基于 PCA 的贝叶斯网络构造算法研究与应用

摘 要

贝叶斯网络是用来表示变量间概率分布的图形模式，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，具有稳固的数学基础，由于其具有图形化的模型表示形式、局部及分布式的学习机制、直观的推理；适用于表达和分析不确定性和概率性的事物；能够对不完全、不精确或不确定的知识或信息做出有效的推理等特性，而成为目前不确定知识表达和推理领域最有效的模型之一。如何通过有效的方法和算法利用现实数据学习贝叶斯网络，并准确地表达蕴含在数据中有价值的信息是目前研究的热点和难点。本文采用基于信息论的方法进行贝叶斯网络的结构学习，并针对其当节点集越大，计算效率越低的缺点采用 PCA 降维，减少节点集的数量，提高算法的效率，主要工作如下：

1、用模糊聚类对连续数据或混合数据进行离散化；对数据集用 PCA 主元分析算法进行降维，减少其中节点的个数；

2、运用 Gibbs 抽样算法对数据集中的缺失数据进行补充，用基于信息论的方法学习贝叶斯网络结构；

3、用分类实验验证基于 PCA 的贝叶斯网络分类器的准确率及算法效率，并对乙烯生产中不同生产规模或不同技术的能耗及物耗相关数据进行贝叶斯数据融合，得到的结果对乙烯生产中能耗物耗水平的评价有一定的参考价值。

关键词：贝叶斯网络，PCA，Gibbs 抽样，信息论，数据融合

A Study and Application of Learning Bayesian Network from Data Approach Based on PCA

ABSTRACT

The Bayesian belief network is a powerful knowledge representation and reasoning tool under conditions of uncertainty. A Bayesian belief network is a directed acyclic graph with a conditional probability distribution for each node, With a solid math foundation. Bayesian networks is one of the most efficient models in the fields of uncertain knowledge expression and inference .It has the following characteristics: the expression form of graph model, partial and distributed study mechanism and directly perceived inference; applicable in expressing and analyzing uncertain and probability things and efficiently reasoning partial, inaccurate and uncertain knowledge or information. In the field of graph model and data mining, the central issue and difficult point is how to learn Bayesian networks and to accurately express valuable information in the data through the efficient methods and algorithm. This paper using the algorithm of learning bayesian network from data on information theory and according to the disadvantage of it, using PCA to reduce the dimensionality of the database amended by Gibbs sampling to cut down the number of nodes of data and improve efficiency of learning

bayesian network.

The main work are as follows:

1、Using fuzzy clustering discretize the continuous attribute and using PCA to reduce the dimensionality of the database to cut down the number of nodes of data;

2、Amending missing data by Gibbs sampling, Learning the structure of bayesian network using method on information theory from data;

3、To verity the accuracy and efficiency of the algrithm of learning bayesian network from data, using bayesian network to classify data; and using bayes data fusion to fuse the data from different installation of ethylene production, the results have a certain extent reference value.

KEY WORDS: bayesian network, PCA, Gibbs sampling, information theory，data fusion

北京化工大学位论文原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：刘晓浩 日期：2009.6.1

关于论文使用授权的说明

学位论文作者完全了解北京化工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北京化工大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位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、汇编学位论文。

保密论文注释：本学位论文属于保密范围，在 2 年解密后适用本授权书。非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名：刘晓浩 日期：2009.6.1

导师签名：[illegible] 日期：2009.6.1

第一章 绪论

1.1 贝叶斯概述

随着信息技术的迅猛发展，各种信息系统和专家系统成为辅助人类决策的重要方法，人工智能技术的推广与应用进一步提高了人类的决策水平和决策效率。如人工神经网络、遗传算法、粗糙集、决策树等技术，在自然科学和社会科学领域都产生了深刻的影响[1]。

人工智能的核心问题之一是如何表达已有知识以及如何应用已有知识进行分析处理或推理，以得到新的知识。由于现实生活中存在大量不确定性知识，对不确定性知识进行表达和推理更是显得尤为重要。事实上，绝大部分智能行为都包含不确定性信息，但处理这些不确定信息将涉及到多种理论和技术，其实现一般比较困难。由于不确定性知识的研究具有非常重大的现实意义，目前它是国际上的一个非常重要的研究热点。

贝叶斯（Reverend Thomas Bayes 1702-1761）学派奠基性的工作是贝叶斯的论文“关于几率性问题求解的评论”。或许因为感觉到自己的学说还有不完善的地方，这一论文在他生前并没有发表，而是在死后，由他的朋友发表的。著名的数学家拉普拉斯（Laplace, P.S）用贝叶斯的方法导出了重要的“相继律”，贝叶斯的方法和理论逐渐被人理解和重视起来。但由于当时贝叶斯方法在理论和实际应用中还存在很多不完善的地方，因而在十九世纪并未被普遍接受。二十世纪初，意大利的菲纳特（B. de Finetti）及其英国的杰弗莱（Jeffreys, H.）都对贝叶斯学派的理论作出了重要的贡献。第二次世界大战后，瓦尔德（Wald, A.）提出了统计的决策理论，在这一理论中，贝叶斯解占有重要的地位；信息论的发展也对贝叶斯学派做出了新的贡献。1958 年英国最悠久的统计杂志 Biometrika 全文重新刊登了贝叶斯的论文，20 世纪 50 年代，以罗宾斯（Robbins,H.）为代表的部分学者提出了用经验贝叶斯方法和经典方法相结合，引起统计界的广泛注意，这一方法很快就显示出它的优点，成为很活跃的一个方向。在这里值得一提的是，八十年代以后，人工智能的发展，尤其是机器学习、数据挖掘的兴起，为贝叶斯理论的发展和应用提供了更为广阔的空间。尽管对于贝叶斯学派哲学上的观点还存在很多异议，但它的思想和方法在社会生活和生产实践中得到越来越广泛的应用却是不争的事实。尤其是近年来，贝叶斯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等成为当前数据挖掘众多方法中最为引人注目的焦点之一[2]。

20世纪90年代之前，贝叶斯网络的研究主要集中在建立贝叶斯网络基础理论体系和不确定性推理[3]方面。该阶段学习贝叶斯网络主要依赖于专家知识。贝叶斯网络结构模型的构建一般是由相关领域的专家根据事物间的因果关系确定的。1986 年 Pearl 首次在专家系统中引进了贝叶斯网。1988 年 Pearl 明确指出影响图中没有决策节点和结果节点就是贝叶斯网，指出 Bayesian 网或许是概率推理中最普及的模型。1989 年 Andreassen 使用 Bayesian 网建造了专家系统 MUNIN（Muscle and Nerve Inference Network）。Shafer1990年指出 Bayesian 网目前已经成为公认的表示概率知识的系统。针对一般的 Bayesian 网，Cooper 证明了概率值的传播计算问题是 NP 难题。Bayesian 网方法由于其理论上的严格性和一致性，以及有效的局部计算机制和直观的图形化知识表达，已经成为人工智能领域的研究热点。1995年，Heckermen等研究者使用贝叶斯方法进行贝叶斯网络学习[4,5]，并把贝叶斯网络用于数据挖掘。这一阶段主要研究如何根据数据和专家知识建立贝叶斯网络，相继出现了许多经典的贝叶斯网络学习算法。

1.1.1 贝叶斯网络结构学习算法

贝叶斯网络学习指的是通过分析数据获得贝叶斯网的过程，它包括结构学习和参数学习两种情况。参数学习指的是已知网络结构，确定网络参数的问题；结构学习则是既要确定网络结构，又要确定网络参数。参数学习假设已知变量间的定性关系，通过数据分析揭示变量间的定量关系；而结构学习则是要同时揭示变量间的定性和定量关系[6]。因此贝叶斯网络结构学习是贝叶斯网络学习的核心。由于仅仅依赖专家的领域知识构建贝叶斯网络结构具有一定的不足和局限性，所以充分利用数据自动构建贝叶斯网络结构已经成为贝叶斯研究领域的重点。

20世纪90年代以来，出现了许多从数据中学习贝叶斯网络结构的算法。这些方法可大致分为两类，一类是基于打分搜索的学习方法，一类是基于条件独立性测试的学习方法。

（1）基于打分搜索的方法，例如 K2 算法、爬山法、SEM 算法等。这些算法首先随机初始化一个网络，根据某种评分函数，给当前网络打分；然后对网络进行局部扰动，对扰动后的网络再评分，保留得评分高的网络，这样循环进行，直到局部扰动后的网络评分变化小于阈值为止。这种方法理论上是可以得到一个最适合当前数据的网络结构的，但实际中，如果要学习一个具有 n 个结点的网络，就需要对 $n!$ 种网络结构进行打分计算，已经证明这是一个 NP 难题。这类算法在实现时可以通过限定节点的个数、假设节点有序、根据专家知识采用某些启发式的搜索规则等来使得打分搜索方法变得可行。

（2）基于依赖分析的方法。一个 Bayesian 网络结构本质上就是表示一个系统中的属性之间的某些依赖关系(有向无环图)，基于这种理论基础，可通过首先发现系统中属性之间的某些依赖关系，而后根据这些依赖关系来建立 Bayesian 网。常用的基于依赖分析的方法有：用 Polytree 来表示概率网的方法、从完全图删除边的方法等。这种方法需要有完整的样本数据集；其次它需要进行指数级的 CI 测试，以发现依赖关系，当节点集较大时，它的计算效率低，所以大多数此类算法都假设结点有序；但这种假设可能会影响最后学习到的网络结构的正确性。对于稀疏网络和具有较大样本数据集的系统来说，这种方法是非常有效的[7]。

1.1.2 国内外研究现状

贝叶斯网络是近 20 年提出和发展起来的，然而，贝叶斯网络的理论基础即贝叶斯定理起源于 18 世纪英国牧师 Tomas Bayes 的一篇论文《论机会学说中的一个问题》。在 19 世纪，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，在统计学中形成一个影响较大的贝叶斯学派。

目前阶段，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。结构学习最早的研究开始于对树——最简单的图类研究。Chow & Liu 于 1968 年介绍了一种根据数据集恢复简单树形结构的方法。Rebane & Pearl 1987 年将 Chow & Liu 的因果树学习发展成为多树结构学习，即单连接网络。然而，如果初始数据不是来源于多树分布，则结构学习的精度将很差，因此在推理应用中是不可靠的。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。如果每个独立关系在网络及其分布中同时存在，那么学习的网络结构就是概率分布的 I-map。对于一般的多连接网络，其拓扑关系就是有向无环图（DAG），基于条件独立性检验的算法可以实现该类网络的学习。Wermuth & Lauritzen 于 1983 年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。1990 年，Spirtes 等提出了不需要变量顺序即可发现有向无环图的算法，这些算法要求学习对象的基本分布是 DAG 同构的。

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 1992 年提出将贝叶斯最大后验概率方法用于多联接网络的结构学习。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。许多学者对结构学习的贝叶斯方法也提出了不同的形式。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。最小描述长度准则最早由 Rissanen 于 1978 年作为统计模型的一个新准则正式提出。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。1993 年，Bouckaert 等进行了 MDL 准则在贝叶斯网络学习中的应用研究。近年来，不少学者尝试了对不完备数据和隐藏变量的结构学习。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 算法。Chickering 于 1996 年将结构搜索空间转化为等价类空间来实现网络结构评估。1998 年，Luis & Miriam 提出一种交互式的学习方法；Campos & Huete 于 2000 年提出一种基于独立性准则的方法。

Friedman 从理论上证明，基于评分的结构学习在分类中可能导致较坏的分类性能。同时，基于评分的方法在实际应用时，复杂性较高，而基于独立性检验的方法从原理上更接近于贝叶斯网的语义特性，在实际中获得较好的效果。基于独立性检验的网络结构学习的一些典型算法包括：Wermuth-Lauritzen 算法、Boundary-DAG 算法、SRA 算法、Constructor 算法等。

动态贝叶斯网络（DBN）表示的是制约动态随机过程中轨迹可能发生变化的概率分布，是复杂随机过程的一种压缩表示。与 BN 一样，DBN 的学习也分为参数学习和结构学习。目前已经有很多 DBN 参数学习的算法，它们一般都能够处理隐藏变量和缺值数据。而 DBN 的结构学习，尤其是有隐藏变量的 DBN 结构学习依然是目前研究的热点之一。1998 年，Friedman 将 SEM 算法扩展到 DBN 的结构学习中，用于从带隐藏变量和缺值的数据中学习 DBN 的结构。给定一个随机变量集随时间的部分观测，其算法能构造一个与观测拟合得很好的 DBN。

利用信息几何理论研究贝叶斯网络是目前较新的研究方向。信息几何是采用（Riemann 流形上的）微分几何方法来研究统计问题的理论。自 1975 年 Efron 首先在统计学中采用微分几何方法以来，许多统计学家在这方面进行了大量的工作。S.Amari 于 1985 年提出了一个单参数的联络：α 一联络，利用它研究了指数簇分布和曲指数簇分布模型的性质，并对其在神经网络方面的应用做了大量研究。1995 年前后，Haiyu Zhu 等人把贝叶斯理论与信息几何理论结合起来，得到了不少有价值的成果。

1.1.3 贝叶斯网络应用现状

国内近几年出现了许多关于使用贝叶斯网络解决实际问题的研究。清华大学电机系的曹冬明等[8]利用贝叶斯网络技术进行故障定位，西安电子科技大学计算

机学院的李伟生等[9]将贝叶斯网络用于规划识别，上海交通大学电子信息学院的邓勇等[10]将贝叶斯网络用于模型诊断，上海大学自动化学院的李明等[11]将贝叶斯网络用于模型诊断串行译码，霍利民、朱永利使用贝叶斯网络进行电力系统可靠性评估，欧海涛、张卫东等对贝叶斯学习的城市交通多智能体系统进行了研究等等。

国外近十年中，贝叶斯网络在许多领域也得到了广泛地应用，相继出现了大量的应用贝叶斯网络解决实际问题的应用系统和文献。例如基于贝叶斯网络的微软公司的办公智能帮助系统广泛应用于微软的 Office97 和 Office2000 系统中；1993 年美国海军实验室应用贝叶斯网络开发出了用于识别船只的系统；PATHFINDER 系统[12]用来帮助进行“淋巴结点”症状的诊断分析，在应用中具有良好的性能；CPCSBN 远程医疗系统[13]有 448 个节点和 908 条弧，优于世界上主要的远程医疗诊断分析方法；ALARM 网[14]，这个贝叶斯网络具有 37 个节点和 46 条边，它描述了在医院的手术室中潜在的细菌问题，常被作为贝叶斯网中结构学习算法检验的标准网；Heckerman[15]将贝叶斯网络成功地用于软件智能化；Hudson 等[16,17]将贝叶斯网络用于模拟军事对抗和预测；Alberola 等[18]把贝叶斯网络用于人类学习中的问题解决研究；Rodrigues 等[19]用贝叶斯网络与制造系统控制相结合；同时贝叶斯网络在 DNA 分析[20]、病理分析[21,22]、信号检测[23]、系统可靠性分析[24]、金融风险分析[25,26]、人类学习机制模拟[27]、图像分析和语音识别[28]、进化计算的优化[29]、软件测试[30,31]、工业废水处理[32]等方面均得到了成功的应用[33]。

1.2 主元分析（PCA）概述

主元分析（Principal Component Analysis, PCA）是工业监控系统中使用得最为广泛的多元统计分析方法。PCA 是一种根据已获取数据的方差进行的最优降维表示，它解决了变量间的相关性[34,35]。对某些过程来说，数据中的大部分信息仅用二维或三维即能获取[36]，主要的过程变化可以用单张图显示出来。无论在低维空间中需要多少维度，使用 T^2 和 SPE 统计量控制图都可以将所有维度的信息表达出来。这些控制图可以帮助操作员和工程师们解释过程数据的重要走向[37]。从 20 世纪 80 年代起，PCA 技术开始应用于工业过程的监测，利用控制图等简单的工具实现初步的诊断功能。90 年代以来，学术界和工业界都对 PCA 的应用进行了深入研究[38,39]，一些公司也介绍了 PCA 对工厂数据的应用[40, 41]，而另一些研究则是基于计算机过程仿真得到的数据而进行的[42-44]。

