27	84	7	201	64	95	12	115	93	92	17	79	119	51
	7	165	47	91	6	120	74	110	17	92	148	100	6	42	198	68
	6	46	49	57	5	47	123	90	9	41	224	65	17	17	414	43
女	4	460	39	44	5	312	14	47	8	116	20	35	13	96	28	24
	10	283	29	61	19	236	47	88	15	164	62	85	21	113	72	45
	6	161	36	72	13	193	75	90	12	174	91	100	25	81	142	76
	5	47	40	60	7	74	108	78	13	45	123	87	16	49	348	89

构造贝叶斯网络找出这些变量之间的因果关系用于数据挖掘 ，具体构造网络的过程如下：确定变量为性别（ A ），学业成绩（ B ）， 家庭经济（ C ），就业形势（ D ），是否打算上研究生（ E ）；根据现有的专家知识我们只选择了最有可能的a和b两种网络结构。 它们的区别是学生的就业形势与家庭经济的因果关系不同，学业成绩和家庭经济的因果关系不同.如图4.1所示：



（a） （b）

图4.1 两个最有可能的网络结构a和b

对于本例而言，采用经典的样本统计法计算各个变量的概率参数，如：对于图4.1(a)中，计算 $p(B|C)$，找出 $p(B,C)$ 和 $p(C)$，由 $p(B|C)=p(B,C)/p(C)$ 利用表中的数据，得到B的条件概率分布函数，其他参数可由同样方法求出。经过计算和分析，调整贝叶斯网络的网络结构和各变量的条件概率分布函数，最终得出图4.1(a)为最优的贝叶斯网络结构。

4.1.2 实验结果分析

对于上案例，用决策树方法重新对数据进行分析，得到的结果与贝叶斯网络

方法进行比较，如图 4.2 所示为贝叶斯网络方法与决策树方法应用于案例的学习曲线。

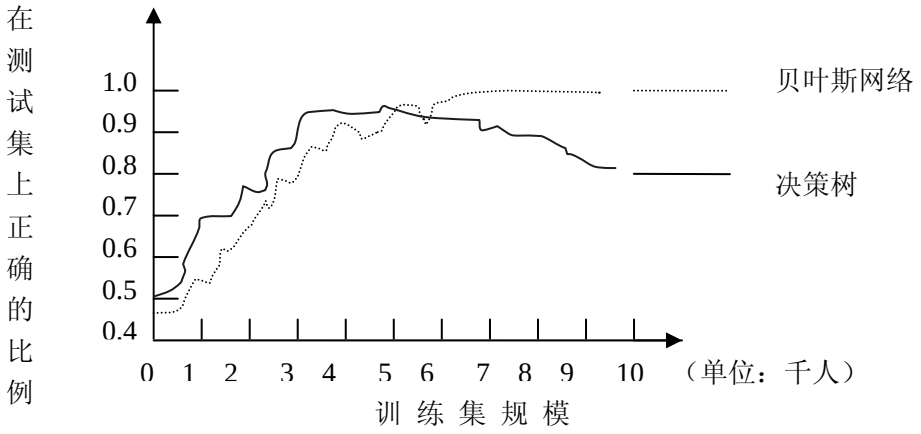


图 4.2 基于上述案例两种不同方法的学习曲线比较

从决策树方法学习曲线来看大约在人数 5000 人之前，在测试集上的比例数值总体是增大的，但超过 5000 人时正确的比例随着人数的增加而减少，表明决策树在数据量较大和数据复杂的情况下，该算法显得力不从心，性能也越来越差。而贝叶斯网络方法学习曲线随着人数的增加，正确的比例在一直加大，当到达 7000 人左右时，比例数值开始趋向于 1，性能越来越好。得到在数据量较大时，贝叶斯网络方法要优于决策树方法。贝叶斯网络具有因果和概率性语义,可以有机地结合先验知识和样本数据，将主观和客观有机地结合起来，表达清晰，因此能更加全面客观地反映数据对象内在的联系与本质，便于实现对数据的挖掘。经过反复验证，该结论具有普遍适用性。

