分类号_____ 密　　级_____

U D C _____ 学校代码_____

学术型　硕士研究生学位论文

基于贝叶斯网络的大数据因果关系挖掘

学院（部、所）： 信息学院

专　　业： 计算机应用技术

姓　　名： 姚衡

导　　师： 曾志勇　王双成

论文起止时间：2015 年 6 月~ 2016 年 3 月

学位论文原创性声明

声明：本人所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。尽我所知，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。

论文作者签名：[illegible]　　日期：2016年 5月31日

学位论文版权使用授权书

本人完全了解云南财经大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文和论文电子版，允许学位论文被查阅或借阅；学校可以公布学位论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存、汇编、发表学位论文；授权学校将学位论文的全文或部分内容编入、提供有关数据库进行检索。

（保密的学位论文在解密后遵循此规定）

论文作者签名：[illegible]　　导师签名：[illegible]

日期：2016 年 5 月31 日　　日期：2016年 5月31日

摘要

2010 年，全球的数据量跨入了 ZB 时代，根据 IDC 预测，至 2020 年全球将拥有超过 35ZB 的数据量，海量数据将直接或者间接的影响我们的日常工作、生活，乃至国家经济以及社会的发展[1]。大数据时代已经到来。随着大数据的快速发展，以概率统计为基础的机器学习在近年来受到工业界和学术界的极大关注，并在互联网、金融、自然语言、生物等领域获得很多重要的应用，其中贝叶斯网络在过去多年也得到了快速发展，并且成为非常重要的一类机器学习方法[2]。

贝叶斯网络是描述随机变量之间因果关系图的模型，是概率理论、因果推理与图形理论的结合，也是传统的基于数据的统计方法和强调知识的人工智能方法的统一[3]，其重要应用之一是随机变量之间的因果知识表示和推理。贝叶斯网络由结构和参数两部分构成，分别用于定性与定量描述变量之间的因果关系，它具有多功能性、有效性和开放性等特征，能够有效的将数据转化成知识，然后利用这些转化后的知识进行推理，来解决现实世界中的不确定性方面的问题，其有效性已在金融风险分析、信息安全、DNA 分析、软件智能化、医疗诊断、系统分析和控制等许多领域得到验证。

目前，对于非时序的常规数据，通常采用贝叶斯网络来挖掘其中的因果关系；而时序的常规的单时间序列一般采用格兰杰方法来挖掘其中特定的因果关系[4]，但是这种方法存在诸多问题。随着大数据时代的来临，大数据技术为我们分析问题和解决问题提供了新的思路和方法。与常规数据集相比，在大数据环境下进行数据挖掘将得到更多更全面的信息。未来从大数据中发现因果关系以及在常规数据中挖掘一般因果关系将是一种趋势。

为了改善传统格兰杰模型在时间序列因果关系挖掘中出现的弊端，改进并完善因果关系挖掘模型，本文提出了在大数据环境下使用二阶贝叶斯网络模型进行因果关系挖掘。该模型采用最小描述长度（Minimum Description Length，MDL）[5]原理来进行打分。通过对期货样本数据分析，并对原始时间序列进行离散化、属性约简、重构等处理后进行二阶贝叶斯网络模型训练，不仅可以挖掘节点与节点之间的因果关系而且可以发现因果关系之间的联系。

本文的主要工作和主要研究成果如下：

1.分析对比现有因果关系挖掘模型和贝叶斯网络结构学习方法，选择基于 MDL 打分原理的贝叶斯网络模型作为本文的研究方法；

2.提出了一种新型的贝叶斯网络模型：二阶贝叶斯网络模型。设计出了新型模型构建的方法，并实现了相关算法。

3.利用二阶贝叶斯网络推理模型对期货时间序列进行仿真实验，实验不仅得

到了单个期货时间序列内部节点之间的因果关系，而且得到了多个时间序列边与边之间的因果关系。
关键词：大数据　贝叶斯网络　最小描述长度　因果关系　数据挖掘

Abstract

The global data entered a new age of ZB in 2010. According to the prediction of IDC, the global data will reach 35 ZB by 2020. Large amounts of data will affect our work , life and even nation’s economy and social development in real time. The Big Data era has come. With the rapid development of large data, probability-based statistical machine learning caused great concern from industry and academia in recent years and obtained many important successful application in the Internet, finance, natural language, biology and other fields, including that Bayesian network also got fast development over the years and became a very important machine learning method.