从 PCA 技术出现发展到今天，在传统(线性、静态)PCA 统计模型的基础上，

形成了多种扩展的和改进的 PCA 统计建模技术，如非线性 PCA、动态 PCA、多尺度 PCA、间歇生产 PCA、多工况及自适应 PCA 等：

（1） 非线性 PCA（Nonlinear PCA）

PCA 是一种线性降维技术，它忽略了过程数据中可能存在的非线性。然而工业过程本质是非线性的。因此，在某些情况下，用于过程监测的非线性方法与线性方法相比表现出更好的性能。Kramer[45]通过运用自适应神经元网络，把 PCA 扩展到了非线性情况。Dong 和 McAvoy[46]研究了一种基于主曲线和神经元网络的、能产生独立主元的非线性 PCA 方法。参考文献[47]对三种用于过程监控的神经元网络方法进行了比较。然而由于被监测变量之间是否为线性关系与 PCA 监测模型的有效性并无直接联系，非线性 PCA 监测方法并未深入系统地得到发展，近年来已少有报道。

（2） 动态 PCA（Dynamic PCA,PCA）

传统 PCA 监控方法都隐含地假定：一个时刻的观测值对于前面时刻的观测值来说是统计独立的。对于典型的工业过程，这一假定只对长采样间隔是有效的。这表明需要一种考虑到数据中序列相关性、用来实现快速采样时间的过程监控方法。处理这个问题的方法之一是用靠前的 h 个观测样本对每个观测向量进行扩充。这种方法称为动态 PCA。通过检验 PCA 和 DPCA，求解田纳西-伊斯曼问题证实了 DPCA 比 PCA 更适于求解相关数据检测故障[48]。另一方面，在实践中已经证明，当有足够的数据来表示过程的全部正常变化时，数据中的序列相关性并不会影响静态 PCA 统计方法的有效性[49]。

（3） 多尺度 PCA（Multiscale PCA）

工业数据本质上具有多尺度特性，反映了不同生产工况和设备状况下的信息。多尺度 PCA 首先利用小波技术，将过程数据进行多尺度分解，以获得不同层次下的过程信息[50]。由于小波系统的正交性，在不同尺度下的分解系数是相互不关联的，并且同一尺度下的系数也互不关联[51]。多尺度 PCA 分别对小波分解后不同层次下的过程信息进行 PCA 建模和监测，最后对总的重构信息进行 PCA 监测。一般而言，多尺度 PCA 更适合刻画生产进行时的状况。

（4） 多向 PCA（Multiway PCA）

对于制药、生化反应和高分子聚合反应等高附加值的生产，一般均以间歇方式进行。在间歇过程中所采集的数据通常可以表示为一个三维的立体数据块。多向 PCA 将这个立体数据块的每一个水平的或垂直的切片拼接成一个矩阵，然后对这个矩阵进行主元分析[52-54]。多向主元分析可以用来

改进过程的操作以及用来理解引起过程不同的主要因素。

PCA 的其他工业应用包括数据协调与粗差检测[57]、软测量建模、过程控制，以及产品设计、工况分析、数据挖掘、动态辨识建模等。近几年来，越来越多的 PCA 及其衍生方法被报道成功应用于工业过程。

第二章 数据预处理

存在不完整的、含噪声的和不一致的数据是现实世界大型的数据库或数据仓库的共同特点。

不完整数据的出现可能有多种原因：有些感兴趣的属性不可用、相关记录可能是错误的、由于设备故障而没有记录等。数据含噪声的原因包括：收集数据的设备出故障、在数据输入时人或计算机出现错误、数据传输中的错误等[58]。

如何预处理数据才能提高数据质量，从而提高从数据中学习到知识结果的质量？怎样预处理数据才能使得学习过程更加有效、更加容易？这就需要用到数据预处理技术。

2.1 数据预处理的内容

数据预处理的主要任务如下：

（1） 数据清理：填写空缺值，平滑噪声数据，识别，删除孤立点，解决不一致性；

（2） 数据集成：集成多个数据库，数据立方体，文件；

（3） 数据变换：规范化（消除冗余属性）和聚集（数据汇总），将数据从一个较大的子空间投影到一个较小的子空间；

（4） 数据归约：得到数据集的压缩表示，量小，但可以得到相近或相同的结果；

（5） 数据离散化：数据规约的一部分，通过概念分层和数据的离散化来规约数据，对数字型数据比较重要。

2.1.1 数据清洗

（1）处理空缺值的方法如下：

- ✓ 忽略元组；
- ✓ 人工填写空缺值；
- ✓ 使用一个全局变量填充空缺值；
- ✓ 使用属性的平均值填充空缺值；
- ✓ 使用与给定元组属于同一类的所有样本的平均值；
- ✓ 使用最可能的值填充空缺值，如用 Bayesian 公式或判定树等基于推理的

方法。

（2）处理噪声数据：噪声是一个测量变量中的随机错误或偏差，常用的方法如下：

- ✓ 分箱（binning）（等深或等宽分箱）
  首先，排序数据，并将它们分到等深的箱中；然后可以按箱的平均值平滑、按箱中值平滑或按箱的边界值平滑。
- ✓ 聚类：检测并且去除孤立点。
- ✓ 计算机和人工检查结合：计算机检测可疑数据，然后对它们进行人工判断。
- ✓ 回归：通过让数据适应回归函数来平滑数据，对连续的数字型数据效果较好。

2.1.2 数据集成

数据集成指将多个数据源中的数据整合到一个一致的存储中。

数据集成的主要内容包括模式集成、实体识别和检测并解决数据值的冲突。模式集成是指整合不同数据源中的元数据；实体识别问题则是指匹配来自不同数据源的现实世界中相同的实体（人工干预或利用字段的元信息，比较字段的描述性元信息，看它们是否相同）；检测并解决数据值的冲突是指检测并处理来自不同数据源的现实世界中的同一实体的不同的属性值。

集成多个数据库时往往会出现冗余数据，主要原因是同一属性在不同的数据库中会有不同的字段名、一个属性可以由另外一个表导出等。有些冗余可以通过检测各个属性之间的相关性被相关分析检测到，所以事先根据数据的元数据或相关性分析对其进行预处理，就能够减少或避免结果数据中的冗余与不一致性，提高数据的质量。

2.1.3 数据变换

数据变换的主要方法如下：

（1）平滑：去除数据中的噪声；

（2）聚集：数据汇总，数据立方体的构建，数据立方体的计算/物化(一个数据立方体在方体的最底层叫基本方体，基本方体就是已知存在的数据，对现有的数据按照不同维度进行汇总就可以得到不同层次的方体，所有的方体联合起来叫做一个方体的格，也叫数据立方体。数据立方体中所涉及到的计算就是汇总)；

（3）数据概化：沿概念分层向上汇总，数据立方体的不同维之间可能存在着一个概念分层的关系；

（4）规范化：将数据按比例缩放，使这些数据落入到一个较小的特定的区间之内。方法有：

- ✓ 最小-最大规范化
- ✓ Z-score 规范化
- ✓ 小数定标规范化

（5）属性的构造：通过现有属性构造新的属性，并添加到属性集中。

2.1.4 数据归约

数据归约的作用是得到数据集的归约表示，它比数据集本身小得多，但却可以产生相同的（或几乎相同的）分析结果。

数据归约的策略有以下三种：

（1）维归约

维规约是用来检测或删除不相关的或基本不相关的属性或冗余属性或维的，用以减少数据量。

属性子集的选择是找出最小属性集，使得数据类的概念分布尽可能的接近使用所有属性的原分布，把不相关的属性全部删除，可以减少出现在发现模式上的属性的数目，使得模式便于理解。主要方法有：启发式的（探索式的 try and error）方法，该方法包括逐步向前选择（从空属性集开始，每次选择都选择当前属性集中最符合的目标（最好的属性）加到当前的属性集中，这样逐步的向前选择，把有用的属性一个一个的添加进来）；逐步向后删除（从属性全集开始，每次删除还在当前属性集中的最不适合的那个属性（最坏的属性），这样一个一个的删除，最后留下来的就是相关的属性）；向前选择和向后删除相结合（每次选择一个最好的属性，并且删除一个最坏的属性）。

（2）数据压缩：

使用一些编码机制来压缩数据集。无损压缩（可以根据压缩之后的数据完整的构造出压缩之前的数据，wrar. zip 等，如字符串压缩）和有损压缩（无法通过压缩之后的数据来完整的构造出压缩之前的数据，如音频/视频压缩，有时可以在不解压缩整体数据的情况下，重构某个片段，主要应用于流媒体传输）。

有损数据压缩的方法主要有小波变换和主成分分析。

（3）数值归约：

使用较小的，替代的数据来估计、替换表示原数据（用参数模型），通过选择

替代的，较小的数据表示形式来减少数据量。方法主要有：

- ✓ 有参方法：使用一个参数模型来估计数据，最后只要存储参数即可，有线性回归方法、多元回归、对数线性模型（近似离散的多维数据概率分布））和无参方法（直方图（将某属性的数据划分为不相交的子集或桶，桶中放置该值的出现频率，其中桶和属性值的划分规则有：等深、等宽、V-最优和 MaxDiff）。
- ✓ 聚类：将数据集划分为聚类，然后通过聚类来表示数据集，如果数据可以组成各种不同的聚类，该技术非常有效，反之如果数据界线模糊，则该方法无效。数据可以分层聚类，并被存储在多层索引树中。
- ✓ 选样：允许用数据的较小随机样本（子集）表示大的数据集。对数据集 D 的样本选择方法有简单随机选择 n 个样本、不放回（由 D 的 N 个元组中抽取 n 个样本）、简单随机选择 n 个样本、回放（由 D 的 N 个元组中抽取 n 个样本，元组被抽取后将被回放，同一元组可能再次被抽取到）、聚类选样（聚类分析和简单随机选样的结合，D 中元组被分入到 M 个互不相交的聚类中，可以在其中的 m 个聚类上进行简单随机选样，$m < M$ ）、分层选样（ D 被划分为互不相交的层，则可通过对每一层的简单随机选样得到 D 的分层选样）。

2.2 数据离散化

离散化是指将连续属性的范围划分为区间，以减少所必需处理的数据的量。主要应用于以下三类数据：名称型（无序集合中的值），序数（有序集合中的值），连续值（实数）。

离散化可以有效的规约数据（基于判定树的分类挖掘）。离散化是通过将属性域划分为区间，减少给定连续属性值的个数，区间的标号可以代替实际的数据值。

数值型数据离散化的主要方法如下：

（1）分箱（binning）：分箱技术递归的用于结果划分，可以产生概念分层；

（2）直方图分析（histogram）：直方图分析方法递归的应用于每一部分，可以自动产生多级概念分层；

（3）聚类分析：将数据划分成簇，每个簇形成同一概念层上的一个节点，一个簇可再分成多个子簇，形成子节点；

（4）基于熵的离散化；

（5）通过自然划分分段：将数值区域划分为相对一致的、易于阅读的、看上

去更直观或自然的区间。

自然划分的 3-4-5 规则：

- ✓ 如果一个区间最高有效位上包含 3，6，7 或 9 个不同的值就将该区间划分为 3 个等宽子区间；如果一个区间最高有效位上包含 2，4 或 8 个不同的值，就将该区间划分为 4 个等宽的子区间；
- ✓ 如果一个区间最高有效位上包含 1，5 或 10 个不同的值，就将该区间划分为 5 个等宽的子区间；再将该规则递归的应用于每个子区间。（对于数据集中出现最大值和最小值的极端分布，为避免上述方法 出现的结果扭曲，可以在顶层分段时，选用一个大部分的概率空间 5%-95%）。

由于直接在具有连续属性的数据集上建立贝叶斯网络模型计算复杂度大，且从连续数据中很难正确学习到变量间的关系，但现实中的大部分数据中都包含有连续属性，因此首先需要对数据进行离散化的处理。

2.2.1 离散化方法分类

离散化方法依据不同的需求沿着不同的主线发展至今，目前已存在很多不同的分类体系。不同的分类体系强调离散化方法间区别的不同方面。主要的分类体系有有监督的和无监督的、动态的和静态的、全局的和局部的、分裂式的（从上至下）和合并式的（从下至上）、单变量的和多变量的以及直接的和增量式的。

根据离散化方法是否使用数据集的类信息，离散化方法可以分为有监督的和无监督的。有监督的离散化方法使用类信息，而无监督的离散化方法不使用类信息。有监督的离散化方法又分为建立在错误率基础上的、建立在熵值基础上的或者建立在统计信息基础上的[60,61]。早期的等宽、等频的离散化方法是无监督方法的典型代表，连续的区间根据使用者给定的宽度或频数划分成小的区间。无监督方法的缺陷在于它对分布不均匀的数据不适用，对异常点比较敏感。为了克服无监督的离散化方法的这些缺陷，使用类信息来进行离散化的有监督的离散化方法逐渐发展起来。

离散化方法也常以动态或静态的分类方法来区分。动态的离散化方法就是在建立分类模型的同时对连续特征进行离散化，如有名的 C4.5；静态的离散化方法就是在进行分类之前完成离散化处理。

根据离散化过程是否针对整个训练数据空间，离散化方法又可分为全局的和局部的。全局的离散化方法使用所有的实例，而局部的离散化方法只是用一部分的实例。

离散化方法还可分为从上至下的和从下至上的，也可称为分裂式的和合并式的。分裂的离散化方法起始的分裂点列表是空的，通过离散化过程逐渐往列表中

加入分裂点，而合并的离散化方法则是将所有的连续值都看作可能的分裂点，再逐渐合并相邻区域的值形成区间。

单变量的离散化方法是指一次只对数据集的一个特征进行离散化，而多变量的离散化是同时考虑数据集的多个特征及其相互关联关系进行离散化，需要考虑更多的因素，算法更加复杂。

另外一种离散化方法的分类是直接式的和增量式的。直接式的离散化方法是根据额外给定的参数（离散化所需得到的区间数等）一次性形成所有的分裂点，而增量式的离散化方法是根据某个准则逐渐的将离散化结果进行改进，直到满足准则的停止条件为止。

Liu Huan和Hussian在其文Discretization: An Enabling Technique. Data Mining and Knowledge Discovery中提出了离散化方法的一个层次框架，主要根据合并分裂及有无监督的分类体系来划分离散化方法，比较全面的概括了离散化方法的分类及各类别中的代表离散化方法，层次框架如图2-1所示。

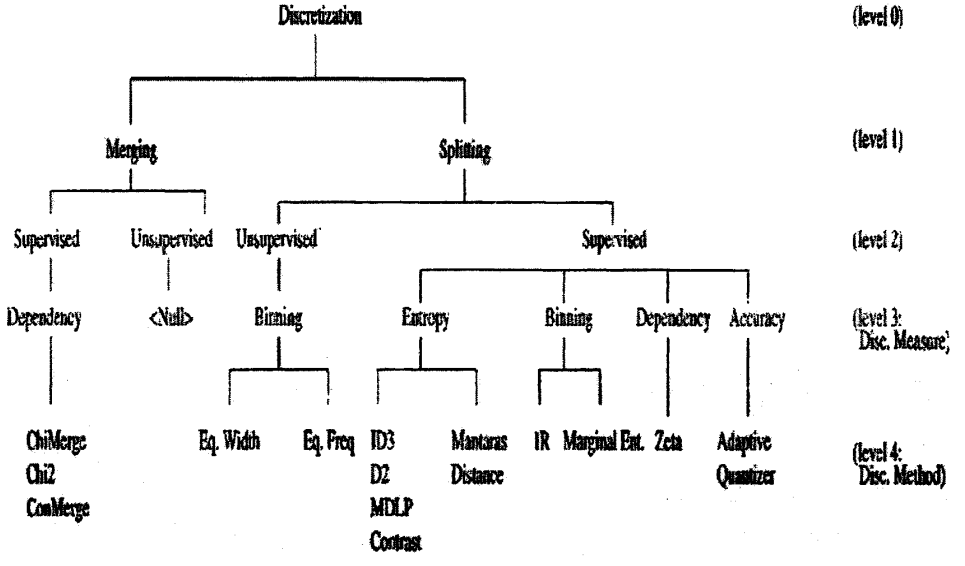

图 2-1 离散化方法层次框架

Fig.2-1 Level framework of the method of discretization

2.2.2 离散化方法

离散化主要是用于分类或者关联分析中的特征离散化。通常来说，离散化结果的好坏依赖于采用的离散化算法，以及需要考虑的其他特征。

一般来说，现在的离散化算法都是考虑单个特征进行离散化，而不考虑对多个特征的关系进行分析同时进行离散化。目前已存在很多对连续特征进行离散化的方法，这些方法既可以是单一算法的，也可以是合成的。单一的算法在完成离

散化的过程中不会用到其他的离散化方法，而合成的方法在离散化的过程中则会同时用到多种离散化算法的思想，是建立在多种单一算法的基础之上的。

（1）无监督的方法

✓ 分箱法

分箱法包括等宽分箱法和等频分箱法，它们是基本的离散化算法。

分箱的方法是基于箱的指定个数自顶向下的分裂技术，在离散化的过程中不使用类信息，属于无监督的离散化方法。等宽分箱法，就是根据箱的个数得出固定的宽度，使得分到每个箱中的数据的宽度是相等的。等频分箱法是使得分到每个箱中的数据的个数是相同的。在等宽或等频划分后，可用箱中的中位数或者平均值替换箱中的每个值，实现特征的离散化。