4.2 农户信用等级评定网络模型

过去，对农户一直没有一套比较客观、科学、易操作的贷款办法，贷与不贷，贷多贷少，基本上有信贷人员凭经验来决策，带来了很大的信贷风险。为改变这种局面，农信社相继推出了农户信用等级评定办法，开展了“信用户”、“信用村”的评定工作，一定程度上降低了贷款的风险，但是在信用等级评定的过程中，虽然客户资料的采集工作做的比较完善，但是在等级评定中主观因素占的比例较大，加上影响个人信用的因素很多，存在的误差比例还是很大，对信贷人员的正确放贷造成了一定的难度。信用等级评定的关键和难点在于指标体系的建立和评定模型的选择，针对目前国内研究农户信用等级评定主要是采用传统的信用评分方法，主观因素比重大，影响了等级评定的准确性。基于此，作者建立了基于贝叶斯网络的农户信用等级评定模型。贝叶斯网络的学习机制高效灵活，可以发现潜在有用的模式或关系，实现对数据实例的分类、聚类和预测，找出影响农户信用等级

的决定因素，利于信用等级的评定，很大程度上降低了信贷风险。实验证明，运用贝叶斯网络提高了信用评级评定的准确性，降低了信贷风险，为信贷人员的放贷提供了科学的依据。

4.2.1 案例实验结果

影响农户是否违约的因素有很多，参照国外成熟的信用评级指标体系的框架，并按照我国农村的实际，来构建信用评级指标体系。

农户的信用等级共分 AAA（最优级）、AA（很好）、A（较好）、BBB（一般）、BB（观察）、B（预警）六个等级，将等级变量取值为 1，2，3，4，5，6 共六个状态。考虑影响等级的因素为：

⑴ 性别（男、女）

⑵ 年龄（<20,20～50,50～70,>70）

⑶ 学历(小学及以下、初中、高中、大学及以上)

⑷ 婚姻状况（未婚、已婚、离异或丧偶）

⑸ 农户类型

小农户：即农村中依靠自己的体力和传统的农业经营经验、观念进行农业经营的农户；

销售农户：即主要是指种粮大户、专业户、养殖大户等；

兼业农户：是指家庭成员至少有一人常年或农闲在外工作，或直接从事二、三产业活动，取得收入的农户；

非农农户：即全家人不从事农业生产经营活动，完全依靠从事二、三产业活动取得收入的农业人口。

⑹ 地域因素（原始农业、传统农业、现代农业）

⑺ 家庭劳力数（1 个、2 个、3 个、4 个及以上）

⑻ 年收入（5000 元及以下、5000～10000 元、10000 元以上）

⑼ 职业情况（种植业、养殖业、手工艺、个体经营）

⑽ 资产状况（10000 元及以下、10000～20000 元、20000 元以上）

⑾ 抵质押物（有、无）

⑿ 信誉情况（优、良、差）

⒀ 行业风险（大、中、小）

实验数据来自于山东省某农村信用社普通农民的客户资料（本着为客户保密原则，姓名一律采用代号），将其预处理后变量及其取值如下表 4.2 所示：

表4.2 农户信用等级评定模型的变量及取值

变量名称	性别	年龄	学历	婚姻状况	农户类型	地域因素	家庭劳力	年收入	职业情况	资产状况	抵质押物	信誉状况	行业风险
取值状态	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2	2	2	2	2	2
		3	3	3	3	3	3	3	3	3		3	3
		4	4		4		4		4				

根据表4.2对原始数据进行预处理后，得到标准的样本数据表4.3：

表4.3 预处理后的部分样本数据

姓名	性别	年龄	学历	婚姻状况	农户类型	地域因素	家庭劳力	年收入	职业情况	资产状况	抵质押物	信誉状况	行业风险
a	1	2	2	2	1	2	2	1	1	1	2	1	2
b	1	2	2	2	1	2	2	1	1	1	2	1	2
c	2	2	3	2	3	2	2	1	1	1	2	1	1
d	2	3	1	2	1	2	1	1	1	1	2	3	3
e	2	3	1	2	2	3	4	2	1	2	2	2	2
f	1	2	3	3	1	2	1	1	1	1	2	2	2
g	1	2	3	1	1	2	1	1	2	1	2	2	1
h	2	2	3	1	1	2	1	1	3	1	2	2	3
i	1	1	3	1	1	2	1	1	2	1	1	2	3
j	1	3	2	2	3	3	3	3	4	3	2	1	1
k	1	2	2	2	4	3	3	3	2	3	2	1	1
l	2	2	3	2	1	2	2	1	2	1	1	1	2
m	1	2	3	2	2	2	2	2	1	2	2	1	2
n	2	3	1	3	2	2	1	1	2	1	2	1	2
o	1	4	1	3	1	2	1	1	1	1	2	1	1
p	1	2	1	2	1	3	3	2	1	2	2	2	2
q	1	2	4	2	2	3	4	3	3	3	2	2	2
r	2	2	2	2	4	3	2	2	2	2	2	2	1
s	2	2	3	2	1	3	2	2	4	2	2	2	1