The Bayesian network is a model describing the causal relationship graph between random variables.It is a combination of probability theory,causal reasoning and graph theory. The Bayesian network is also a unification of the traditional statisticalmethod which based on data and the artificial intelligence method which emphasizedknowledge.One of its important applications is the causal knowledge representation and reasoning of the random variables. The Bayesian network consists of two parts: structure and parameters, which is used to the qualitative and quantitative description of the causal relationship between the variables, respectively. It has the characters of versatility, effectiveness and openness. The Bayesian network can be effectively transformed data into knowledge and use this knowledge to reason when dealing withuncertain problems in the real-world. Its effectiveness has been verified in the financial risk analysis, information security, DNA analysis, softwareintelligent, medical diagnostics,system analysis and control, etc.

Currently, we usually use Bayesian network to dig into the causal relationship of conventional non-sequential data and use Granger method to dig into a specific causal relationship of the conventional sequential single time series . However, there are many problems with this approach . With the advent of the era of big data, big data technology provides new ideas and methods for us to analyze and solve problems . Compared with conventional data set, data mining in large data environment will provide more comprehensive information. Founding causality from large data and general causation mining in general data will be a trend in the future.

In order to improve the shortcomings of the traditional Granger models in causality of time series mining and take a step forward to improvecausality mining model, this paper propose to use the second-order Bayesian network model to do causality miningin big data environments . This model uses minimum description length (MDL) principle to mark. By futures sample data analysisand the second-order Bayesian network model training of the original time series after processing the discretization, attribute reduction, reconstruction , we can not only mine causalitybetween the nodes , but also find the relationship between causality.

The main works and research results of this paper are as follows:

1. The analysis and comparison of existing causalitymining models and Bayesian network structure learning method. To select Bayesian network model based on the principles of MDL scoring as a research method;

2. Propose a novel Bayesian network model: second-order Bayesian network model. Design a new method of model construction, and realized correlation algorithm.

3. To make simulation experiment on futures time series by using the second-order Bayesian network inference model . The experiment has not only got causal links between the internal nodes of Bayesian network in individual futures time series , but also got causal links between the edges of Bayesian network in multiple time series.

Keywords: Big DataBayesian networkMinimum Description Length (MDL) CausalityData Mining

目　录

第一章　引言

第一节　研究背景及意义

2010 年，全球的数据量跨入了 ZB 时代，跟据 IDC 的预测，至 2020 年全球将拥有 35ZB 的数据量，大量数据将实时地影响我们的工作、生活，甚至国家经济和社会发展。大数据时代已经到来。通常意义上，大数据是指在一段时间内用常规的硬件及软件无法对其进行获取、处理和分析的数据集合[6]。大数据不仅改变了人们的生活和工作方式，同时也改变了企业的经营模式并引起了科学研究领域的变革。大数据具有数据量巨大、数据类型多样、流动速度快和价值密度低的特点，大数据技术为我们分析问题和解决问题提供了新的思路和方法，其研究渐渐成为热点[7]。与常规数据集相比，在大数据环境下进行数据挖掘将得到更多更全面的信息。

随着大数据的快速发展，以概率统计为基础的机器学习在近年来受到工业界和学术界的极大关注，并在互联网、金融、自然语言、生物等领域获得很多重要的成功应用，其中贝叶斯网络在过去多年也得到了快速发展，成为非常重要的一类机器学习方法。贝叶斯网络是描述随机变量之间因果关系图的模型，是概率理论、因果推理与图形理论的结合，也是传统的基于数据的统计方法和强调知识的人工智能方法的统一，它已超出了基于逻辑理论基础的人工智能以及基于规则科技的专家系统范畴，其重要应用之一是随机变量之间的因果知识表示和推理[8]。贝叶斯网络由结构和参数两部分构成，分别用于定性与定量描述变量之间的因果关系，它具有多功能性、有效性和开放性(贝叶斯网络结构是一个能够集成其它智能技术与统计方法的平台)等特征，能够有效地转化数据为知识(具有形象直观的知识表示形式，并可与专家知识有机结合)，并利用这些知识进行推理(具有类似人类思维的推理方式)，以解决现实世界中的不确定性方面的问题，其有效性已在金融风险分析、信息安全、DNA 分析、软件智能化、医疗诊断、系统分析和控制等许多领域得到验证，贝叶斯网络在时间序列因果分析方面也有着广阔的应用前景[9]。