这两种方法简单，易于操作，但都需要人为地规定划分区间的个数这个参数。同时，使用等宽法的缺点在于它对异常点比较敏感，倾向于不均匀地把实例分布到各个箱中，有些箱中包括许多实例，而另外一些箱中又一个实例都没有。这样会严重地损坏特征建立好的决策结构的能力。而等频的方法虽然避免了上述问题的产生，却可能将具有相同类标签的相同特征值分入不同的箱中以满足箱中数据的固定个数的条件。

可以在离散化前首先设定某个阈值将异常数据移除来解决等宽法对异常点敏感的问题。针对等频法易将具有相同类标签相同特征值的实例分入不同箱中的问题，可用的解决方法是先用等频法将特征值进行分箱，然后对各个相邻分箱的边界值进行调整，使得相同的值可以被分入同一个箱中。

✓ 根据直观划分离散化

对于实际数据的离散化，用户希望离散后得到的区间是相对一致、易于阅读、看上去直观或者自然的区间。根据直观划分的离散化方法使用 3-4-5 规则可以将特征的数值区间划分为相对一致和自然的区间。

一般地，3-4-5 规则根据最高有效位的取值范围，递归逐层地将给定的数据区域划分为 3、4 或者 5 个相对等宽的区间。现实世界中的数据常包含特别大的正的和负的离群值，基于最小和最大数据值的自顶向下的离散化方法可能导致扭曲的结果。而根据直观划分离散化的方法则首先将数据集中的数值得大多数的数值区间（如，第 5 个百分位数到第 95 个百分位数）挑出来作为顶层离散化区间，按照 3-4-5 规则进行离散化，再将超出顶层离散化区间的特别高的和特别低的值用类似的方法单独处理成不同的区间。

这种方法的优点主要在于它适用于实际数据的简单且有实际意义的离散化，并且特别考虑了离群点的离散化。

✓ 基于聚类分析的离散化

基于聚类分析的离散化方法也是一种无监督的离散化方法。此种方法包含两个步骤，首先是将某特征的值用聚类算法（如 K-means 算法）通过考虑特征值的分布以及数据点的邻近性，划分成簇或者组。然后是将聚类得到的簇进行再处理，可分为自顶向下的分裂策略和自底向上的合并策略。分裂策略是将每一个初始簇进一步分裂为若干子簇，合并策略则是通过反复地对邻近簇进行合并。聚类分析的离散化方法也需要用户指定簇的个数，从而决定离散产生的区间数。

现阶段，无监督的方法还比较少，在没有类信息的情况下，要得到好的离散化结果比较困难，并且离散化的结果也比较难衡量。但是实际数据集在多数情况下又是没有类标签的，可以考虑先使用聚类算法人为地为数据集添加类标签，然后再用添加了类标签的数据集进行离散化指导离散化过程。

（2）有监督的方法

✓ 1R 方法

1R 是一种使用分箱的有监督的方法。它把连续的区间分成小的区间，然后再使用类标签对小区间的边界进行调整。每个区间至少包含 6 个实例，最后一个区间除外（最后一个区间包含所有未被列入其他区间的实例）。从第一个实例开始，把前 6 个实例列入第一区间，如果下一个实例与此区间中大多数实例的类标签相同，则把此实例加入区间中，再判定下一个实例按照前述操作能否加入刚才的区间中，否则形成下一个含 6 个实例的新的区间，对下一个实例重复类似的操作，直至结束。然后把区间中的大多数实例的共同类标签作为此区间的类标签，如果相邻区间经过此操作后有相同的类标签，则应把这两个相邻区间合并。

1R 离散化方法也是分箱方法，操作仍然比较方便，但又不再需要用户人为指定箱的个数，也克服了无监督的分箱方法的不使用类信息的缺陷，并且能避免把具有相同特征值相同类标签的实例分入不同的小区间中。

✓ 基于熵的离散化方法

由于建立决策树时用熵（Shannon, C. and Weaver, W. 1949）来分裂连续特征的准则在实际中运行得很好，考虑将这种思想扩展到更通常的离散化中，通过反复地分裂小区间直到满足停止的条件。由此产生了基于熵的离散化方法。熵是最常用的离散化度量之一。基于熵的离散化方法使用类分布信息计算和确定分裂点，是一种有监督的、自顶向下的分裂技术。

ID3（Quinlan, J.R. 1986）和 C4.5 是两种常用的使用熵的度量准则来建立决策树的算法。ID3 通过贪心算法搜寻给定数据区间内的具有熵值最小的数据点作为断点。该方法将区间内的每一个数值作为候选断点，计算其熵值，然后从中选出具有最小熵值的数据点作为断点，将区间一分为二，然后再对得到的区间递归的应用以上方法进行离散化。当得到的每个区间中的类标签都是一样时，即停止

离散化过程。这个准则也可以根据需要适当放宽要求。它的缺点在于其使用的停止准则的原则化合理性是有待考虑的，分类和离散化毕竟是两个不能完全等同的处理过程。而使用 ID3 和 C4.5 的方法来离散化完全是照搬利用这两种算法建立决策树的思想。

在 ID3 法的基础上，又产生了 D2 和 MDLP（minimum description length principle，最小描述长度准则）的基于熵的离散化方法。D2 法与 ID3 法的不同之处在于，ID3 法是动态的，在建树的过程中进行离散化，而 D2 法则不是。此外，D2 法与 ID3 法没有明显的不同，D2 法也是计算数据集中每个取值的熵，选取熵值最小的点作为端点，将区间一分为二，再对得到的每一个区间递归地二分。离散化过程在满足 D2 法的停止准则时停止。这些停止准则包括待分裂的区间中含的实例数少于 14 个、或者得到的区间数大于等于 8 个等等。

MDLP 的思想是假设断点是类的分界，依此得到许多小的区间，每个区间中的实例的类标签都是一样的，然后再应用 MDLP 准则衡量类的分界点中哪些是符合要求可以作为端点，哪些不是端点需要将相邻区间进行合并。由此选出必要的断点，对整个数据集进行离散化处理。这种离散化方法的停止准则相对于 D2 更完善，并且选出的断点是区分类的点。

✓ 基于卡方的离散化方法

前面提到的离散化方法均采用自顶向下的分裂策略，即先把整个数据集当作一个区间，再逐步选出端点对大的区间进行分裂得到小的区间，而基于卡方的离散化方法是采用自底向上的策略，首先将数据取值范围内的所有数据值列为一个单独的区间，再递归地找出最佳邻近可合并的区间，然后合并它们，进而形成较大的区间。在判定最佳邻近可合并的区间时，会用到卡方统计量来检测两个对象间的相关度。

最常用的基于卡方的离散化方法是 ChiMerge 方法（Kerber, R. 1992），它是一种自动化的离散化算法[62]。它的过程如下：

首先将数值特征的每个不同值看作一个区间，对每对相邻区间计算卡方统计量，将其与由给定的置信水平确定的阈值进行比较，高于阈值则将相邻区间进行合并，因为高的卡方统计量表示这两个相邻区间具有相似的类分布，而具有相似类分布的区间应当进行合并形成为一个区间。合并过程递归地进行，直至计算得到的卡方统计量不再大于阈值，也就是说，找不到相邻的区间可以进行合并，则离散化过程终止，得到最终的离散化结果。

应用卡方统计量检验两个对象是否相关时，需要人为设定置信水平参数，由统计学知识算出一个与计算量相比较的阈值。对于置信水平的设置要合理，过高会导致过分离散化，过低又会导致离散化不足。对此 Liu and Setiono 等人作了改

进，并形成了 Chi2 算法，它不再使用固定显著性水平，而是使显著性水平逐步下降，这样可以使得在满足不一致率准则的前提下，更多的区间被合并。Chi2 算法分为两个阶段进行离散化，并通过检验系统的不一致率，确定是否终止。

用不一致率检验离散化程度的 Chi2 算法，优于显著性水平固定设定的 Chimerge 算法，但该算法也有不足之处，主要有：

（1） 需设定不一致率的下限值,为此需经多次试验，工作量大，且难以优选；

（2）不一致率并不能全面反映分类特性，而这些特性体现了对样本数据离散化程度的要求；

（3） 有多个变量(特征) 需要离散化时，Chi2 算法没有考虑离散化顺序对结果的影响。

除上面介绍的几种离散化方法之外，还有很多其他的离散化的方法，包括在前述方法基础上针对其缺陷进行改进的离散化方法，以距离度量等其他思想作为指导进行离散化的方法，以及在一种离散化方法中揉合应用多种单一的离散化方法以有效结合各种单一离散化方法的优点克服某些缺点的方法。由于篇幅限制，本文并未详细介绍所有的离散化方法[63]。

现阶段，基于监督的离散化方法的应用范围不是很广，原因是由于很多数据本身是没有类别标签的，无法用基于监督的离散化方法对数据进行预处理。笔者考虑到这方面的原因，为了使系统的应用范围更加普遍，采用无监督的方法进行数据离散化。

在聚类问题中，涉及很多事物之间界限模糊的情况，这时需要将模糊的概念用于聚类分析而产生的模糊聚类来解决，用模糊聚类对数据进行离散化，更符合现实数据的特点，因而具有更高的准确性，所以本文采用模糊聚类的方法对数据进行离散化处理。

笔者采用对数据集中的每个连续属性分别进行数据离散化的方法，每个属性都设有三个标度：5 标度、7 标度、9 标度可供选择，在实际运用的过程当中，可由专家根据领域知识选择各个属性离散化的标度。

算法如下所示：

（1） 随机初始化隶属度矩阵

$$U=[u_{ij}]: \quad U^0;$$

（2）根据$U^{(k)}$计算聚类中心向量$C^{(k)}=[c_j]$：

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \bullet x_i}{\sum_{i=1}^{N} u_{ij}^m},$$

（3）更新隶属度矩阵$U^{(k)}$，记为$U^{(k+1)}$如下：

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2}_{m-1}}$$

（4）如果$\|U^{(k+1)} - U^{(k)}\| < \varepsilon$则停止迭代，否则回到第二步[64]。

由于每个属性的单位及数量级均不相同，所以在离散化前要将属性中心化跟标准化：

$$x'_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j}$$

其中，x_{ij}为某属性的属性值，$\overline{x_j}$为某属性的平均值，即$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$；$S_j$为该属性的标准差，$S_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(xij - \overline{x_j})^2}$。

算法用 VS2005 平台的 C#语言开发，离散化的最终结果由 DataGridView 显示，某次离散化结果如图 2-1 所示：

	V1	V2	V3	V
▶	-1.1721852...	0.86162542...	-1.3667837...	-:
	-1.1721852...	-0.1556546...	-1.3667837...	-:
	-1.1721852...	0.29991054...	-1.3667837...	-:
	-1.1721852...	0.29991054...	-1.2184747...	-:
	-1.1721852...	0.86162542...	-1.3667837...	-:
	-0.3074201...	1.69091499...	-1.2184747...	-:
	-1.1721852...	0.86162542...	-1.3667837...	-:
	-1.1721852...	0.86162542...	-1.2184747...	-:

图 2-2 数据离散化结果

Fig.2-2 Result of data discretization

2.3 补充缺失值

基于依赖的贝叶斯网络结构学习算法需要有完整的数据集，才能从数据中正确的学习到变量之间的因果关系及网络结构，而现实数据集中常常有存在缺失数据的现象，因此，在离散化数据的基础上，需要补充缺失的数据。

2.3.1 空值语义

对于某个对象的属性值未知的情况，称它在该属性的取值为空值(null value)。空值的来源有许多种，因此现实世界中的空值语义也比较复杂。总的说来，可以把空值分成以下三类：

（1）不存在型空值。即无法填入的值，或称对象在该属性上无法取值，如一个未婚者的配偶姓名等；

（2）存在型空值。即对象在该属性上取值是存在的，但暂时无法知道。旦对象在该属性上的实际值被确知以后，人们就可以用相应的实际值来取代原来的空值，使信息趋于完全。存在型空值是不确定性的一种表征，该类空值的实际值在当前是未知的。但它有确定性的一面，诸如它的实际值确实存在，总是落在一个人们可以确定的区间内。一般情况下，空值是指存在型空值；

（3）占位型空值。即无法确定是不存在型空值还是存在型空值，这要随着时间的推移才能够清楚，是最不确定的一类。这种空值除填充空位外，并不代表任何其他信息。

数据缺失在许多研究领域都是一个复杂的问题。对数据挖掘来说，空值的存在，造成了以下影响：

（1）系统丢失了大量的有用信息；

（2）系统中表现出的不确定性更加显著，其中蕴涵的确定性成分更加难以把握；

（3）包含空值的数据会使挖掘过程陷入混乱，导致不可靠的输出。

数据挖掘算法本身更致力于避免数据过分适合所建的模型，这一特性使得它难以通过自身的算法去很好地处理不完整数据。因此，空缺的数据需要通过专门的方法进行推导、填充等，以减少数据挖掘算法与实际应用之间的差距。

2.3.2 造成数据缺失的原因

在各种现实的数据库中，属性值缺失的情况经常发生甚至是不可避免的。因

此，在大多数情况下，数据集是不完备的，或者说存在某种程度的不完备。造成数据缺失的原因是多方面的，主要有以下几种：

（1）有些信息暂时无法获取。例如在医疗数据库中，并非所有病人的所有临床检验结果都能在给定的时间内得到，致使一部分属性值空缺出来。又如在申请表数据中，对某些问题的反映依赖于对其他问题的回答等；

（2）有些信息是被遗漏的。可能是因为输入时认为该信息不重要、忘记填写了或对数据理解错误而遗漏，也可能是由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失了；

（3）有些对象的某个或某些属性是不可用的。也就是说，对于这个对象来说，该属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等；

（4）有些信息（被认为）是不重要的。如一个属性的取值与给定语境是无关的，或训练数据库的设计者并不在乎某个属性的取值（称为 dont-care value）；

（5）由于一些信息获取的代价太大；

（6）系统实时性能要求较高，即要求得到这些信息前迅速做出判断或决策；

2.3.3 数据缺失机制

对缺失数据进行处理前，了解数据缺失的机制和形式是十分必要的。将数据集中不含缺失值的变量（属性）称为完全变量，数据集中含有缺失值的变量称为不完全变量，Little 和 Rubin 定义了以下三种不同的数据缺失机制：

（1）完全随机缺失（Missing Completely at Random，MCAR），数据的缺失与不完全变量以及完全变量都是无关的；

（2）随机缺失（Missing at Random，MAR），数据的缺失仅仅依赖于完全变量，与不完全变量无关；

（3）非随机、不可忽略缺失（Not Missing at Random,NMAR，or nonignorable），不完全变量中数据的缺失依赖于不完全变量本身，这种缺失是不可忽略的。

2.3.4 空值处理的方法

处理不完备数据集的方法主要有以下三大类：

（1）删除元组

也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表。这种方法简单易行，在对象有多个属性缺失值、被删除的含缺失值的对象与信息表中的数据量相比非常小的情况下是非常有效的，在缺少类标号（假设是分类任务）时通常使用。然而，这种方法是有很大的局限性的。它以减

少历史数据来换取信息的完备，这会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息。在信息表中本来包含的对象很少的情况下，删除少量对象就会严重影响到信息表信息的客观性和结果的正确性；当每个属性空值的百分比变化很大时，它的性能会变得非常差。因此，当遗漏数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。

（2） 数据补齐

这类方法是用一定的值去填充空值，从而使信息表完备化。通常基于统计学原理，根据决策表中其余对象取值的分布情况来对一个空值进行填充，譬如用其余属性的平均值来进行补充等。数据挖掘中常用的有以下几种补齐方法：

✓ 人工填写（filling manually）

由于最了解数据的还是用户自己，因此这个方法产生数据偏离最小，可能是填充效果最好的一种。然而一般来说，该方法很费时，当数据规模很大、空值很多的时候，该方法是不可行的。

✓ 特殊值填充（Treating Missing Attribute values as Special values）

将空值作为一种特殊的属性值来处理，它不同于其他的任何属性值。如所有的空值都用"unknown"填充。这样将形成另一个有趣的概念，可能导致严重的数据偏离，一般不推荐使用。