t	1	2	3	2	1	2	2	1	2	1	2	2	2
u	1	2	2	2	2	2	2	1	1	2	2	2	2
v	1	1	1	2	2	3	3	2	1	2	2	1	2
w	1	2	3	2	1	2	2	1	3	2	2	1	2
x	1	2	3	1	1	2	1	1	2	1	2	2	3

通过对数据实例的学习，采用基于互信息的条件独立性检验删除部分边，设门槛值$\varepsilon = 0.03$,根据专家知识和先验信息得到最小无向图，运用MAP和MDL的准则，结合样本数据等，得到相应的农户信用等级评定贝叶斯网络结构模型如图4.3所示。从信用等级评定模型可以看出，农户的学历、信誉状况、年收入、抵押物和行业风险直接影响农户的信用等级，属于主导因素，应给予较大的权重。农户的学历越高，信用等级越高；信誉状况越好，信用等级越高。如果有抵押物，则违约的风险就越小，但从原始数据看，农户贷款有抵押物的很少，基本上都是信用贷款，可以不予过多考虑。从模型中还可以看出，农户在经营副业过程中的行业风险对农户的信用等级也有直接影响，需要信贷人员在评级过程中应全面考察行业风险。此外，农户的资产状况对农户的信用等级没有直接的关联性，应赋予较小的权重。不存在直接或间接关系的因素我们可以不予考虑。