目前，对于非时序的常规数据，通常采用贝叶斯网络来挖掘其中的因果关系；而时序的常规的单时间序列一般采用格兰杰方法来挖掘其中特定的因果关系，但是格兰杰方法存在以下一些问题：①进行格兰杰因果关系检验的一个前提条件是时间序列必须具有平稳性，否则可能会出现虚假回归的问题。②格兰杰因果关系检验的结论只是一种预测，是统计意义上的“格兰杰因果性“，而不是真正意义

上的因果关系，不能作为肯定或否定因果关系的根据。③格兰杰因果关系是建立在线性回归基础之上，因此，从某种意义上可以说格兰杰因果关系是线性因果关系[10]。

随着大数据时代的来临，大数据技术为我们分析问题和解决问题提供了新的思路和方法。与常规数据集相比，在大数据环境下进行数据挖掘将得到更多更全面的信息。未来从大数据中发现因果关系以及在常规数据中挖掘一般因果关系将是一种趋势。为了改善传统格兰杰模型在时间序列因果关系挖掘中出现的弊端，近一步改进并完善因果关系挖掘模型，本文提出了在大数据环境下使用二阶贝叶斯网络模型进行因果关系挖掘。该模型采用最小描述长度（Minimum Description Length，MDL）原理来进行打分。通过对期货样本数据分析，并对原始时间序列进行离散化、属性约简、重构等处理后进行二阶贝叶斯网络模型训练，不仅可以挖掘节点与节点之间的因果关系而且可以发现因果关系之间的联系。

第二节　国内外研究综述

一、大数据国内外发展动态

1989年，第11届国际人工智能大会在美国召开，会上第一次提出了KDD（数据库中发现知识）的概念[11]。1995年召开了全球第一届KDD大会，随着数据挖掘的发展，参与大会的人员逐年增多，KDD国际大会每年都要举办一次。1998年，有30多家公司参与了在美国举行的第四届KDD大会并展示了自己的相关产品，例如，SPSS公司开发的数据挖掘软件Clementine；IBM公司的数据挖掘解决方案；另外还有SAS公司以及Oracle公司的相关产品。

大数据为公司带来了巨大的经济利益，谷歌、IBM、微软等巨头纷纷开始研究大数据相关技术。仅2009年，大数据业务为谷歌带来了500亿美元的收入[12]。Facebook在2011年第一次公布了其新的数据处理分析平台，该平台通过优化数据处理环节，将数据分析周期从两天缩短到10秒以内，和之前使用的hadoop技术相比，效率提高了几万倍。

2012年，Hadoop与大数据技术大会在美国召开，大会的主题是：大数据共享和技术开放。大会上提出了一系列问题，如：数据科学和大数据之间的边界，数据计算的基本模式，大数据带来的架构挑战等等。会上成立了“大数据共享协会”，主要目的在于分享大数据技术，推动大数据的发展[13]。

目前，我国国内的大数据相关技术研究主要集中在数据挖掘算法、数据挖掘理论以及实际应用等领域，涉及到各行各业，包括金融、电信、互联网等行业。

研究单位主要集中在各大高校、研究所以及一些大公司，尤其是一些IT巨头公司，例如华为、阿里巴巴、百度和腾讯等等，这些公司为我国的大数据发展贡献了重要力量。

二、贝叶斯网络研究现状

目前常见的挖掘因果关系的贝叶斯网络结构学习方法有两种，一种是基于搜索-评分，另一种是基于独立性测试。

基于搜索-评分的方法的主要思想类似于贪心算法，它的原理是在其节点的解构空间中指定搜索方法和评分标准，构建贝叶斯网络。这种方法虽然能够得到准确的贝叶斯网络结构，但是却无法保证结果是全局最优的。这类算法的代表有爬山法、K2 算法等等。

基于独立性测试是一种根据变量之间的条件独立性关系来构建贝叶斯网络结构的方法，在给定的数据集中计算变量之间的相关性[14]。该方法比较容易理解，它把独立性测试和贝叶斯网络结构的搜索分开。另外该方法有一个明显的缺点是对独立性测试产生的误差很敏感，并且有时候独立性测试次数会随着变量个数的增加而成指数级增长，时间复杂度非常高。该方法的代表算法有 TPDA 算法，PC 算法等等。

另外，有一种基于扰动的方法也被广泛的研究，该方法由 cooper 发明，他在 1999 年发表的文献[15]中第一次提出在因果关系的挖掘中应该结合真实数据和扰动数据。该方法在因果关系挖掘中表现良好，但缺点是扰动节点需要人工选取。文献[16]中讨论了在扰动的前提下对局部因果结构进行学习的方法。文献[17] 和文献[18]分别提出了两种不同的扰动模型，文献[19]提出了用随机目标节点来进行扰动的方法，该方法的效果具有不确定性。文献[19]文中同时给出随机扰动模型，该模型假设被扰动的节点是已知的，这种假设在真实的数据集上进行了实验，效果很明显。