✓ 平均值填充（Mean/Mode Completer）

将信息表中的属性分为数值属性和非数值属性来分别进行处理。如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值；如果空值是非数值型的，则根据统计学中的众数原理，用该属性在其他所有对象的取值次数最多的值(即出现频率最高的值)来补齐该缺失的属性值。另外有一种与其相似的方法叫条件平均值填充法（Conditional Mean Completer）。在该方法中，缺失属性值的补齐同样是靠该属性在其他对象中的取值求平均得到，但不同的是用于求平均的值并不是从信息表的所有对象中取得的，而是从与该对象具有相同决策属性值的对象中取得的。这两种数据的补齐方法，其基本的出发点都是一样的，以最大概率可能的取值来补充缺失的属性值，只是在具体方法上有一点不同。与其他方法相比，它是用现存数据的多数信息来推测缺失值的。

✓ 热卡填充（Hot deck imputation，或就近补齐）

对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似性进行判定。该方法概念上很简单，且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准，主观因素较多。

✓ K 最邻近距离法（K-means clustering）

先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的K个样本，将这K个值加权平均来估计该样本的值。

✓ 使用所有可能的值填充（Assigning All Possible values of the Attribute）

这种方法是用空缺属性值的所有可能的属性取值来填充，能够得到较好的补齐效果。但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大，可能的测试方案很多。另有一种方法，填补遗漏属性值的原则是一样的，不同的只是从决策相同的对象中尝试所有的属性值的可能情况，而不是根据信息表中所有对象进行尝试，这样能够在一定程度上减小原方法的代价。

✓ 组合完整化方法（Combinatorial Completer）

这种方法是用空缺属性值的所有可能的属性取值来试，并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法，能够得到较好的约简结果；但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大。另一种称为条件组合完整化方法（Conditional Combinatorial Complete），填补遗漏属性值的原则是一样的，不同的只是从决策相同的对象中尝试所有的属性值的可能情况，而不是根据信息表中所有对象进行尝试。条件组合完整化方法能够在一定程度上减小组合完整化方法的代价。在信息表包含不完整数据较多的情况下，可能的测试方案将巨增。

✓ 回归（Regression）

基于完整的数据集，建立回归方程（模型）。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。当变量不是线性相关或预测变量高度相关时会导致有偏差的估计。

✓ 期望值最大化方法（Expectation maximization，EM）

EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E步（Excepctaion step，期望步），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M步（Maximzation step，极大化步），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在E步和M步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

✓ 多重填补（Multiple Imputation，MI）

多重填补方法分为三个步骤：

①为每个空值产生一套可能的填补值，这些值反映了无响应模型的不确定性，每个值都被用来填补数据集中的缺失值，产生若干个完整数据集合；

②每个填补数据集合都用针对完整数据集的统计方法进行统计分析；

③对来自各个填补数据集的结果进行综合，产生最终的统计推断，这一推断考虑到了由于数据填补而产生的不确定性。该方法将空缺值视为随机样本，这样计算出来的统计推断可能受到空缺值的不确定性的影响。该方法的计算也很复杂。

✓ C4.5 方法

通过寻找属性间的关系来对遗失值填充。它寻找具有最大相关性的两个属性，其中没有遗失值的一个称为代理属性，另一个称为原始属性，用代理属性决定原始属性中的遗失值。这种基于规则归纳的方法只能处理基数较小的名词型属性。

就几种基于统计的方法而言，删除元组法和平均值法差于 hot deck、EM 和 MI；回归是比较好的一种方法，但仍比不上 hot deck 和 EM；EM 缺少 MI 包含的不确定成分。值得注意的是，这些方法直接处理的是模型参数的估计而不是空缺值预测本身。它们合适于处理无监督学习的问题，而对有监督学习来说，情况就不尽相同了。譬如，对于无监督学习可以删除包含空值的对象用完整的数据集来进行训练，但预测时却不能忽略包含空值的对象。另外，C4.5 和使用所有可能的值填充的方法也有较好的补齐效果，人工填写和特殊值填充则一般不推荐使用。

补齐处理只是将未知值补以主观估计值，不一定完全符合客观事实，在对不完备信息进行补齐处理的同时，会或多或少改变原始的信息系统。而且，对空值不正确的填充往往将新的噪声引入数据中，使挖掘任务产生错误的结果。因此，在多数情况下，仍旧希望在保持原始信息不发生变化的前提下对信息系统进行处理，这就是第三种方法：

（3）不处理

直接在包含空值的数据上进行数据挖掘。这类方法包括贝叶斯网络和人工神经网络等。

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。在这个网络中，用节点表示变量，有向边表示变量间的依赖关系。贝叶斯网络仅适合于对领域知识具有一定了解的情况，至少对变量间的依赖关系比较清楚。否则直接从数据中学习贝叶斯网的结构不但复杂性较高（随着变量的增加，指数级增加），网络维护代价昂贵，而且它的估计参数较多，为系统带来了高方差，影响了它的预测精度。当在任何一个对象中的缺失值数量很大时，存在指数爆炸的危险。

人工神经网络可以有效的处理包含空值的数据，但人工神经网络在这方面的研究还有待进一步深入展开，笔者在这里就不多介绍了[65]。

大多数数据挖掘系统都是在数据挖掘之前的数据预处理阶段采用第一、第二类方法来对空缺数据进行处理的。并不存在一种处理空值的方法可以适合于任何问题。无论用哪种方式填充，都无法避免主观因素对原系统的影响，并且在空值过多的情形下将系统完备化是不可行的。现阶段人工神经网络方法在数据挖掘中的应用仍很有限。值得一提的是，采用不精确信息处理数据的不完备性已得到了广泛的研究。不完备数据的表达方法所依据的理论主要有可信度理论、概率论、模糊集合论、可能性理论，D-S 的证据理论等。

2.3.5 Gibbs 抽样补充缺失数据

从理论上来说，贝叶斯考虑了一切，但是只有当数据集较小或满足某些条件（如多元正态分布）时完全贝叶斯分析才是可行的。而前文中也已经提到过，直接从含有缺失数据的数据集中学习贝叶斯网络的结构复杂性较高，给系统带来较高的方差，会影响最终网络结构的正确性，而且如果数据集中的缺失数据很多是，学习算法存在着指数爆炸的危险，因此在进行贝叶斯网络结构学习前，笔者运用 Gibbs 抽样算法对数据集中存在的缺失数据进行补充[66]。

Gibbs 抽样(Gbbis sample)是已知联合概率分布 $P(X1, X2, ..., Xn)$，求某函数 $f(\{X\})$ 的期望的方法：

（1）给出一组初始抽样，例如可随机产生等：

（2）利用联合概率分布和当前抽样，计算每一变量的条件概率分布：

$$P(x_i \mid \{xj \mid j \in [1,n], j \neq i\}) = \frac{P(x1, x2, ...xn)}{\sum_{X} P(x1, x2, ..., Xi, ...xn)}$$

（3）由（2）的条件概率分布重新生成一组抽样，计算 f，返回（2），迭代至 $f(\{X\})$ 的平均值收敛。

最终得到的收敛值就是对 $f(\{X\})$ 期望的近似[67]。

文献[66]中提到的使用 Gibbs 抽样修复丢失的数据、基于依赖分析方法进行贝叶斯网络结构学习和调整的方法首先随机初始化丢失的数据，并建立最大似然树[68]作为初始贝叶斯网络结构，然后进行数据集和贝叶斯网络的迭代修正、调整，直到结构趋于稳定或满足给定的终止条件为止。每一次数据集修正后进行贝叶斯网络结构调整，使调整后的贝叶斯网络适合于当前的数据集，并且不会陷入局部最优结构。Gibbs 抽样迭代收敛到平稳分布，因此结构序列将收敛到平稳分布的贝叶斯网络结构。联合概率可按贝叶斯网络结构进行分解，对一个变量的抽样只

需考虑对应的条件概率因子即可，而条件概率因子中条件变量的数量与所有变量的数量没有联系，解决了满条件分布(full conditional distribution)所带来的问题，从而能够显著提高抽样效率。

算法见第四章 4.3 节。

第三章 主元分析

对高维数据的分析，常常是一项很复杂的工作。特征提取利用变量变换分析可以得出结果集中的部分变量包含数据中大部分有用信息，比其他的一些变量要更“重要”。通常把更“重要”的变量称为特征。如果次要的变量可忽略不计，那么就达到了降维的目的。此外，从统计学意义上讲，次要分量有可能来自于噪声。因此，除去这些分量可能本身也是净化数据的过程。

特征提取过程是将数据空间变换为特征空间。虽然与最初的数据空间相比，特征空间维数降低了很多，但它仍然保留了数据内容的大多数本质信息[69]。降维的方法有很多，比如主成分分析(PCA)、因素分析以及特征聚类等[70]。

3.1 主元分析的基本原理

主元分析是在有一定相关性的m个样本值和n个参数所构成的数据阵列的基础上，通过建立较小数目的综合变量，使其更集中的反应原数据阵列中包含的信息的方法。其基本方法是根据数据在各负荷向量上的方差大小来确定主元方向的主次地位，按主次顺序得到彼此之间线性无关的各个主元。

为了透彻理解主元分析法的基本原理，分别下面从代数与几何两方面对主元分析法进行阐述。

3.1.1 主元分析的代数原理

主元在代数上是m个随机数据变量$x_1, x_2, ..., x_m$的一些特殊的线性组合。

对P维随机矩阵$x=(x_1, x_2, ..., x_p)^T$，它的第i个主成分为：

$$y_i = l_i^T x = l_{1i} + l_{2i} + ... + l_{pi} x_p$$

同时各主成分满足以下条件：

$y_1, y_2, ..., y_p$两两互不相关，且其对应方差从大到小排列。

另外，为了使各主元方差不因系数向量乘以某个常数而增大，要求各系数向量为单位向量，即$l_i^T l_i = 1$。

各主元可根据如下定理来确定：

定理 2.1：

设总体 $x=(x_1,x_2,...,x_p)^T$ 的协方差阵为 Σ，Σ 的顺序排列的特征根为：$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0, e_1, e_2, ..., e_p$ 为对应的单位化的正交特征向量，则第 i 个主成分为：

$$y_i = e^T x$$

此时：

$$Var(y_i) = e_i^T \Sigma e_i = \lambda, i = 1,2,...p$$

$$Cov(y_i, y_j) = e_i^T \Sigma e_j = 0, i \neq j$$

可见，主元分析法是对过程数据矩阵的最优重构，各数据样本点将以最为均匀、集中的方式分布于各主元向量的周围。

3.1.2 主元分析的几何意义

设在多维空间中，总体 X 以常密度分布于以 μ 为中心的椭球上。

令 $\mu=0$，将椭球平移至坐标原点处，则椭球面方程为：

$$\frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + .. + \frac{1}{\lambda_m} y_m^2 = c^2$$

其中密度常数 $c^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$。

该椭圆主轴沿 e_1 的方向，其余较次要的坐标轴沿 $e_2,...,e_m$ 的方向。

在几何上，由随机变量线性组合而成的各主元代表了一个新的坐标系，它是以 $x_1, x_2, ..., x_m$ 为坐标轴的原坐标系旋转后得到的。新坐标轴代表数据变异最大的方向，并且提供了对协方差结构的一个更为精炼的刻画[71]。

形象地说，PCA 过程是对原有坐标进行坐标平移和旋转变换，使得新坐标的原点和样本点集合的中心重合，新坐标的第一坐标轴对应于数据变异的最大方向，第二坐标轴对应于数据变异的次最大方向，依次类推。若经舍弃少量信息后，由 m 个新坐标轴所构成的子空间能够十分有效地表示原 $p(p>m)$ 维数据的变异情况，则原来的 p 维变量空间就被降至 m 维[72]。

3.2 PCA 基本算法

设有n个样本，每个样本有m个属性，记为：

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}=(x_1,x_2,...,x_m)$$

一般来说，求主成分的计算过程算法如下：

（1）对样本数据的标准化；

为了实现样本数据的标准化，应求样本数据的均值和方差。标准化的实质是对X的列进行中心化和标准化变换，如下：

$$x_{ij}^{'}=\frac{x_{ij}-\overline{x_j}}{S_j}$$

其中，x_{ij}为某属性的属性值，$\overline{x_j}$为某属性的平均值，即$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$；$S_j$为该属性的标准差，$S_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(xij-\overline{x_j})^2}$。

（2）计算相关矩阵；

对给定的n个样本，计算指标变量的相关系数矩阵

$$R=\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mm} \end{pmatrix}=\frac{1}{n}X'X$$

其中$r_{ij}=\frac{1}{n}\sum_{i=1}^{n}X_{ij}X_{ik}=\frac{1}{n}x_j^{'}x_k \qquad j,k=1,2,...,m$

（3）求特征值和特征向量；

设求得的相关矩阵为R，求解特征方程：

$$|R-\lambda f|=0$$

通过求解特征方程，可得到k个特征值$(i=1\sim m)$，和对应于每一个特征值的特征向量：$Q_i=(a_{i1},a_{i2},...,a_{ip}) \qquad i=1\sim k$

（4）求主成分；

通过上述方法可求得$k(k\le n)$个主成分。称$\lambda_i\Big/\left(\sum_{i=1}^{k}\lambda_i\right)$为第$i$个主成分的贡

献率，记为β_i。即：

$$\beta_i = \lambda_i \Big/ \sum_{i=1}^{k} \lambda_i$$

在k个主成分中，称前q个主成分的贡献率之和为前q个主成分的累积贡献率，记为α：

$$\alpha = \sum_{i=1}^{q} \lambda_i \Big/ \sum_{i=1}^{k} \lambda_i$$

主成分的个数可以通过积累贡献率来确定。通常以累积贡献率$\alpha \geq 0.85$为标准。这样所选定的q个主成分基本保留原来n个变量的信息。在决定主成分的个数时应在$\alpha \geq 0.85$的条件下，尽量减少主成分的个数[73]。

鉴于基于依赖的方法构造贝叶斯网络对稀疏网络更加有效，且当数据的节点集较大时计算效率低的缺点，笔者引入主元分析技术为数据集降维，然后用降维后的数据构建网络，提高学习贝叶斯网络结构算法的效率、简化网络结构。

对于使用PCA降维和不使用PCA降维的数据学习贝叶斯网络结构的效率及分类结果见第四章实验分析部分。

第四章 贝叶斯网络结构学习

贝叶斯网络是不确定条件下表示因果关系的有力工具，是概率表中每个节点的有向无环图。贝叶斯网络中的每个节点代表一个数据集中的变量，两节点之间的边代表着节点之间的依赖关系。

4.1 贝叶斯网络结构学习算法

近年来，贝叶斯网络结构学习成为一个非常活跃的研究领域，出现了许多贝叶斯网络结构学习的算法，这些算法通常有两种：基于打分搜索的方法和依赖分析的方法。这两种方法有着它们各自的缺点和优点。通常来说，对于稀疏网络基于依赖分析的方法比打分搜索的方法要更为有效，当数据的条件概率满足特定假设时也能学习到正确的网络结构。但是许多基于依赖方分析法的算法都要求指数级的CI测试和许多更高要求的CI测试（有很大条件集的CI测试）。有很大条件集的CI测试可能是不可靠的除非数据集的属性集很大。另外，尽管打分搜索的方法不能找到最优的结构，但它仍然比基于依赖的方法使用范围要广。

4.1.1 基于打分搜索的方法

（1）Chow-Liu的树构造算法

Chow和Liu提出的树构造算法对图模型的学习领域有着很长时间的影响。它以一个可能的分布P作为输入，构造一个树结构作为输出，算法的复杂度仅仅为$O(N^2)$（N是节点的个数）。他们证明了依赖树的分布P'是与P最为接近的，并且当P的隐含结构本身就是一个树时，算法能保证找到此树的结构。

这种方法比两种图模型学习方法出现的要早，但算法本身的思想其实就是找到一个得分最高的结构，它以对每对结点进行依赖分析结尾，而后者是依赖分析的方法。原因是这种对节点对的分析能够保证最高的分数并且避免局部搜索。

算法的优点是仅仅需要$O(N^2)$次的节点对的独立性判断并且每次计算仅用二阶性质。但是算法对于多连通图的构造不是很有效，因为条件集被包含在节点对的独立性测试中，这意味着更高阶的性质必须被运用。

（2）Poly树构造算法

Poly树构造算法是对Chow-Liu算法的直接的扩展。Poly树是无环结构，也叫单连通图，因此图中两个节点间至多存在一条路径。Rebane和Pearl证明当概率

分布有完全图且映射含有Poly树结构，则他们的算法通常能够找到此完全图。他们还出了通过识别阻塞点来给边定向的方法，这个思想被许多其他的算法运用来为边定向。

（3）K2算法

K2是贝叶斯网络学习算法中一个典型的打分搜索算法。它的输入为一个供学习的数据集和一组有序的节点，输出贝叶斯网络的结构。K2算法运用了一个贝叶斯评分函数，它的目的是找到基于输入数据集的最可能的贝叶斯主网络结构，即使得$P(B_s \mid D)$最大化。