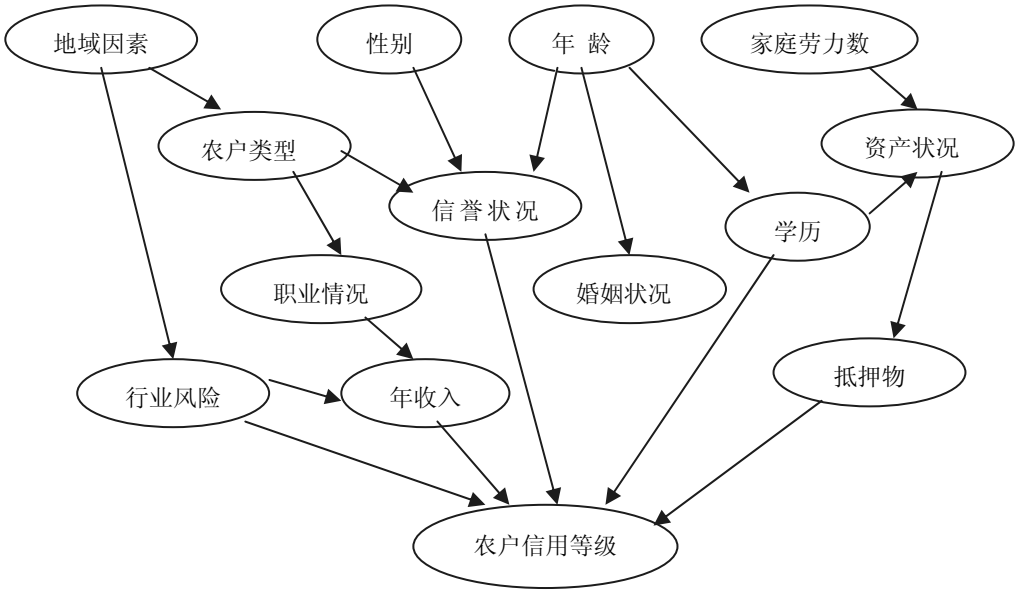


图4.3 基于贝叶斯网络的农户信用评价结构图

4.2.2 实验结果分析

传统的信用评分法，是指在对贷款人的信用等级进行评定以及决定是否发放贷款时，主要依靠商业银行信贷人员进行主观判断打分，从而决定给予农户一定

的信用等级的方法。这种主观评价方法的缺点是相当明显的：(1)主观性比较强，信用等级不准确，对是否放贷及贷款数额形成一定的误导；(2)在对其信用等级不能准确计量的同时，贷款风险控制的成本就会提高，造成违约贷款的数量增加。因此，主要依赖信贷人员判断的信用评定方法，不仅无法对个人信用程度进行准确的计量，而且无法有效地降低单笔贷款的管理成本。下面是运用贝叶斯网络方法与传统的信用评分法的准确率对比如下，新样本来自同一地区另一农村信用社的农户资料。显然，贝叶斯网络方法的运用提高了信用等级评定的准确性。

表4.5 农户信用等级评定准确性对比统计表（训练样本）

方法	检验个数	正确	错误	准确率
贝叶斯网络	100	92	8	92％
传统的信用评分法	100	79	21	79％

表4.6 农户信用等级评定准确性对比统计表（新样本）

方法	检验个数	正确	错误	准确率
贝叶斯网络	100	89	11	90％
传统的信用评分法	100	78	22	78％

4.3 结论

由于贝叶斯网络(Bayesian Networks ,BN)具有诸多的优点，近年来已经成为数据挖掘引人注目的研究方向,将贝叶斯网络应用于数据挖掘当中，能充分挖掘数据的隐藏信息和内在本质,应用前景非常广泛。

将贝叶斯网络方法运用到农户信用等级评定，通过构造模型，找出影响信用等级的主要因素，由于贝叶斯网络的学习机制高效灵活，可以有机地结合先验知识和样本数据，发现潜在有用的模式或关系，避免由于主观因素可能造成的偏见，实现对数据实例的分类、聚类和预测，提高了等级评定的准确性，给农户的信用评级提供了新的方法和可靠的依据，很大程度上降低了信贷风险。由于我国区域经济发展不平衡，尤其农村更是如此。当然，可以将其结果推广到个人信用等级的评定模型建立中。在客户信息的采集过程中，影响信用等级的因素有较大不同；再加上由于采集人的不负责任以及农户的弄虚作假，数据的真实性有了差距，给信用评级的准确性造成一定的偏差，模型的适用性受到一定的限制。

目前，贝叶斯网络不仅在数据挖掘方面，而且在医学、金融、农业、环境科学、软件测试等众多领域显示出了它不可估量的作用。

结束语

常用的数据挖掘方法有许多，贝叶斯网络(Bayesian Networks ,BN)方法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算法研究的热点问题。贝叶斯网络是一种进行不确定性推理和知识表示的有力工具，当与统计方法结合使用时，显示出许多关于数据处理的优势。本文首先介绍了贝叶斯网络和数据挖掘的研究背景和现状，第二章介绍了贝叶斯网络的基本理论，结合信息论的有关知识，讨论了贝叶斯网络中重要的条件独立性关系研究，并学习和研究了贝叶斯网络在完备数据和不完备数据两种情况下的结构和参数学习方法。第三章首先介绍了数据挖掘的相关知识以及目前主流的一些算法以及它们的优缺点，着重总结了贝叶斯网络在数据挖掘中的优势，探讨性的提出了一种启发式的在数据挖掘中利用样本数据构建贝叶斯网络的算法思想，第四章将其运用到实际案例中，并将得到的结果与决策树方法和传统的信用评分方法进行比较，得到了将贝叶斯网络应用于数据挖掘当中，能够充分挖掘数据的隐含信息和内在本质,具备良好地预测能力等优点。此外，在第二个案例中，将贝叶斯网络运用到农户的信用等级评定中，给农户和个人的信用等级评定提供了新的方法和可靠的依据，提高了等级评定的准确性，降低了信贷风险，为信贷人员的放贷提供了科学的依据。

虽然贝叶斯网络在数据挖掘中的研究和应用取得了巨大的成就，但作为一种新的理论，目前还有许多有待解决的问题。

1. 如何利用贝叶斯网络的自学习能力，从试验数据中自动构建贝叶斯网络结构和参数，这样将会增加模型的准确性和完备性。

2. 现有的贝叶斯网络存在的一个局限性是它没有考虑原因节点影响结果节点的滞后时间，所以，有必要引入动态贝叶斯网络，将时间因素引入贝叶斯网络的建模中，使贝叶斯网络的推理与时间相关。