第三节　研究内容及安排

一、论文的主要工作及创新点

目前，对于非时序的常规数据，通常采用贝叶斯网络来挖掘其中的因果关系；而时序的常规的单时间序列一般采用格兰杰方法来挖掘其中特定的因果关系，但是这种方法存在诸多问题。随着大数据时代的来临，大数据技术为我们分析问题和解决问题提供了新的思路和方法。与常规数据集中相比，在大数据环境下进行

数据挖掘将得到更多更全面的信息。未来从大数据中发现因果关系以及在常规数据中挖掘一般因果关系将是一种趋势。

为了改善传统格兰杰模型在因果关系挖掘中出现的弊端，近一步改进并完善因果关系挖掘模型，本文提出了在大数据环境下使用二阶贝叶斯网络模型进行因果关系挖掘。该模型采用最小描述长度（Minimum Description Length，MDL）原理来进行打分。通过对期货样本数据分析，并对原始时间序列进行离散化、属性约简、重构等处理后进行二阶贝叶斯网络模型训练，不仅可以挖掘节点与节点之间的因果关系而且可以发现因果关系之间的联系。

本文的主要工作和主要研究成果如下:

1.分析对比现有因果关系挖掘模型和贝叶斯网络结构学习方法，选择基于MDL 打分原理的贝叶斯网络模型作为本文的研究方法；

2.提出了一种新型贝叶斯网络模型：二阶贝叶斯网络模型。设计出了新型模型构建的方法，并实现了相关算法。

3.利用二阶贝叶斯网络推理模型对期货时间序列进行仿真实验，实验不仅得到了单个期货时间序列内部节点之间的因果关系，而且得到了多个时间序列边与边之间的因果关系。

二、论文的结构安排

本文共分成六个部分，具体内容如下：

第一章 阐述了本文的研究背景及意义，介绍目前大数据和贝叶斯网络国内外相关发展动态，概括了本文的主要研究工作及组织结构。

第二章 介绍大数据的发展及现状，阐述大数据对数据挖掘方法、工具及其他方面带来的影响。

第三章 介绍贝叶斯网络的基本知识，贝叶斯网络在因果关系挖掘方面的应用。

第四章 介绍实验数据来源，数据的预处理以及数据集的构建。

第五章 阐述新型贝叶斯网络模型，设计算法并如何实现，解释相关的实验结论。

第六章 总结本文的主要工作，并对未来需要进一步完成的工作进行了说明和展望。

第二章　大数据概述

第一节　大数据的产生和发展

癌症是一种疾病的类型，这种疾病是由癌细胞不受控制的分裂并且入侵人体组织所造成的[20]。迄今为止人类发现的癌症种类超多 100 种，其中大部分以癌细胞病变所在的器官来命名。癌症隐藏在人体的细胞之中，在正常情况下，人类的身体控制着新细胞的产生，并用新细胞来替换衰老的或已死去的细胞。但是对于癌症患者来说，应该死亡的细胞并没有死去而应该产生的新细胞却没有产生。这些额外产生的细胞形成了肿瘤。肿瘤分为两种，一种是良性肿瘤，它不会变成癌症；另一种是恶性肿瘤，它会变成癌症。恶性肿瘤在人的身体内扩散并入侵人体组织。

2013 年，在美国总共有 160 万人被确诊为癌症，超过 58 万人死于癌症[21]。

2014 年,美国估计有 235000 人被诊断为乳腺癌,40000 人将死于这种疾病[22]。乳腺癌中最常见是导管癌，治疗乳腺癌最常见的方法有手术、化疗、放疗、免疫治疗和疫苗治疗。通常会选取其中一种或者多种方式进行治疗。国家食品和药物管理部门批准了超过 60 种治疗乳腺癌的药物。治疗的过程和药物的选择由医生和病人的会诊结果来决定，其中包含了很多的因素。一种 AFDA(美国食品药品管理局)批准的治疗乳腺癌的药物叫枸橼酸它莫西芬[23]。数以百万计的乳腺癌患者使用三苯氧胺来进行治疗。由于使用三苯氧胺的成功率高达 80%，所以这种药物是治疗乳腺癌的首选药物。三苯氧胺 80%的有效率给患者带来了极大的希望，但是有一个很重要的事实被忽略了，那就是 80%的有效率并不是指有 80%的概率有效，而是在所有患者中有 80%的人 100%有效而在另外 20%的患者中是 100%无效。这个重大的发现改变了很多人的生活。对数据进行分析，专家们就能提前确定三苯氧胺是否对一名患者有效。在大数据时代来临之前,这种分析是不可能完成的。除了数据，以前计算资源也是一个很大的问题，因为并不是所有的专家都能使用超级计算机的。还有一个很重要的原因就是近年来算法和建模技术日趋成熟。这个故事强调了一个激动人心的事情：我们拥有越来越多的数据、计算资源和算法模型用来进行数据挖掘[24]。拥有数据挖掘所得的知识我们可以重新制定癌症的治疗方案并且改变我们生活中的方方面面。有了大数据挖掘的知识就可以避免对所有的患者使用同样的药物，针对不同的患者给出个性化的治疗方案。