K2算法是非常著名的一个算法，运用它能学习到ALARM网的正确结构（在Macintosh II电脑上大约17分钟能从10000条记录的数据中学习出丢失一条边、多另一条边的网络），ALARM网是公认的检验贝叶斯网络结构学习算法有效性的网络。

（4）HGC算法

HGC算法是另一个基于打分搜索的贝叶斯网络学习算法。此算法所做工作的意义在于通过学习数据包含的属性跟打分方法的假设，找到两种假设，称之为参数模块化和结果相等性，而这些往往被其他的研究者所忽略。通过运用这些由其他研究者提出的假设，Heckerman等人提出了一个结合用户知识和统计数据的一个直接的方法。

（5）CB算法

CB算法是较早的一个试图解决打分搜索算法中要求节点顺序由专家给出问题的算法。由于基于依赖分析的算法能够为边定向，Singh和Valtorta提出了依赖分析方法和打分搜索算法相结合的混合算法。它首先用改进的PC算法得到节点顺序，然后用改进的K2算法学习贝叶斯网络结构。因此，CB算法能够避免要求节点有序的条件。用ALARM网数据测试CB算法，结果产生了有两条丢失边，两条额外边、两条错误定向边的网络结构。

（6）Suzuki's算法

Suzuki's贝叶斯主要网络学习算法基于最小描述长度准则MDL，此准则可以选出一个在网络结构的复杂度和与数据的拟合度之间作出平衡的网络结构规则。此算法的意义在于：不像其他打分搜索算法，它能够避免局部搜索并且能够保证找出最优的结构。由于搜索空间太大，Suzuki提出了一种分支和维度技术，使用它能够在一条边被加入到网络中后判断是否对此分支更进一步的研究是必要的。用ALARM网数据测试此算法，实验结果表明当有100、200、500和1000条记录时，Suzuki's算法比K2算法更为有效、精确。然而，当数据集中的记录增加到上千条时，它的效率就会变得很差。

（7）Lam-Bacchus算法

Lam-Bacchus算法是另一个基于最小化描述长度（MDL）评分函数的算法，它跟Suzuki's算法运用MDL评分函数的方式是一样的。此算法的意义在于不要求节点有序并且它能够运用纯粹的打分搜索方法给边定向，这点与许多基于依赖分析的方法运用基于判别阻塞点为边定向的方法和一些混合算法是有很大不同的。ALARM网数据的实验结果表明，Lam-Bacchus算法能够得到一个存在三条丢失的边、两条错误定向的边的网络。

（8）Friedman-Goldszmidt算法

Friedman-Goldszmidt算法用两种不同的评分函数学习网络结构，例如，MDL评分函数和贝叶斯评分函数。这样做能够同时学习到贝叶斯网络结构和条件概率分布（网络的参数）。Fiedman和Goldszmidt证明他们的方法优于忽略学习网络参数的结构学习算法。和Lam-Bacchus算法一样，此算法能够运用纯粹的打分搜索方法为边定向。当向网络中添加一条边时，它将会选择能得到最高分数的边的方向。此算法用平均信息量距离衡量其与真实模型之间的差别。

（9）WKD（Wallace,Korb and Dai）算法

WKD算法运用最小信息长度评分函数来学习贝叶斯网络结构。最小信息长度（MML）评分函数与最小距离长度方法近似，也能用于上文所述的算法中。WKD算法与PC算法的效率和准确率相同，所不同的只是PC算法需要节点有序，而WKD则不需要。

4.1.2 基于依赖的方法

（1）Wermuth-Lauritzen算法

算法依据节点顺序对每一个节点进行检验，如果节点V_i在节点V_k之前，则对节点V_i和节点V_k用条件独立测试判断它们是否相互依赖，如果是则向图中添加边$V_i \to V_k$。算法能够确保找到数据集的最小I-图。但是由于它需要进行高阶的CI测试，所以此算法适应于具有较大样本及属性集很小的数据集以确保CI测试可行。因此，Wermuth-Lauritzen算法实际上可行性不高。

（2）Bounday DAG算法

Bounday DAG算法是在给定节点顺序和联合概率函数（或足够大的数据集）来构造贝叶斯网络的一种简单的方法。它用目的搜索的方法避免了如Wermuth-Lauritzen算法中所说的大部分的高阶CI测试。然而，找到一个节点的马尔科夫分界线需要考虑这个节点顺序之前的节点的所有子集，所以Bounday DAG算法需要指数级的CI测试。

（3）SRA算法

SRA算法是对Bounday DAG算法的直接扩展，在SRA算法中，可以用节点部分有序或其他领域知识来代替节点完全有序的要求。当寻找被加入图中的节点时，算法用部分有序的信息和局部搜索方法决定哪一个节点将被加入到图中。此算法要求指数级的CI测试。

（4）SGS（Spirtes,Glymour and Scheines）算法

此算法是不需要节点有顺的贝叶斯网络结构学习算法。它能够自动为出CI测试学习得到的网络结构中的边定向。算法要求指数级的CI测试。

（5）PC算法

1991年，Spirtes和Glymour优化了他们的SGS算法，使得其对于模型为稀疏图（图中含有的边较少的模型）时学习贝叶斯网络结构时更为有效。跟其他算法一样，将PC算法用于构造ALARM网的实验，用10000条记录且节点无序，产生的网络有三条丢失的边及两条多余的边。

4.1.3 其他算法

作为打分搜索算法的而一个分支，模型平均技术也被广泛应用于模型的学习当中。此方法认为一个数据集中所包含的潜在的信息有时是不确定的，也就是说，没有一个模型能完全正确的表示一个数据集。因此，该算法会找到几个可能的网络结构并用这些网络的“平均网络”作为最后的输出。

所有以上介绍的算法都假设数据集是完整的[74]。

4.2 基于信息论的贝叶斯网络结构学习

贝叶斯网络的结构学习方法有两种：基于打分搜索方法的和基于依赖分析的方法。本文所用的是基于依赖分析的方法。因此条件依赖关系在本文的算法中扮演着很重要的角色。通过运用马尔科夫条件，能够得到一个贝叶斯网络的条件依赖关系的集合，这些条件依赖关系又影响了其他条件依赖关系。通过运用D分离的概念,所有正确的条件依赖关系都能从贝叶斯网络的拓扑结构中直接导出。

4.2.1 信息论基本概念

信息熵是对信息不确定性的一种度量。从信息论角度来看，信息就是用来消

除不确定性的东西，信息的载体称为消息，含有信息的消息集合称为信源，信源的信息熵，就是信源提供的整个信息的总体度量。所以如果消息消除的不确定性越大，信源的信息熵就越小，信息间的相互依赖性就越大；反之信息间的相互独立性就越大[75,76]。

设 X,Y,Z 为三个不相交的变量集，则称：

$I(X,Y)=\sum_{i=1}^{r}\sum_{j=1}^{q}p(x_i,y_j)\log(\frac{p(x_i,y_j)}{p(x_i)p(y_j)})$ 为 X,Y 的互信息。

$I(X,Y|Z)=\sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{s}p(x_i,y_j,z_k)\log(\frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)})$ 为给定 Z 的条件下，X 和 Y 的互信息。

其中，r,q,s 为 X,Y,Z 的状态个数，$P(x_i,y_j,z_k)$ 为 X,Y,Z 状态为 (x_i,y_j,z_k) 时的概率。

互信息 $I(X,Y)$ 和 $I(X,Y,Z)$ 具有如下性质：

（1）对称性，即 $I(X,Y)=I(Y,X)$ 和 $I(X,Y|Z)=I(Y,X|Z)$；

（2）非负性，即 $I(X,Y)\geq 0$ 和 $I(X,Y|Z)\geq 0$。而且，当且仅当在给定条件 Z，X 和 Y 条件独立时有 $I(X,Y|Z)=0$。

4.2.2 基于信息论的贝叶斯网络学习

Bayesian网是一个有向无环图<DAG>，它可以表示为一个三元组 $G=(N,E,P)$。其中 N 是一组节点的集合，$N=\{x_1,x_2,...,x_n\}$，每个节点代表一个变量（属性）。E 是一组有向边的集合，$E=|<x_i,x_j>|x_i\neq x_j$并且$x_i,x_j\in N|$，每条边$<x_i,x_j>$表示x_i,x_j具有依赖关系 $x_i\rightarrow x_j$。P 是一组条件概率的集合，$P=|p(x_i|\pi_i)|$，$p(x_i|\pi_i)$表示x_i的父节点集π_i对x_i的影响。

设某个系统有 N 个属性节点$\{x_1,x_2,...,x_n\}$，由联合概率公式可得：

$$p(x_1,x_2,...,x_n)=\prod_{i=1}^{n}p(x_i|x_1,x_2,...,x_{i-1}) \qquad (1)$$

现在假设如对每个节点x_i，已知存在某个子集$\pi_i\subseteq\{x_1,x_2,...,x_{i-1}\}$，在给出$\pi_i$时，$x_i$和$\{x_1,x_2,...,x_{i-1}\}-\pi_i$是条件独立的，即有：

$$p(x_i|x_1,x_2,...,x_{i-1})=p(x_i|\pi_i) \qquad (2)$$

由（2）式，（1）可等价为：

$$p(x_1,x_2,...,x_n)=\prod_{i=1}^{n}p(x_i|\pi_i) \qquad (3)$$

由（3）式可知，Bayesian网实质上就是一个联合概率分布 $p(x_1,x_2,...,x_n)$ 所有条件独立性的图形化表示，其中 π_i 为 x_i 的父亲节点集。

依赖模型 M 定义为一组条件独立的集合，设 X,Y,Z 是全集 U 的三个不相交的子集，$M=\{I(X,Z,Y)\}$。其中 $I(X,Z,Y)$ 表示在给定 Z 的条件下，X 独立于 Y，即：$p(X|Y,Z)=p(X|Z)$ 和 $p(Y|X,Z)=p(Y|Z)$。

在依赖模型 M 中，设 X,Y,Z 是全集 U 的三个不相交的子集，条件独立 $I(X,Z,Y)$ 满足对称性、分解率、减缩率、交换律、弱归并率等性质。

从依赖模型 M 出发来构造Bayesian网络的结构是要达到的目标。而利用DAG中D分离的概念来表示依赖模型中的条件独立就可以达到此目的。

✓ V-结构

设 a,b,c 是有向无环图G中三个不同的节点，如果满足 $a\to b\in G$ 或 $c\to b\in G$，且 a 和 c 之间在G中不存在有向边，则称三元组 (a,b,c) 为图G的一个V-结构，b 称为汇聚节点。

✓ D分离

在有向无环图G中，X,Y,Z 是 U 的三个不相交的子集，如果从 X 中一个节点到 Y 中的一个节点的所有路径之间，存在节点E满足以下条件之一：

（1）E不是一个汇聚节点，且E属于Z中；

（2）E是一个汇聚节点，且E或任何E的子节点都不属于Z，则称 X 和 Y 对于 Z 为D分离，记为 $<X|Z|Y>_G$。

这样就可以用 $<X|Z|Y>_G$ 来表示依赖模型中条件独立信息 $I(X,Z,Y)$，从而得到了一个依赖模型的图形化表示方法。

设 M 为概率依赖模型，$<X|Z|Y>_M$ 表示依赖模型 M 所蕴含的依赖关系 $I(X,Z,Y)$，定义：

（1）当 $<X|Z|Y>_M\geq<X|Z|Y>_G$ 时，则称G是 M 的依赖图，记为D-图；

（2）当 $<X|Z|Y>_M\leq<X|Z|Y>_G$ 时，则称G是 M 是的独立图，记作I-图；

（3）当 $<X|Z|Y>_M<=><X|Z|Y>_G$ 时，则称G是 M 的理想图，记为P-图。

由定义可知，对于任一依赖模型 M，最好是能够找到它所对应的P-图，但实际上这并不是总能实现的，所以通常的做法是从I-图着手，找到一个对应于 M 的最小I-图。

设一个有向无环图G是 M 的一个I-图，若删除G中的任何一条边后，使得G不再是 M 的I-图，则称G为 M 的最小I-图。

显然，最小I-图能够最多地表示依赖模型 M 中的依赖关系，即是最后希望学习到的贝叶斯网络结构。

可以证明，满足对称性、分布性、交换律和弱归并律的依赖模型 M，从完

全图中删除所有条件独立性成立的边，则产生一个唯一的最小I-图[76]。

4.3 三步法、Gibbs 抽样构造贝叶斯网络

由于基于打分搜索的方法容易陷入局部最优，而且本身是一个NP难题，所以笔者选择用基于依赖分析的贝叶斯网络结构学习方法。

本文所用的算法是提出的三步法构造贝叶斯网络，此算法是由Jie Cheng、David Bell、Weiru Liu于2002年提出的，由于该算法表现出的良好的效果和效率，使得后来许多学者的研究都基于该算法之上。

算法步骤如下：

（1）随机补充缺失数据；

（2）用普瑞姆算法生成最大似然树构造初始贝叶斯网络；

（3）在已有网络结构基础上用Gibbs抽样修正缺失数据：

$$\hat{x}_{im}=\begin{cases} x_i^1, 0<\lambda\le w(1) \\ \cdots \\ x_i^h, \sum_{j=1}^{h-1} w(j)<\lambda\le \sum_{j=1}^{h} w(j) \\ \cdots \\ x_i^{r_1}, \lambda>\sum_{j=1}^{r_1-1} w(j) \end{cases}$$

其中，x_{im} 是属性 x_i 在其第 m 条记录的待修正值，$\hat{x}_{im}$ 为其修正值，$x_i^1,\ldots,x_i^{r_1}$ 为 x_i 的 r_1 个可能取值，λ 为随机生成数，

$$w(h)=\frac{\hat{p}^{(k)}(x_i^h\mid\pi_{x_{im}},D_{(i,m)}^{(k)})}{\sum_{j=1}^{s}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})+\sum_{j=1}^{l}\hat{p}^{(k)}(x_i^j\mid\pi_{x_{im}},D_{(i,m)}^{(k)})},h\in\{1,\ldots,r_1\};$$

$$\hat{p}^{(k)}(x_i^u\mid\pi_{x_{im}},D_{(i,m)}^{(k)})=\frac{(1/N)}{(N(\pi_{x_{im}})+N(x_i^u)(1/N))};$$

（4）用修正后的数据集学习贝叶斯网络结构：

①对所有互信息大于阈值且在当前图中无边的节点对 $n1$、$n2$：

a. 找出它们邻接路径上的邻居节点，设 $n1$、$n2$ 的这些邻居节点的节点集分别为 $S1$ 和 $S2$；

b. 令集合 $S1$ 和 $S2$ 中势较小的一个作为条件集合 C；

c. 计算条件互信息 $v=I(n1,n2\mid C)$，如果 $v<\varepsilon$，则返回分离；否则，如果 C 只

包含一个节点，那么转去步骤e，否则，对每一个i令

$Ci = C \setminus \{C$中的第i个节点$\}$，　　$vi = I(n1, n2 | Ci)$；

d. 如果$v_{min} < \varepsilon$，则返回分离，否则返回步骤c；

e. 如果$S2$没有用过，那么用$S2$作为条件集C，返回步骤c；否则，返回失败。

f. 如果这对节点在当前图中能够被分离则检测下一对节点对，否则，向图中添加连接这对节点的边。

②对每一条图中存在边的节点对，如果除了这条边之外它们之间还存在其他路径，那么暂时从图中移掉这条边，然后对这对节点进行a~f的检验；如果这对节点不能被分离，则仍将前面移掉的边加入图中，否则永久移除这条边；