3. 在贝叶斯网络的学习和应用中，如何选取合适的实验数据是较为困难的问题。选择一个选取数据库的标准，对于建立网络模型的准确性和参数学习的有效性都有积极的意义。

4. 如何根据数据和专家知识，进行有效的网络结构学习是NP难题，尤其是数据不完备时，网络结构的学习，现有的算法有的计算复杂度高，有的适用范围较为狭窄，如何寻找更有效的网络学习算法是未来发展的方向。

5. 由于挖掘的内容和知识有限，不是很深入，其应用须进一步拓宽。

随着科学的不断进步和发展，理论和算法也将不断成熟和完善，贝叶斯网络在数据挖掘中的应用前景必将越来越广阔。

总之，在本次论文研究即将告段落的时候，既有一份成功的喜悦，也有对研究中不足之处的遗憾。但毫无疑问的是，我们将在此次研究的基础上再接再厉，在基于贝叶斯网络的数据挖掘应用领域中继续探索。由于时间仓促，加上本人水平和学识有限，论文的研究中还存在很多不完善的地方，缺点和问题在所难免，敬请各位专家、老师和同学批评、指正。

致谢

在攻读硕士学位期间，无论在学习方面，还是生活方面，导师张卓奎教授都给予学生巨大的鼓励和支持，衷心感谢张老师的精心指导和谆谆教诲！从论文的选题到论文的撰写与定稿，都倾注了恩师大量的心血和精力，在此向张老师致以最崇高的敬意！

在本论文的写作过程中，自始至终得到张老师的悉心指导和热情鼓励，攻读近三载，在张老师的无微不至的指导下，无论是在知识更新方面，还是在科学研究创新方面，都使我受益匪浅。张老师渊博的知识，严谨的治学态度，求实的科研作风和朴素的生活习惯，将使我受益终生。上研究生期间，我的成长和进步，还有本文的最终完成，都凝聚着张老师的辛勤汗水，在此表示最真挚的感谢。

同时，特别感谢师母陈慧婵老师在我攻读硕士学位期间对我的关心和照顾，我还要感谢马爱利同学对我的帮助，我们的交流使我得到很多的启发，我深深感谢刘晨华等同学以及我的室友田苗、赵文红、段伟同学对我的帮助，感谢他们在我学习期间给我的鼓励，我取得成绩离不开他们的关心。

最后感谢所有在我攻读硕士学位期间给予帮助和支持的老师和同学，向所有给过我关心和支持的朋友们表示忠心的感谢，并祝他们身体健康、万事如意！

参考文献

[1] 杨莉，孙华昕，朱宏超.基于贝叶斯网络的多目标优化算法.华北电力大学学报[J].2007，34（1）.128-131

[2] 王双成，张明，陈乃激.基于因果语义定向的贝叶斯网络结构学习.计算机工程与应用[J].2007，43（8）.29－31

[3] 田凤占,黄丽,于剑等.包含隐变量的贝叶斯网络增量学习方法.电子学报[J].2005，33（11）.1925－1928

[4] 冀俊忠，阎静，刘椿年.基于I-B&B-MDL的贝叶斯网结构学习改进算法.北京工业大学学报[J].2006，32（5）.437－441

[5] 冀俊忠，刘椿年，江川等.贝叶斯网及其概率推理在智能教学中的应用.北京工业大学学报[J].2002，28（3）.353－357

[6] 周忠宝,董豆豆,周经伦.贝叶斯网络在可靠性分析中的应用.系统工程理论与实践[J].2006（6）.95－100

[7] 冀俊忠，沙志强，刘椿年.贝叶斯网模型在推荐系统中的应用研究.计算机工程[J].2005,31(13).32-34

[8] 田子德.贝叶斯网络在自适应超媒体系统中应用研究.情报科学[J].2006,24(7).1049-1052

[9] 叶跃祥,糜仲春,王宏宇等.基于贝叶斯网络的不确定环境下多属性决策方法.系统工程理论与实践[J].2007(4).107-113

[10] 臧玉卫，王萍，吴玉华.贝叶斯网络在股指期货风险预警中的应用.科学学与科学技术管理[J].2003（10）.122－125

[11] 何盈捷，刘惟一.由Markov网到Bayesian网.计算机研究与发展[J].2002，39（1）.87－98

[12] Agrawal R, Imiclinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases [A]. Proceedings of the 1993 ACM SIGMOD Conference [C]. Washington, DC. 1993, 5:207-216

[13] Agrawal R, Imiclinski T, Swami A. Database Mining: A performance perspective. IEEE Trans. Knowledge and Data Engineering. 1993(5):914-925

[14] Srikant R, Agrawal R. Mining Association Rules in Large Relational Tables. Proceedings of the 1995 Very Large Databases Conference.