如今我们身处大数据时代，Jared Dean 认为大数据时代起源于 2001 年《纽约时报》上的一些辩论[24]。早在 1990 年“大数据”一词就已经被使用。Francis X. Diebolt 的论文《大数据动态因子模型对宏观经济的测量和预测》于 2003 年

被发表，这是世界上第一篇发表的关于大数据的学术论文[25]。大数据时代所描述的是一个时代，迅速扩张的数据量已经远远超出了大多数人的想象。国际数据公司（IDC）估计整个 2012 年全球总共产生了 2.8 zettabytes 的数据，到 2016 年,这个数字将翻倍。面对如此海量的数据,我们的挑战不是在于如何得到数据，而是在于如何在海量数据中找出我们想要的数据并且发现一些不为人知的知识。

下面这个时间表列出了了大数据时代发展过程中发生的一些重要的事件：

1991 年

众所周知的互联网诞生了。HTTP 协议成为共享信息的标准。

1995 年

Sun 公司发布了 Java 平台。Java 诞生于 1991 年，成为了继 C 语言之后全球最流行的编程语言。它在 Web 应用开发的领域中占据了主导地位。这些应用是记录和存储网络流量的来源

全球定位系统（GPS）全面投入运作。全球定位系统最初是由美国国防部高级研究计划局为军队开发的一套系统，最早应用于 1970 年。该项技术在应用中已经无处不在，汽车导航、航空导航、手机定位等等。

1998 年

Carlo Strozzi 开发了一种开放式关系数据库，称为 NoSQL。十年之后兴起了使用 NoSQL 数据库来处理、存储海量非结构化数据的运动。

Larry Page 和 Sergey Brin 创立了 Google.

1999 年

麻省理工学院的 Auto-ID 中心创始人 Kevin Ashton 发明了“物联网”一词。

2001 年

维基百科诞生了。它彻底改变了人们参考信息的方式。

2002 年

IEEE 发布了蓝牙 1.1 版。蓝牙是一种短距离无线数据传输技术标准。蓝牙的产生使得大量可穿戴设备能够和计算机进行通讯。如今几乎所有的便携式设备都拥有一个蓝牙接收器。

2003 年

根据 IDC 和 EMC 的研究,2003 年一年中产生的数据量超过了之前人类历史上所产生数据量的总和。据估计，仅 2011 年就产生了 1.8ZB 的数据（1.8 ZB 相当于 2000 亿部超清电影,每部两小时）

LinkedIn 被创建了，这是一个受欢迎的专业人士社交网站。2013 年,该网站拥有大约 2.6 亿用户。

2004 年

维基百科在 2 月份达到 500000 篇文章,七个月之后超过了 100 万篇文章。

Mark Zuckerberg 和他的剑桥同事创建了 Facebook（社交网络服务）。2013 年，Facebook 的用户超过了 11.5 亿。

2005 年

Apache 的 Hadoop 项目由 Doug Cutting 和 Mike Caferella 创建。项目的名称来自 Cutting 儿子的一只玩具大象。

2007 年

苹果发布了 iPhone 手机，为智能机创建一个强大的消费者市场。

2008 年

连入互联网的设备数量超过了全球人口总数

2011 年

在电视节目《危险边缘》中，IBM 的超级计算机击败了两名人类玩家。这台超级计算机能够在几秒钟之内扫描并分析 4 TB 的数据。

NoSQL 数据库的查询语言 UnQL 开始被使用。

IPv4 的可用地址被分配完。这个事件显示了联网设备的数量以及需求。

2012 年

美国政府宣布大数据研究发展项目,包括六个部门的 84 个项目。

IDC 和 EMC 估计有 2.8 ZB 的数据被创建,但是只有不到 3%的数据被用于大数据分析。他们预测到 2020 年将产生 40 ZB 的数据，这个数字相当于全世界所有海滩上的沙子颗粒数量的 57 倍。