③对每一条图中已存在边的节点对，如果除了这条边之外它们之间还存在其他路径，那么暂时从图中移除这条边，并对节点进行以下$a' \sim g'$的检验，如果两个节点不能被分离，则将这条边重新加入图中；否则，将这条边永久的移除；

a' 找到这对节点node1和node2在它们邻接路径上的相邻节点，将node1放入集合N1中，node2放入集合N2中；

b' 找到N1中节点的在node1和node2邻接路径上的、并且不在N1中的邻居节点，将它们放入集合$N1'$；

c' 找到N2中节点的在node1和node2邻接路径上的、并且不在N2中的邻居节点，将它们放入集合$N2'$；

d' 如果集合$N1 \cup N1'$的势小于$N2 \cup N2'$，即：$|N1 \cup N1'| \lhd |N2 \cup N2'|$，则设$C = N1 \cup N1'$；否则让$C = N2 \cup N2'$；

e' 计算条件互信息$v = I(node1, node2 | C)$，如果$v < \varepsilon$，返回分离，否则如果集合C中只包含一个节点则返回失败；

f' 设$C' = C$，对于每一个$i \in [1, |C|]$，设$C_i = C \setminus \{the\ \ i^{th} node\ \ of\ \ C\}$，$v_i = I(node1, node2 | C_i)$，如果$v_i < \varepsilon$则返回分离，否则如果$v_i < \varepsilon + \delta$（$\delta$为一个很小的值）则使得$C' = C' \setminus \{the\ \ i^{th} node\ \ of\ \ C\}$；

g' 如果$|C'| \lhd |C|$则让$C = C'$，转e'；否则返回失败。

④用碰撞识别V结构的方法来对网络中的边定向，对不能构成V结构的边用打分的方法对其进行定向[74]。

普瑞姆算法如下所示：

设TV是最小生成树的已选顶点集合，T是最小生成树的已选边集合。

普瑞姆算法首先任选图 G 中的一个顶点u，将u加入TV中；然后将一条代价最小的边(u,v)加入T中，使得$T \cup \{(u,v)\}$仍然是一棵树。

重复上述步骤，直到T包含$n-1$条边为止。

注意，(u,v)关联的两个顶点必有一个在TV中，另一个不在TV中。

普瑞姆算法的框架如下：

```
TV = {0}；// 从顶点0开始构造，假设G至少有一个顶点
for ((T = ∅； T包含的边少于n-1条；将(u,v)加入T)
{
   令(u,v)为满足u ∈TV且v∉ TV的代价最小的边；
  if (不存在这样的边) break;
  将v加入TV；
}
if (T包含的边少于n-1条)
    cout << "不存在最小生成树" << endl ；
```

设$\overline{TV}$是不属于TV的顶点集合。

对于任何$v \in \overline{TV}$，定义near[v]为TV中使cost(near[v], v)最小的顶点（若(v, w)∉E，则设cost(v,w)= ∞），则$\text{cost}(\text{near}[v],v) = \min_{u \in TV}\{\text{cost}(u,v)\}$。

下一个加入TV的顶点v应满足：$v \in \overline{TV}$，且

$$\text{cost}(\text{near}[v],v) = \min_{x \in TV} P\{\text{cost}(near[x],x)\}。$$

下一条加入T的边自然是(near[v],v)。

设w是$\overline{TV}$中与v邻接的顶点。顶点v加入TV之后，如果cost(v, w) < cost(near[w],w)，则将near[w]改为v。

4.4 实验分析

由于本文运用 PCA 的方法减少了数据集的维数，所以不能用经典的 ALARM 网来验证本文所构造贝叶斯网络结构的正确性，因此笔者将构造的贝叶斯网络用作分类器，跟 TAN 分类器、朴素贝叶斯分类器（NB 分类器）、和一般的贝叶斯网络分类器（BN 分类器）的分类结果作比较。实验共分为采用完整数据集的实验和采用不完整数据集的实验两部分。

因为本文采用 Gibbs 抽样补充缺失数据，所以 4.4.2 节将本算法用在不完整数据集上，检测算法在缺失数据集上构造贝叶斯网络结构的效果。

4.4.1 完整数据集部分

用 IRIS 实际数据[77]、Zoo Data、Glass Identification Data 作为网络学习的数据集，这三组数据为 UCI 数据集中的三个用于分类的数据集，它们均为完整数据集。

（1） IRIS

IRIS 数据集有 150 组数据，共分三类，是鸢尾花的三种类型（Iris setosa，Iris versicolor and Iris virginica），每类 50 个数据；由 4 个属性组成，分别为萼片的长跟宽和花瓣的长跟宽，有类别标签，是用来分类的数据。

IRIS 数据集的所有属性都是连续的，所以需要对其进行数据离散化处理，在本实验中，数据离散化的标度都是笔者随机选择的。

（2） Zoo Data

Zoo 数据不是真实数据，其中有 17 个属性，101 组数据，有类标签。

它是根据动物的一些特征，例如有无毛发、有几条腿等来判别动物的类别的。它的属性有：动物名字、头发、羽毛、蛋、奶水、是否会飞、是否水生、肉食、牙齿、脊柱、呼吸、有毒、鳍、腿、尾巴、群居、大小。数据集的属性大部分是布尔型变量，全部都是离散的。

（3） Glass Identification Data

此数据集是根据玻璃中含有的各种氧化物含量的不同而界定的 6 种玻璃类型数据，它含有 10 个属性，有类别标签。共 214 组数据。

其中属性有：ID、折射率、钠含量、镁含量、铝含量、硒含量、钾含量、钙含量、钡含量、铁含量。

其中后十个属性都是连续变量，需要进行离散化处理。

实验一，用经过 PCA 降维后的数据构造贝叶斯网络并进行分类的结果与未经过 PCA 降维的数据分类结果的准确率进行比较，其中，IRIS 和 ZooData 在 PCA 降维时积累贡献率阈值设为 0.05，在对 Glass Identification Data 进行降维时此值设为 0.1，结果如表 4-1 所示：

表4-1 PCA对贝叶斯分类器准确率的影响

Table 4-1 Affact of PCA in Bayesian classifier

	IRIS	Zoo Data	Glass Identification Data
经过PCA降维	96	97.86	76.72
不经过PCA降维	94	94.29	72.45

用经过 PCA 降维后的数据和不经过降维的数据集分别进行贝叶斯网络结构的学习，所用时间如表 4-2 所示：

表4-2 两种算法的效率比较

Table 4-2 Comparison of efficiency of two algrithms

	IRIS	Zoo Data	Glass Identification Data
经过PCA降维	12．5ms	15.8ms	16.95ms
不经过PCA降维	35．165ms	47.6ms	48.34ms

笔者所用计算机的配置为：

CPU：Intell PentiumI M processor 1.50GHz

内存：768M

硬盘：40G

系统为 Windows XP

本文所用的贝叶斯网络学习算法进行 CI 测试最坏情况下的时间复杂度为 $O(N^4)$。

由表 4-2 可知，采用 PCA 降维后，算法所用时间约占原构造算法时间的 34.58%，贝叶斯网络结构的学习效率有所提高。

经过 PCA 降维，IRIS 数据集的属性由 4 个减少为三个；Zoo Data 的属性由 18 个减少到 12 个；Glass Identification Data 的属性由 11 个减少为 8 个。属性数量的减少使得网络结构更为简单，并且通过表 4-1 可以看出，经过 PCA 降维后进行分类的结果准确率不低于不经过降维直接由数据集学习得到的贝叶斯网络分类结果的准确率，并且算法的效率有了一定的提高。

IRIS 数据降维前后贝叶斯网络结构如图 4-1 和 4-2 所示：

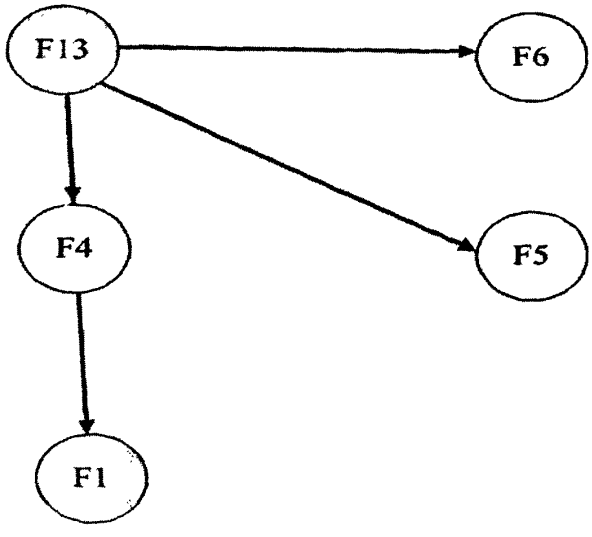

图 4-1 降维前贝叶斯网络结构图

Fig.4-1 Structure of Bayesian network without using PCA

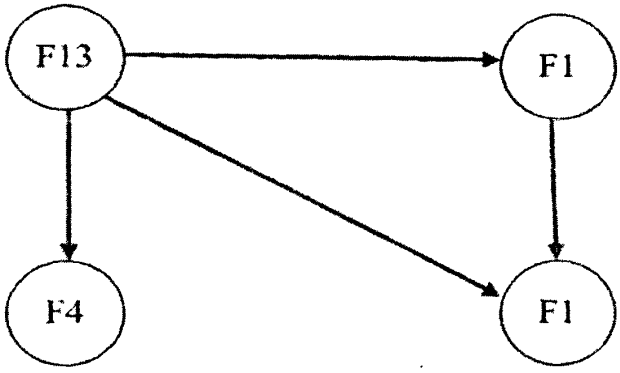

图4-2 降维后的贝叶斯网络结构图

Fig.4-2 Structure of Bayesian network by using PCA

Zoo 数据降维前后贝叶斯网络结构如图 4-3 和 4-4 所示：

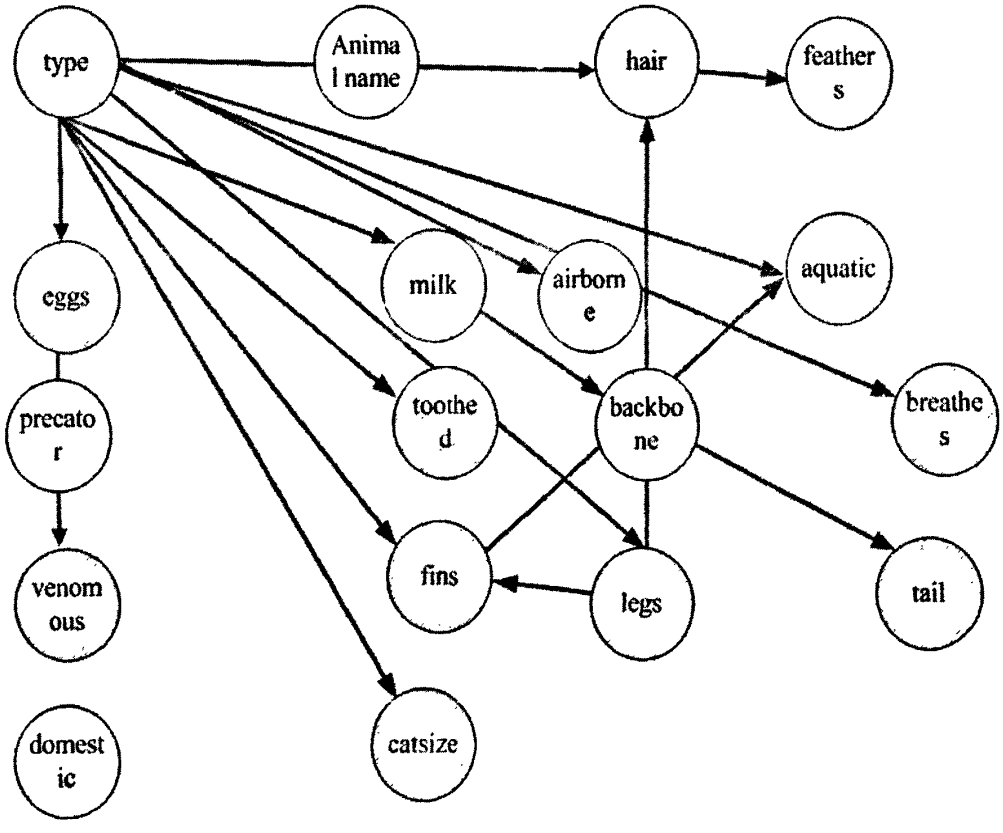

图 4-3 降维前贝叶斯网络结构图

Fig.4-3 Structure of Bayesian network without using PCA

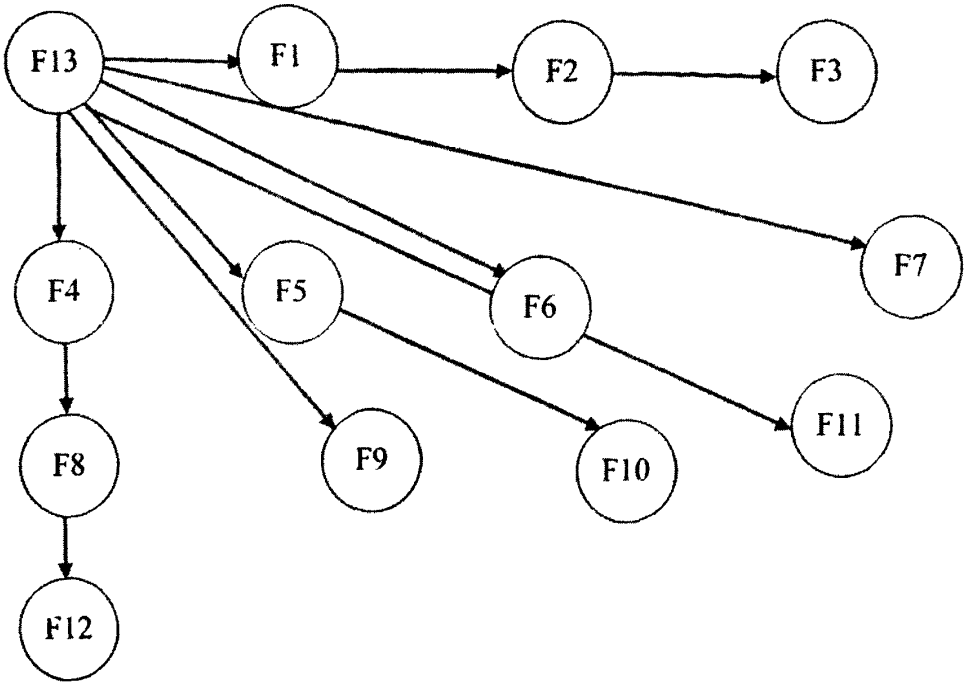

图4-4 降维后的贝叶斯网络结构图

Fig.4-4 Structure of Bayesian network by using PCA

Glass Identification Data 降维前后构造的贝叶斯网络结构如图 4-5 和 4-6 所示：

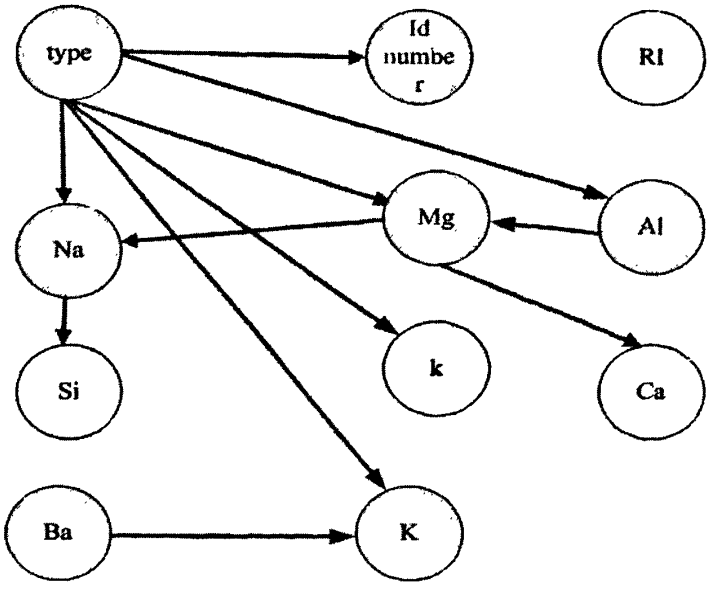

图 4-5 降维前贝叶斯网络结构图

Fig.4-5 Structure of Bayesian network without using PCA

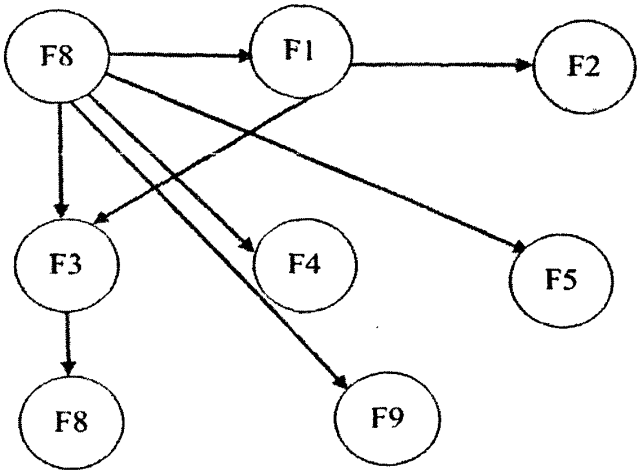

图4-6 降维后的贝叶斯网络结构图

Fig.4-6 Structure of Bayesian network by using PCA

其中图 4-2 中的节点 V4、图 4-4 中的节点 F13 及图 4-6 中的节点 F8 是类别标签节点，其余节点为原数据节点的线性变换，无实际意义。

实验二，用经过PCA降维后的数据构造的贝叶斯网络分类器（BN）与朴素贝叶斯（NB）分类器、TAN分类器分类对以上三组数据进行分类，分类准确率的比较如表4-3所示：