[15] Cheung D W, Han J, Ng V, et al. Maintenance of discovered Association rules in large databases: An incremental updating technique. Proceedings of the 1996

International Conference on Data Engineering.

[16] 程继华，施鹏飞，郭建生. 模糊关联规则及挖掘算法. 小型微型计算机系统[J]. 1999，20（4）. 270－274

[17] Srikant R, Agrawal R. Mining Generalized Association Rules. Proceedings of the 1995 Very Large Databases Conference.

[18] Agrawal R ，Srikant R. Fast Algorithm for Mining Association Rules in Large Databases. In Research Report RJ9839, IBM Almaden Research Center, San Jose, CA.1994,6.

[19] De Campos L M,Huete J F.A new approach for learning belief networks using independence criteria[J]. International Journal of Approximate Reasoning, 2000, 24,11-37

[20] Marco Ramoni, Paola Sebastiani, Patameter Estimation in Bayesian Networks from Incomplete Databases. Intelligent Data Analysis,1998,2:139-160

[21] Pang-Ning Tan, Michael Steinbach, and Vipin Kupnar. Introduction to Data Mining [M]. Addison Wesley.2005

[22] P.Spirtes, C.Glymour, R .Scheines. Causation, Prediction and Search. Lecture Notes in Statistics 81. Springer-Verlag.1993

[23] P.Spirtes, C .Glymour, R .Scheines. An Algorithm for Fast Recovery of Sparse Causal Graphs. Social Science Computer Review.1991 (9):62-72

[24] S. Acid, D .Campos, L .M. et al. Approximations of Causal Networks by Poly trees: an Empirical Study. In B. Bouchon-Meunier, editor. Advances in Intelligent Compiting, Lectures Notes in Computer Science 945. Springer-Verlag, 1995. 149-158

[25] J .Cheng, R .Greiner, J. Kelly, D .A. Bell, W .Liu. Learning Bayesian Networks from Data: An Information Theory Based Approach. The Artificial Intelligence Journal. 2002(137):43-90

[26] Chickering D, Heckerman D. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Machine Learning,1997,29: 181-212

[27] Friedman N. The Bayesian structure EM algorithm [A].P roc of the 14th Inf'1 Conf on Uncertainty in Artificial Intelligence [C]. San Francisco: Morgan Kaufmann Publishers, 1997. 125-133

[28] Larranaga P, Poza M , Yuram endi Y, *et al*. Learning Bayesian network structure for the best ordering with genetic algorithms [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1996, 18 (9): 487 -493

[29] German S, German D. Stochastic relaxation, Gibbs distributions and the Bayesian

restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984, 6 (6):721-742

[30] Dempster P, Laird N M, Rubin D B. Maximum Likelihood From Incomplete Data Via the EM Algorithm [J]. Royal Stat. Soc., 1997, 39(l) :1-38

[31] Tom M. Mitchell. Machine Learning [M].McGraw-Hill Companies Inc.2003

[32] Gamez J A, Puerta J M, Searching for the best elimination sequence in Bayesian networks by using ant colony optimization [J]. Pattern Recognition Letters，2002，23(1-3):261-277

[33] Chickering D. Learning equivalence classes of Bayesian network structures. In: Proc of Twelfth Conference on Uncertainty in Artificial Intelligence.1996, 150-15

[34] 蔡自兴,徐光祐.人工智能及其应用(第三版)[M].北京.清华大学出版社.2004

[35] 董琳,邱泉,于晓峰 吴韶群 孙立骏 译.数据挖掘、实用机器学习技术[M].北京.机械工业出版社.2005

[36] 姜哲,金奕江,张敏,杨磊等译.人工智能—一种现代方法（第二版）(Stuart Russell. Peter Norvig. Artificial Intelligence. A Modern Approach)[M].北京.人民邮电出版社.2004