2013 年

随着智能手机、平板电脑以及 Wi-Fi 的普及，大量的数据以更快的速度被产生。越来越多的个人和公共数据使得数据更具价值。

在过去的 20 年中，数据的处理方式已经从根本上改变。我们每一天都创造了大量的数据。在这些海量的数据中隐藏着大量未知的知识需要我们去发现。

第二节　大数据的定义

目前，大数据一词已经被越来越多的人熟知，但是关于大数据的定义却没有一个统一的说法。大数据是一个非常抽象的概念，它不仅仅是“海量数据”那么简单，通常意义上，大数据是指在一段时间内用常规的硬件及软件无法对其进行获取、处理和分析的数据集合[26]。从不同行业不同的角度来看，大数据有着不同的含义。

2010 年，Apache 给出了大数据的相关定义：普通的计算机软件无法在可接

受的时间范围内捕捉、管理、处理的规模庞大的数据集[27]。2011 年，全球著名的咨询机构麦肯锡公司对 Apache 给出了大数据定义进行的补充：大数据是指数据量的大小超出了典型数据库采集、存储、管理和分析能力的数据集。麦肯锡指出大数据有两个重要特征，首先，数据量是不断变化的，随着时间的增长，技术的进步数据量会不断的增长。其次，数据无法通过传统的数据库进行管理。

国际数据公司(IDC)对大数据的定义有着不同的见解，IDC 认为大数据指的是利用全新的技术及框架体系，通过快速获取、分析和发现大量数据中隐藏的经济价值[28]。大数据有四个特征：海量数据、形式众多、数据生成速度快、高价值低密度，这种定义得到了广泛的认同。

学术界和工业界对于大数据究竟如何定义已有多年的争论，但是大数据如何定义并不是关键，关键是如何从海量数据中挖掘隐藏的价值。

第三节　大数据环境下的数据挖掘算法

随着大数据的发展，越来越多的人开始重视大数据。目前，大数据技术已经在很多领域有着广泛的应用，例如语音识别、图像识别、推荐系统等等，这些高科技应用都离不开数据挖掘算法。数据挖掘算法用来从海量数据中挖掘有价值的知识。

过去，传统的数据挖掘算法在常规数据上的应用非常成功，但是传统的算法有一个缺点，它们的计算通常是基于内存的，随着大数据的发展，数据量成指数级增长，基于内存的数据挖掘算法显然已经无法满足要求，如何开发出新的数据挖掘算法来适应大数据处理的需求已经成为一个非常热门的研究方向。

一、特征选择

通常在进行数据挖掘时会发现数据集往往很大，其中有很多明显没用的数据，如果直接将未处理的数据送入计算机进行模型的学习会导致效率非常低下，所以一般在进行数据挖掘前都要对数据集进行处理，去除不相关的维度，填充缺失值，删除重复样本等等，这样可以显著提高算法效率，减少计算时间。

在大数据环境下，如何处理高维、稀疏数据集是一个巨大的难题。互联网流量、手机通信记录产生了大量高维数据可以运用张量分解来进行大数据分析。Kolda 提出了 Tucker 分解法，该方法可以大幅提高内存的使用率，从而克服了传统张量分解无法解决的效率问题。

Tucker 分解法在分解时可以判断内存的使用情况，根据内存自动选择最合适的执行方法[29]。该算法在内存空间允许的情况采用计算速度最快的方法，避免

产生大量中间结果，自动选择执行顺序，在保证精度的前提下节省内存空间，并且该算法还能避免溢出问题。大多数情况下，在线学习模型需要访问训练集的所有特征空间，但如果训练集是高维数据集时，要获得全部数据集变得非常困难，这使得在线学习模型很多情况下无法应用于实际中。Hoi 等人为了克服这种局限，提出了稀疏正则化和截断技术，这两种技术解决了如何从高维特征集中选择部分特征来代替全部特征进行预测。Hoi 的研究同时还证明了在线特征选择算法在解决现实的大数据挖掘问题上比一些经典的批处理特征选择算法具有很良好的扩展性。

传统的自组织映射（SUM）[30]也是一种特征提取算法，但是该算法存在一个缺点：当数据集较大时，运行速度比较慢。为了克服该缺点，Sagheer 等人提出了一种新的算法 FSUM，该算法基于 SUM 进行了改进。FSUM 算法的主要思想是：由于大数据的主要信息都集中在特征空间的某个区域，那么如果能找出这个区域，从这个区域中提取特征，那么将大幅提高算法效率。