表4-3 三种分类器的分类准确率的比较

Table 4-3 Comparison of accuracy of three classifier

分类器	NB分类器	TAN分类器	BN分类器
IRIS	83.93	87.17	96
Zoo Data	95.05	94.06	97.86
Glass Identification Data	48.60	72.43	76.72

贝叶斯网络分类器本身在各种分类器中就表现良好，由实验一可知，使用 PCA 降维后构造的贝叶斯网络与未使用降维数据学习得到的网络分类结果正确率相差不大，而这样构造的网络分类结果比其他分类器正确率高很多，同时使用降维后数据构造的网络还具有节点少、结构简单、学习效率高等优点。

4.4.2 缺失数据集部分

笔者将 4.4.1 节中所用的三个数据集随机产生缺失数据，数据的缺失率分别为 10%、20%、30%，然后用本文所述方法进行 PCA 降维后构造贝叶斯网络，分类结果如表 4-4 所示：

表 4-4 缺失率对分类器效果的影响

Table 4-4 Affact of missing rate in accuracy of classifier

	IRIS	Zoo Data	Glass Identification Data
缺失率为10%	92.3	94.5	70.56
缺失率为20%	90.5	92..35	69.7
缺失率为30%	88.4	89.91	66.52

由表 4-4 可看出，数据集中的缺失值越多时，其对依赖关系的影响越大，对学习到的贝叶斯网络结构的正确性影响也就越大。

表 4-5 与完整数据集分类效果的比较

Table 4-5 Comparison with classifier with complete data

	IRIS	Zoo Data	Glass Identification Data
有缺失数据	90. 4	92. 253	68. 93
完整数据集	96	97. 86	76. 72

表 4-5 中给出的有缺失数据时贝叶斯网络分类器的分类正确率是缺失率为10%、20%和30%时的分类正确率的平均值，可以看出，本文所用的基于 Gibbs 抽样的贝叶斯网络构造算法在数据有缺失值时表现较好，分类正确率与用完整数据集构造的贝叶斯网络的分类正确率相差不大。

第五章 贝叶斯网络在乙烯能耗指标评价中的应用

5.1 背景介绍

五年内节能20%已作为一项约束性指标写进了“十一五”规划，而石油化工行业能耗占全国工业能耗的1/3左右，其中乙烯工业是石化行业的龙头，所以研究面向乙烯流程的能耗指标体系的建立及应用，可以为建立石化行业科学的节能降耗标杆提供示范与借鉴。为实现这一目标，笔者运用所收集的乙烯全行业能耗相关生产数据、经济数据、设计数据、专家经验、操作规程等海量数据与信息，采用数据/信息融合技术提取能耗特征，建立按照装置、工厂、企业乃至行业的多层次能耗指标体系，评估企业能耗分布，找出节能降耗的方向与主要目标，指导节能降耗，对分解落实量化的节能目标具有非常重要的战略意义。

数据融合是近年来信息技术和自动化技术领域出现的一项高新前沿技术，信息融合采用各种新技术（军事、计算机、控制理论、人工智能、通信技术、信息处理等）对多源信息进行综合处理，经过关联-互联-跟踪-融合计算等多个环节，从中提取有用的信息。数据融合技术模拟人的思维能力，融合信息包括两个部分，一是传感器提供的信息，二是包括人的经验和环境提供的社会信息。该技术目前广泛应用于C^3I系统和控制领域[78]。

本文将数据融合技术应用于建立能耗指标评价系统中的装置能耗指标中，通过数据融合技术建立起实时可以验证的能耗指标，可以改变目前的生产操作考核模式，使生产流程在满足生产负荷、产品质量要求的情况下，使能耗与物耗最合理，产生最佳的社会与经济效益，提高我国石油化工企业的生产管理与操作水平，提高企业的核心竞争力，确保我国支柱产业之一的石油化工行业能够科学和谐地可持续发展。

5.2 贝叶斯数据融合技术

贝叶斯推理方法可以对多传感器测量数据进行融合，以计算出给定假设为真的后验概率[79, 80]。设有n个传感器，它们可能是不同类的，但它们共同对一个目标进行探测。再设目标有m个属性需要进行识别，即有m个假设或命题Ai=1,2,...,m。贝叶斯融合算法在实现上分多级进行。

（1）在传感器一级，将测量数据依其获取的信息特征与要识别的目标属性联系进行分类，最终给出关于目标属性的一个说明B1,B2,...,Bn,它依赖于测量数

据和传感器分类算法[80]；

（2）计算每个传感器的说明(证据)在各假设为真的条件下的似然函数；

（3）依据贝叶斯公式计算多测量证据下各个假设为真的后验概率。最后一步是判定逻辑，以产生属性判定结论。过程如图 5-1 所示：

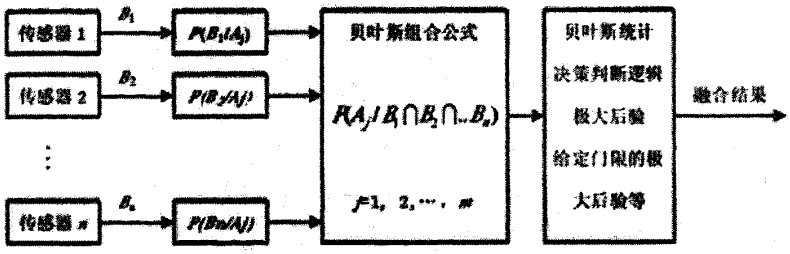

图5-1 基于贝叶斯推理的数据融合

Fig.5-1 Data fusion based on Bayesian

5.3 贝叶斯网络应用结果

本文用贝叶斯数据融合技术对中石化的 21 个不同乙烯生产厂家（不同生产规模及生产技术）的乙烯生产中能耗、物耗的数据进行融合，最后的结果可为乙烯生产中能量、物资的消耗提供相关的参考依据。

21 个乙烯生产厂家及其乙烯生产规模和所用生产技术如表 5-1 所示：

表 5-1 乙烯生产厂家及其规模和技术

Table 5-1 The size and technology of ethylene producers

序号	乙烯厂家	生产能力（万吨）	采用的乙烯技术
1	燕山	71	Lummus 顺序分离
2	大庆	60	KBR 前脱丙烷前加氢
3	齐鲁	80	Lummus 顺序分离
4	扬子	70	Lummus 顺序分离
5	上海 1#乙烯	14.5	三菱重工前脱丙烷后加氢
6	上海 2#乙烯	70	SmW 公司专利技术
7	兰化 1#乙烯	24	SmW 前脱丙烷前加氢
8	兰化 2#乙烯	46	美国 KBR 专利技术
9	辽化	20	中国成达双塔脱甲烷、丙烷
10	盘锦	18	上海惠生
11	抚顺	14	Lummu 顺序分离

12	东方	16	TPL 专利技术
13	独山子	22	Lummu 顺序分离
14	天津	20	Lummu 顺序分离
15	茂名	100	前脱丙烷前加氢
16	中原	18	Lummu 顺序分离
17	吉化有机厂	15	大连工学院技术
18	吉化聚乙烯厂	70	Linde 前脱乙烷
19	广州	21	SmW 前脱丙烷前加氢
20	赛科	90	Lummu 顺序分离
21	扬巴	60	SmW 前脱丙烷前加氢

首先，由于乙烯生产中能耗、物耗相关的属性都为连续属性，所以要对这些数据进行预处理，包括离散化处理及补充缺失数据；然后用预处理后的数据学习构造贝叶斯网络；最后用学习到的网络结构推导出各个属性对融合结果的后验概率作为其在融合过程中的权重，所有厂家乙烯生产能耗及物耗相关属性数据融合结果如图 5-2 至 5-6 所示，其中横轴为时间，纵轴为该属性在时间上的取值。

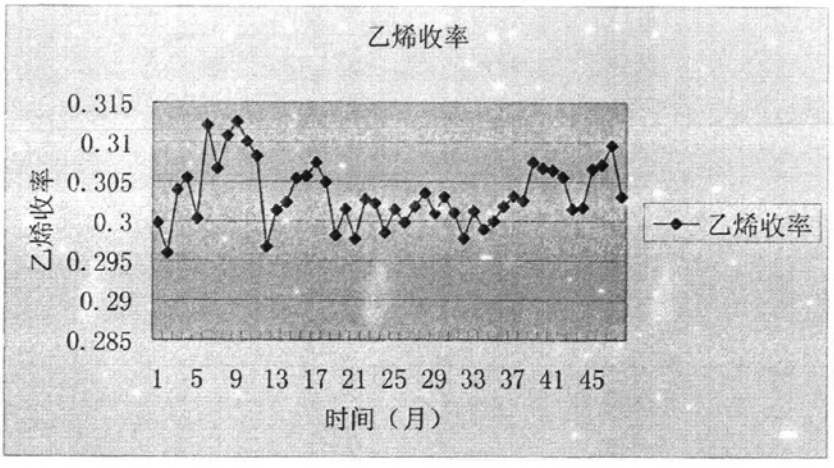

图5-2 不同厂家乙烯收率的数据融合

Fig.5-2 Data fusion of ethylene yield from different factories

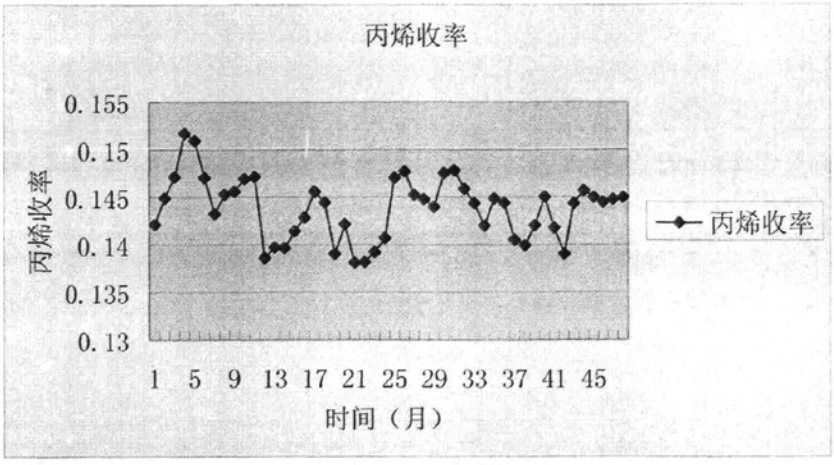

图 5-3 不同厂家丙烯收率的数据融合

Fig.5-3 Data fusion of propylene yield from different factories

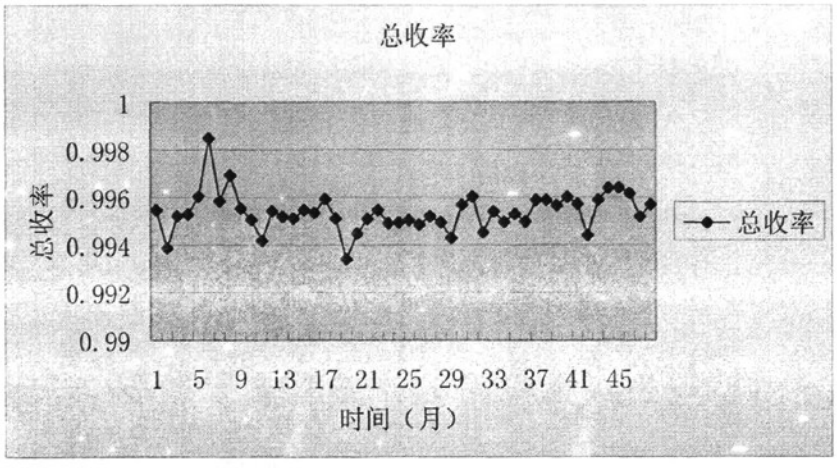

图 5-4 不同厂家总收率的数据融合

Fig.5-4 Data fusion of total yield from different factories

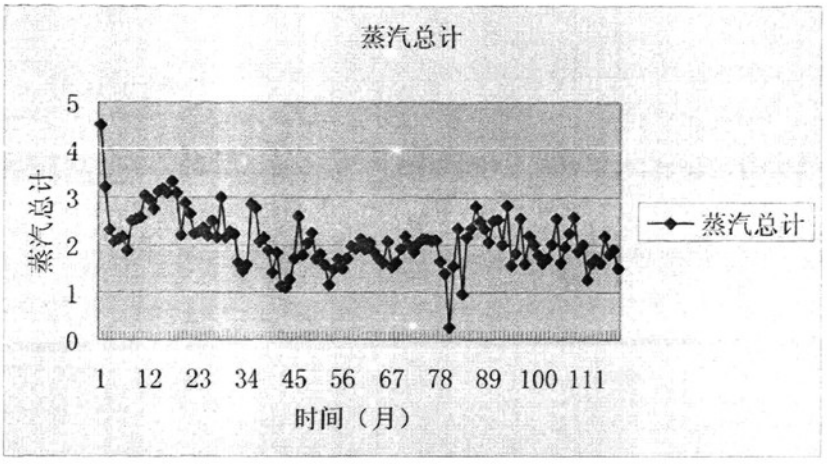

图 5-5 不同厂家蒸汽总计数据融合

Fig.5-5 Data fusion of total steam from different factories

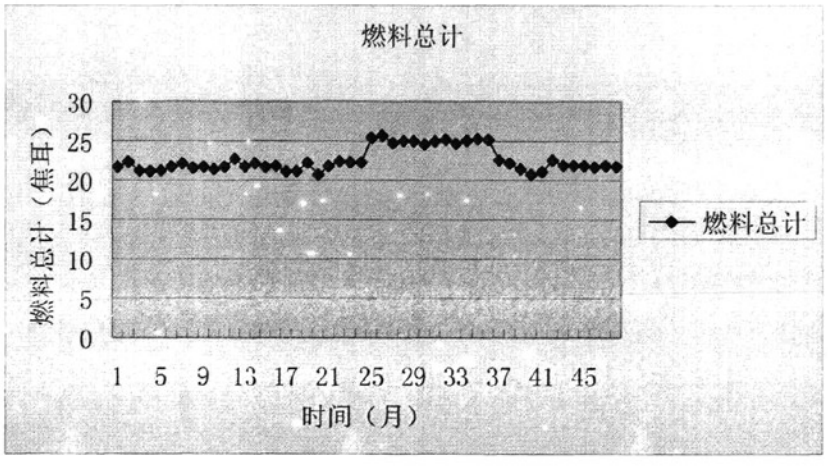

图 5-6 不同厂家燃料总计数据融合

Fig.5-6 Data fusion of total fuel from different factories

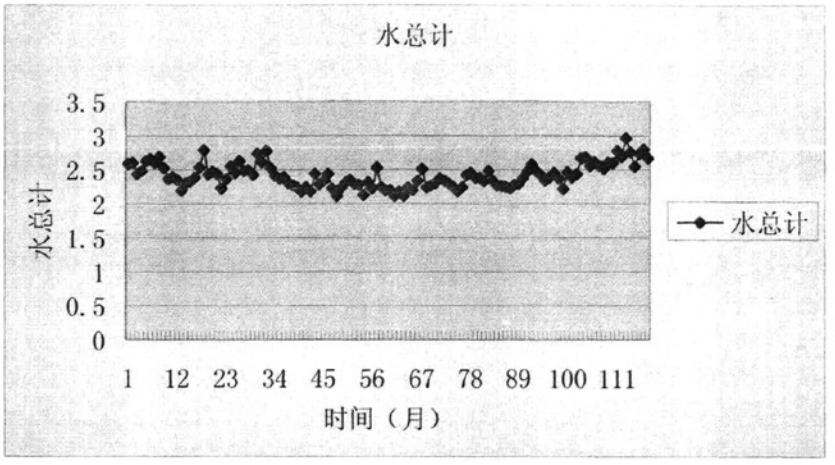

图 5-7 不同厂家水总计数据融合

Fig.5-7 Data fusion of total water from different factories

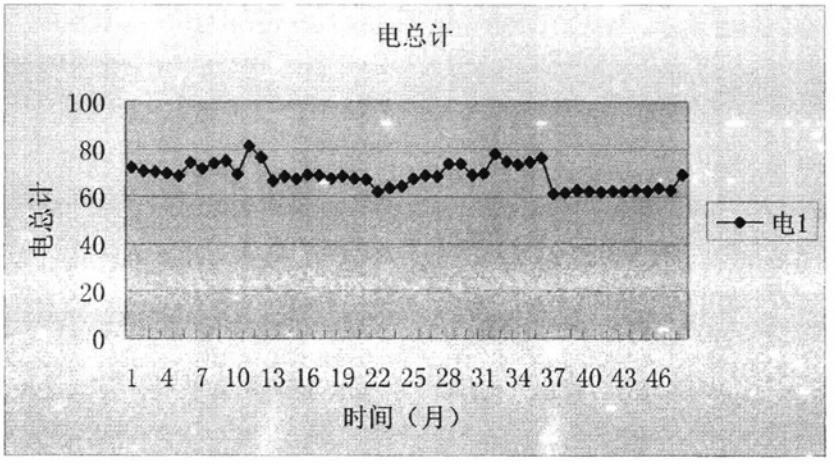

图 5-8 不同厂家电总计数据融合

Fig.5-8 Data fusion of total electricity from different factories

其中图 5-7 和图 5-8 的纵轴单位均为 GJ。

以上得到的结果可作为乙烯生产过程中能耗和物耗相关属性在一定程度上的参考依据。

第六章 总结

由于贝叶斯网络具有图型化的表示形式和直观的推理，而成为不确定知识推理方法中的一个有力的工具，它已经被广泛应用于医学、生物医学、故障诊断等领域。目前国内外许多学者和研究机构都对贝叶斯网络进行了深入的研究，主要集中在以下几个方面：基于贝叶斯网络的学习；基于贝叶斯网络的推理；基于贝叶斯网络的应用。