[37] Nils J. Nilsson 著.郑扣根, 庄越挺 译.人工智能[M].北京.机械工业出版社 .2000

[38] David Heckerman. Bayesian Networks for Data Mining. Machine Learning [J]. 1997（3）.213～244

[39] 林士敏,田凤占,陆玉昌.贝叶斯学习、贝叶斯网络与数据采掘.计算机科学[J]，2000，27（10）.69～72

[40] 闫志勇,李明等.贝叶斯网络在自适应教育超媒体中的应用.计算机工程与应用[J].2002(8).217～219

[41] 薛万欣，刘大有等. Bayesian 网中概率参数学习方法.电子学报[J].2003(11).1686～1689

[42] 朱毅峰，吴晶妹，宋玮等.银行信用风险管理[M].北京.中国人民大学出版社.2006

[43] 殷剑峰，王唯翔，程炼等译.高级信用风险分析[M].北京.机械工业出版社.2005

[44] 李剑川,陶俊勇等.基于贝叶斯网络的智能故障诊断方法.中国惯性技术学报[J].2002，10（4）.24－28

[45] 陈心.农民信用等级评定初探－农民信用评定等级,支农信贷实现双赢.浙江

金融[J].2006，4

[46] 黄解军，万幼川，潘和平.贝叶斯网络结构学习及其应用研究.武汉大学学报（信息版）[J].2004，29（4）.315－318

[47] G. F. Cooper, “A simple constraint-based algorithm for efficiently mining observational databases for causal relationships”, *Data Mining Knowledge Discovery*,1997 (1):203–224.

[48] D. Heckerman, “Bayesian networks for data mining”, *Data Mining Knowledge Discovery*, 1997 (1):79–119.

[49] 林士敏，王双成，陆玉昌. Bayesian方法的计算学习机制和问题求解.清华大学学报（自然科学版）[J].2000，40（9）.61－64

[50] 林士敏，田凤占，陆玉昌.贝叶斯网络的建造及其在数据挖掘中的应用.清华大学学报（自然科学版）[J].2001，41（1）.49－52

[51] 羌磊，肖田元，乔桂秀.一种改进的Bayesian网结构学习算法.计算机研究与发展[J].2002，39（10）.1221－1226

[52] 张忠玉，刘惟一，张玉琢. Bayesian网的信息熵.云南大学学报（自然科学版）[J].2002，24（3）.183－185

[53] 胡兆勇，屈梁生.贝叶斯网络的一种仿真算法.系统仿真学报[J],2004,16（2）.286－288

[54] Pearl J. Fusion, propagation, and structuring in belief networks. Artificial intelligence [J].1986, 29(3):241-288

[55] Chow C, Liu C. Approximating discrete probability distribution with dependence trees [J]. IEEE Trans. Inf. Theory, 1968, 14: 462-467

[56] Geiger D, Verma T, Pearl J. Identifying independence in Bayesian networks [J]. Communications of the Association for Computing, 1990, 20:507-534

[57] Wermuth N, Lauritzen S L. Graphical and recuisive models for contingency tables [J]. Biometrika, 1983, 70: 537-552

[58] Cooper G, Herskovits E.A Bayesian method for the induction of Bayesian networks from data [J].Machine Learning, 1992,9: 309-347

[59] Bouckaert R R.Belief networks construction using the minimum description length principle[J].Lecture Notes in Computer Science,1993,747:41-48

[60] Lam W, Bacchus F. Learning Bayesian Belief Network: AN Approach based on the MDL Principle [J]. Computational Intelligence,1994,10 (4):269-293

在读期间的研究成果

在硕士研究生期间,主要从事贝叶斯网络在数据挖掘中的研究和应用,取得的研究成果如下：

[1] 李艳美，张卓奎.基于贝叶斯网络的数据挖掘方法.计算机仿真.已录用.(核心)

[2] 李艳美,张卓奎.基于贝叶斯网络的个人信用等级评定模型.西安电子科技大学2007研究生学术年会理学院数学系论文集.19－23

西安電子科技大學

地址：西安市太白南路2号

邮编：710071

网址：www.xidian.edu.cn