还有一种特征选择算法是 FRFS[31]，该算法由 Ana 等人提出。该算法是一种带阈值的基于模糊粗糙集的特征选择算法。该算法中 Ana 等人加入了一个阈值，该阈值是特征选择时选择特征个数的上限，这样能够有效减少特征选择的个数，从而达到减少计算时间的目的，实验证明该算法性能优异。

SAGA 算法吸取了模拟退火算法和遗传算法的优点，专门用于处理特征最优化选择的 NP 难问题，该算法性能比较优异，能有效降低时间复杂度[32]。

Pa 等人提出了一种专门用于分类问题的特征选择算法，该算法基于支持向量机（SVM）[33]。我们都知道影响一个分类算法精度的因素有很多，其中特征的选择以及数据样本的大小有着至关重要的作用，如果特征选择的好，则能大幅提高分类算法的精度。因此在分类之前进行特征选择非常重要。Su 等人提出了另一种用于分类的算法，该算法利用局部学习思想将复杂的非线性问题转化成对应的线性问题，然后基于 SVM 的最大间隔思想进行特征选择[34]。实践证明，在有着大量不相关维度的数据集上，该算法的表现良好。

综上所述，大数据有数据量大、维度多和数据复杂等特征，所以进行特征处理是首要工作，它将直接影响到后期挖掘工作的精度和效率。到目前为止虽然有很多特征选择算法，但是在实际应用中还是有很多问题难以解决，因此，如何使用特征处理算法是进行大数据挖掘的首先需要解决的问题。

二、大数据分类

常规的分类算法有很多，例如逻辑回归(LR)、支持向量机(SVM)、决策树(DT)等等。这些分类算法在常规数据集上表现良好，但是由于这些算法通常都是基于

内存的，当数据量急剧扩大后便无法处理了。因此，如何在大数据环境下进行分类处理是一个普遍存在的问题。

(1)支持向量机

Corinna 等人在 1995 年提出了一种新型的算法[35]：支持向量机(SVM)。该算法一经提出立刻获得了广泛的关注，它在一些特定的样本条件下表现出的分类精确度让人惊叹。SVM 算法属于监督型学习模型，在分类、模式识别等领域有着重要的应用。SVM 算法的基本原理是在样本空间中找出一个超平面，并且在超平面的两侧固定的距离给出间隔，在超平面的两侧分别是不同的类，由于 SVM 给出了间隔的概念，使得模型不易过拟合，分类的精确度大大提高。另外，在支持向量机算法中提出了使用对偶理论来解决算法推导过程中最优化的难题。

支持向量机算法还有一个亮点是引入了核函数。大部分情况下，样本集往往无法低维空间进行线性划分，但是将样本映射到高维空间后往往就能划分，而通过选择合适的核函数进行映射就能达到这一目的。

虽然支持向量机算法在常规数据集上有着优异的表现，然而它有一个明显的缺点：计算机复杂度很高，并且计算基于内存。因此，在大数据环境下，面对海量数据的时候，支持向量机算法往往无能为力。为了解决这个问题，有人提出了基于支持向量机的在线学习方法。该方法通过迭代计算的思想将海量数据进行分割，分别对分割后小数据集进行计算，然后将结果进行融合。另外，还有人提出在进行模型计算之前先进行特征降维。由于支持向量机算法有很大一部分计算开销是在低维和高维之间通过核函数进行的映射上，因此如果能将维度降低的话就能大幅度减少计算复杂度。

(2)决策树

决策树算法是一种分类算法，由于其优异的表现被广泛采用，是目前分类算法中使用最广泛的算法之一。决策树的一个缺点是计算需要在内存中，因此，在大数据时代，面对海量数据的时候决策树算法往往使得计算开销很大。

Fra 等人对此提出了一种改进的方法，可以将原本进行决策树训练时需要将所有样本集都放进内存中变成只需将其中一部分数据放入内存计算即可，将基于内存变成了基于硬盘[36]。实验证明该方法在处理大数据问题上的速度要比传统的决策树算法来的更快。

Yang 等人通过对决策树算法改进使得其可以进行在线学习，这种增量学习的方法使得决策树能够适应大数据时代数据爆炸的情况[37]。但是这种方法有一定的局限性，当数据样本不均衡或者有大量噪音时容易导致局部最优，必须重新进行全局训练。