本文就贝叶斯网络结构学习中基于依赖的学习方法的以下特点：

（1）当结点集较大时，计算效率低；

（2）大多数此类算法都需假设节点有序，但这种假设可能会影响最后学习到的网络结构的正确性；

（3）对于稀疏网络来说，基于依赖分析的学习方法是非常有效的。

对数据集先用 PCA 进行降维，减少数据集属性的个数，以此来提高基于依赖分析方法的效率。

笔者用 VS2005（C#）语言开发，用三组 UCI 数据来验证所学贝叶斯网络分类器的正确性，并将贝叶斯网络用作数据融合，对乙烯生产中不同装置能耗相关的相同属性进行数据融合，所做的主要工作如下：

（1）对数据集进行数据预处理，包括对数据集运用 PCA 方法降维，减少属性集维数；用模糊离散化方法对数据集中存在的连续属性进行离散化处理；用 Gibbs 抽样算法补充数据集中的缺失数据；

（2）用降维后的数据运用基于依赖分析方法学习贝叶斯网络结构；

（3）将学习到的贝叶斯网络结构用作分类器，用实际数据验证学习到网结结构的正确性，并对乙烯生产中不同规模及技术的能耗、物耗相关的数据进行融合，得到的结果可作为相关能耗标准的参考依据。

参 考 文 献

[1] 黄解军. 贝叶斯网络结构学习及其在数据挖掘中的应用研究[D].武汉: 武汉大学, 2005

[2] 黄友平. 贝叶斯网络研究[D]. 北京: 中国科学院研究生院, 2005. 1-5

[3] 关菁华. 基于依赖分析的贝叶斯网络结构学习和分类器的研究与实现[D]. 吉林: 吉林大学, 1979

[4] Heckerman D. A Bayesian approach for learning causal networks[C]. Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal QU: Morgan Kaufmann, 1995, 285-295

[5] Chickering D, Geiger D, Heckerman D. Learning Bayesain networks: Search methods and experimental results[C]. In Fifth International Worshop on Artificial Intelligence and Statiscs, 1995, 112-128

[6] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006

[7] 聂文广, 刘惟一, 杨运涛, 等. 基于信息论的 Bayesian 网络结构学习算法研究[J]. 计算机应用, 2005, 25(1): 1-3, 10

[8] 曹冬明, 张伯明, 邓佑满, 等. 一种新型故障定位方位方法的研究[J]. 电力系统自动化, 1999, 23(7): 12-14

[9] 李伟生, 王宝树. 实现规划识别的一种贝叶斯网络[J]. 西安电子科技大学学报(自然科学版), 2002, 29(6): 741-744

[10] 邓勇, 施文康, 陈良州. 基于模型诊断的贝叶斯解释及应用[J]. 上海交通大学学报, 2003, 37(1): 5-8

[11] 李明, 邓家梅, 曹家麟. 基于贝叶斯网络的串行译码方法[J]. 通信技术, 2001, 4: 38-40

[12] Lucas P J F. Expert knowledge and its role in learning Bayesian networks in medicine:An appraisal[J]. LECT NOTES ARTIF INT, 2001,2101: 156-166

[13] Onisko A, Lucas P, Druzdzel M J. Comparison of rule-based and Bayesian network approaches in medical diagnostic systems[J]. LECT NOTES ARTIF INT, 2001, 2101: 283-292

[14] Beinlich I A, Suermondt H J, Chavez R M, et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks[C]. In: Proceedings of the 2th Second European Conference on Artificial Intelligence in Medicine, London, England, 1989: 247-256

[15] Heckerman D, Mamdani A, Wellman M. Real-world applications of Bayesian Networks[J]. Communications of the ACM, 1995, 38

[16] Ho K M, Scott P D. Zeta: A global method for discretization of continuous variables. In: Proceedings of ĶDD97, Newport beach CA, USA, 1997, 191-194

[17] Neil M, Fenton N, Forey S, et al. Using Bayesian belief networks to predict the reliability of military vehicles[J]. COMPUT CONTROL ENG, 2001, 12(1): 11-20

[18] Alberola C, Tardon L, Ruiz-Alzola J. Graphical models for problem Solving[J], COMPUT SCI ENG, 2000, 2(4): 46-57

[19] Rodrigues M A, Liu Y, Bottaci L, et al. Learning and diagnosis in manufacturing processes through an executable Bayesian network[J]. LECT NOTES ARTIF INT, 2000, 1821:

390-395

[20] Sillanpaa M J, Corander J. Model choice in gene mapping: what and Why[J]. TRENDS GENET, 2002, 18(6): 301-307

[21] Raval A, Ghahramani Z, Wild DL. A Bayesian network model for protein fold and remote homologue recognition. BIOINFORMATICS, 2002, 18(6): 788-801

[22] Geman S, Kochanek K. Dynamic programming and the graphical representation of error-correcting codes. IEEE T INFORM THEORY, 2001, 47(2): 549-568

[23] Raval A, Ghahramani Z, Wild DL. A Bayesian network model for protein fold and remote homologue recognition. BIOINFORMATICS, 2002, 18(6): 788-80

[24] McCabe B. Bilief networks for engineering applications[J]. INT J TECHNOL MANAGE, 2001, 21(3-4): 257-270

[25] Gemela J. Financial analysis using Bayesian networks[J].APPL STOCH MODEL BUS, 2001, 17(1): 57-67

[26] Giudici P. Bayesian data mining with application to benchmarking and credit Scoring[J]. APPL STOCH MODEL BUS, 2001, 17(1): 69-81

[27] Millan E, Perez-de-la-Cruz J L, Suarez E. Adaptive Bayesian networks for multilevel student modelling[J]. LECT NOTES COMPUT SC, 2000, 1839: 534-543

[28] Socher G, Sagerer G, Perona P. Bayesian reasoning on qualitative descriptions from images and speech[J]. IMAGE VISION COMPUT, 2000, 18(2): 155-172

[29] Muhlenbein H, Mahnig T. Evolutionary optimization using graphical models. NEW GENERAT COMPUT, 2000, 18(2): 157-166

[30] Pham T V, Worring M, Smeulders A W M. Face detection by aggregated Bayesian network classifiers. PATTERN RECOGN LETT, 2002, 23(4): 451-461

[31] Wooff D A, Goldstein M, Coolen F P A. Bayesian graphical models for sofeware testing. IEEE T SOFTWARE ENG, 2002, 28(5): 510-525

[32] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Mateo, California, Morgan Kaufmann, 1988

[33] 李小琳. 面向智能数据处理的贝叶斯网络研究与应用[D].吉林: 吉林大学, 2005, 6-8

[34] Jackson J E. Quality control methods for two related variables[J]. Industrial Quality Control, 1956, 7: 2-6

[35] Jackson J E. Quality control methods for several related Variables[J]. Technometrics, 1956, 1: 359-377

[36] MacGregor J F. Statistical process control of multivariate processes[C]. In Proc.of the IFAC Int. Symp. On Advanced Control of Chemical Processes, New York: Pergamon Press, 1994, 427-435

[37] Russell E L, Chiang L H, Braatz R D. Data-driven Techniques for Fault Detection and Diagnosis in Chemical Processes, Springer-Verlag, London, 2000

[38] Kresta J V, Marlin T E, MacGtrgor J F. Multivariable statistical monitoring of process operating performance[J]. Can.J of Chem. Eng, 1991, 69: 35-47

[39] Piovoso M J, Kosanovich K A, Pearson R K. Monitoring process perfomance in real time[C]. In Proc.Of the American Control Conf., Piscataway, New Jersey, IEEE Press, 1992: 2359-2363

[40] Kosanovich K A, Piovoso M J, Dahl K S, et al. Multi-way PCA applied to an industrial batch process[C]. In Proc.of the American Control Conf.,Piscataway, New Jersey, IEEE

Press, 1994, 1294-1298

[41] Wise B M, Gallagher N B. The process chemomtrics approach to process monitoring and fault detection[J]. J.of Process Control. 1996, 6: 329-348

[42] Dunia R, Qin S J. Joint diagnosis of process and sensor faults using principal Component analysis[J]. Control Engineering Practice, 1998, 6: 457-469

[43] Kaspar M H, Ray W H. Chemometric methods for process monitoring and high-performance controller design. AIChE J., 1992, 38: 1593-1608

[44] Luo R, Misra M, Himmelblau D M. Sensor fault detection via multiscale Analysis and dynamic PCA[J]. Ind, Eng, Chem.Res., 1999, 38: 1489-1495

[45] Kramer M A. Nonlinear principal compenent analysis using autoassociative Neural networks. AIChE J., 1991, 37: 233-243

[46] Dong D, McAvoy T J. Nonlinear principal component analysis-based on Principal curves and neural networkds, Comput. Chen. Eng. 1996(20): 65

[47] Dunia R, Qin S J, Edgar T F, et al. Identification of faulty sensors using principal component analysis. AIChE J., 1996, 42: 2797-2812

[48] Ku W, Storer R H, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis[J], Chem.Intell.Lab.Syst., 1995, 30: 179

[49] Kouti T, MacGregor J F. Multivariate SPC methods for process and product Monitoring[J]. Quality Technology, 1996, 28: 409-428

[50] Bakshi B R. Multiscale PCA with application to multivariate statistical processMonitoring[J]. AIChE J, 1998, 44: 1596-1610

[51] Daubechies I. Ten lectures on wavelets[J]. SLAM, Philadelphia, 1992

[52] Konsanovich. Improved process understanding using multiway principal Component analysis[J]. Ind.Eng. Chen.Res, 1996, 35: 138-146

[53] Boque R, Smilde A K. Monitoring and diagnosis batch processes with mulriway covariates regression models[J]. AIChE, J., 1999, 45: 1504-1520

[54] Chen J H, Liu K. On-line batch process monitoring using dynamic PCA and Dynamic PLS models[J]. Chem.Eng.Sci., 2002, 57: 63-75

[55] Chen J, Liu J. Process monitoring using principal component analysis in different operating time processes[J]. Preprints of 14th IFAC World Congress, Beijing, N: 91-96

[56] Lane S. Monitoring of multi-product process[J]. Preprints of 14th IFAC World Congress, Beijing, N: 97-102

[57] Tong H, Crowe C M. Detection of gross errors in data reconciliation by Principal component analysis. AIChE J., 1995, 41: 1712-1722

[58] Data Mining Concepts and Techniques,Second Edition. Jia wei Han, Micheline Kamber. 北京：机械工业出版社，2007

[59] 数据挖掘中的数据预处理[EB/OL].(2007-12-17). [2009-4-1]. http://hi.baidu.com/dingzhoufang/blog/item/927a0afa8813818a9f51463b.html.

[60] Dougherty J, Kohavi R, Sahami M.. Supervised and unsupervised discretization of continuous features[C]. In Proc. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, 1995, 194–202

[61] Ian H. Witten, Eibe Frank,. 数据挖掘实用机器学习技术[M]. 北京: 机械工业出版社, 396-305

[62] Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms[M].

IEEE press, 2003. 19-22, 54-58

[63] 汪庆, 张巍, 刘鹏. 连续特征离散化方法综述[J]. 上海：上海财经大学.

[64] 模糊离散化算法 [EB/OL].[2008-09-28].http://home.dei.polimi.it/matteucc/Clustering/tutorial_ht.

[65] 关于数据缺失问题的总结[EB/OL].(2008-07-01).[2009-04-03]. http://hi.baidu.com/hihsw/blog/item/9ff6ab44beb65b44500ffe78.html

[66] 王双成,苑森淼.具有丢失数据的贝叶斯网络结构学习研究[J]. 软件学报, 2004, 15(7): 1042-1048

[67] Gibbs 抽样.[EB/OL].(2006-05-20).[2009-4-5]. http://pandawendao.spaces.live.com/blog/cns!6327f5bc5215a21c!143.entry.

[68] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997, 29(3): 131-161

[69] 赵雅明, 金祥林, 刘智勇. 因子分析法在试卷分析中的应用[J]. 数理统计与管理, 1995

[70] 方吴丰. 基于 PCA 并行算法的学生评教系统的设计与实现[D].东北师范大学, 2008

[71] 毛振华. 基于主元分析的自适应过程监控方法研究[D]. 浙江: 浙江大学, 2008

[72] 方开泰. 实用多元统计分析[M]. 上海: 华东师范大学出版社.1992

[73] Hastie T, Tibshirani R, Friedman J. 统计学习基础[M]. 北京: 电子工业出版社, 2004.

[74] Cheng J, Bell D, Liu W. Learning Bayesian Networks from Data An Efficient Approach Based on Information Theory. Artificial Intelligence, 2002, 137(1-2): 43-90

[75] 姜丹, 钱玉美. 信息论与编码[M]. 北京: 科学出版社, 1992

[76] 聂文广, 刘惟一, 杨运涛, 等. 基于信息论的 Bayesian 网络结构学习算法研究[J]. 计算机应用, 2005, 25(1): 1-3, 10

[77] Bezdek J C, Keller J M, Krishnapuram R, et al. Will the Real IRIS Data Please Stand Up[J]. IEEE Trans on Fuzzy System, 1999, 7(3): 368-369

[78] 孟宪尧, 白广来, 伞宝钢, 等. 贝叶斯数据融合技术在机舱故障智能诊断中的应用[J]. 大连海事大学学报, 2002,28(3): 10-13

[79] 潘巍, 王阳生, 杨宏戟. 多模态信息融合的一般功能模型设计——基于融合功能与信息层次[J]. 北京: 首都师范大学信息工程学院, 2006

[80] Xiang Y, Pant B, Eisen A, et al. Multiply sectioned bayesian networks for neuromuscular diagnosis. Artificial Intelligence in Medicine, 1993, 5(4): 293–314

致谢

首先向我的导师朱群雄教授及耿志强老师表示最崇高的敬意和最诚挚的谢意。本文的研究工作是在两位老师的精心指导和悉心关怀下完成的，在三年的学习中，无不倾注着老师辛勤的汗水和心血。他们严谨求实的学术作风、渊博的知识、创新的思维、无私的奉献精神使我深受启迪，同时也培养了我刻苦钻研的精神和独立工作的能力。

同时，也要感谢同课题组的各位老师和全体同学的大力支持和帮助，组中营造的活跃而又浓厚的学术氛围使我的课题得以顺利的完成。在此，向他们表示深深的由衷的谢意。

我还要感谢我的家人和朋友，他们的理解和鼓励是我最坚强的后盾。

最后感谢所有帮助过我的人，非常感谢。

研究成果及发表的学术论文

发表及已接受的论文

刘晓洁. 基于 PCA 的贝叶斯网络分类器研究. *国外电子元器件*. 2009, 第 10 期发表

作者和导师简介

导师简介：

朱群雄，男，江苏无锡人，工学博士、教授、博士生导师。主要研究方向智能系统与数据挖掘、信息集成与决策支持，Web-VRGIS。联系方式：zhuqx@mail.buct,edu.cn。

耿志强，男，河南人，工学博士，副教授。主要研究方向为过程建模与优化。联系方式：gengzhiqiang@mail.buct.edu.cn。

作者简介：

刘晓洁，女，1983 年生于甘肃省，团员，汉族。联系方式：sharon0717@126.com

教育经历：

2002 年 9 月-2006 年 6 月：北京化工大学，电子信息科学与技术，学士。

2006 年 9 月-2009 年 6 月：北京化工大学，计算机应用技术，硕士。