Ben 等人提出对决策树算法进行改进，使其能够进行分布式计算，并且控制决策树的分支个数，及时对一些分支进行减支，这样不仅能提高效率而且能有效控制决策树过拟合的问题[38]。

(3)神经网络

神经网络算法最早出现在 1940 年，该算法是科学家们在研究人脑的思维方式时得到启发而发明的。该算法模拟人脑的神经元，将数据集中的每一个维度作为神经网络中的一个神经元，然后通过可调节的带有权值的边将这些神经元连接起来。人工神经网络属于监督式的学习算法，其中反向传播神经网络(BP 神经网络)应用最为广泛。BP 神经网络的中间层数、各神经元之间的权值以及学习参数都可以根据实际需求来设定，因此该算法拥有很强的非线性映射能力，从理论上来讲 BP 神经网络可以逼近任何函数，但是该算法也有很明显的缺点：容易过拟合；容易陷入局部最优陷阱；计算复杂度高。

虽然神经网络算法有众多的缺点，但是由于该算法天生具有分布式并行计算的能力，因此在大数据时代，该算法被广泛的研究与应用，比如最近几年非常流行的深度神经网络。深度神经网络目前已在图像识别、语音识别等领域有着惊人的表现。

综上所述，在大数据时代，很多传统的分类算法无法适应新型、海量的数据，因此如何改进传统的算法使其适应大数据环境是当下需要研究的问题。

三、大数据聚类

目前聚类算法主要有基于划分的、基于层次的、基于密度的和基于网格的。

基于划分的代表算法有 k-means，该算法的思想是给定一个初始值 k，计算机从数据集中随机抽取 k 个样本做为核，然后依次计算其他样本和 k 个核之间的距离，按距离的大小将样本进行划分，全部样本划分完毕之后重新抽取 k 个样本做为新的核，重复上述步骤直至划分的区域收敛为止。该算法简单有效，已在图片分割、商品归类、客户相似度分析等问题中成功应用。

基于层次算法的思想是初始时每个点都是中心，然后判断两两之间的距离，如果符合要求则进行合并，直到最后所有的中心之间的距离均不满足合并的要求为止。该算法简单有效，但缺点是只能发现凸的簇，一旦分错则不可逆，并且需要人工设置满足合并的距离条件。

基于密度算法的思想是预先给定半径（域的大小），如果域内数据点的最小个数满足条件则通过密度可达，实现聚类。该算法的优点是它基于密度，因此无需人工设定簇的个数，特别适合对未知的数据进行聚类，而且该算法能发现任意形状的簇并且对噪音不敏感。缺点是需要人工设定域半径和域内数据点的最小个

数。

随着大数据的发展，数据量的爆炸使得人们对算法的时间和空间复杂度要求越来越高，由于聚类算法天生可并行化，因此受到了广泛的青睐。除了算法本身的性能以外还有一个不可忽视的问题就是 I/O 瓶颈问题，由于进行模型训练的数据量巨大，I/O 已经成为制约算法效率的瓶颈。对于这个问题有人提出了使用压缩算法对训练样本进行压缩后再进入模型训练，这样能有效减少数据样本所占空间，改善 I/O 性能。另外，研究人员发现聚类算法只需部分样本进行训练就可得到很好的结果，因此有人提出了首先使用二次抽样的方法来构建数据样本，然后再进行训练，实验证明这种方法十分有效。

伴随着互联网及其他各种信息技术的发展，数据的表现形式越来越多，聚类面对的挑战不仅仅是数据量暴增的问题，更重要的是如何处理高维问题。例如在对图片、声音、视频等数据进行聚类时，由于这类数据的形式繁多，并且这些特殊数据的维度信息也特别多，需要从数百上千甚至上万个方面来展现。因此如何处理高维数据并进行聚类分析已经成为一个新的研究方向。目前对于高维数据的聚类分析问题中，解决数据高维问题最简单的方法就是对高维数据进行降维。针对高维问题进行改进的聚类算法有很多，比如有基于降维的聚类，划分子空间后再进行降维，另外还有专门针对图的聚类方法。无论何种降维方法都需要将高维数据映射到低维空间中，信息量的损失不可避免，因此如何保证聚类性能的前提下进行尽可能大的降维幅度是目前该领域需要研究讨论的主要问题。

综上所述，在大数据时代，传统的聚类算法不仅面临着数据量巨大，维度高，数据形式复杂等众多问题，而且还面临如何进行并行化改进的挑战。

第四节　大数据处理流程

随着数据获取途径的增多，数据的形式越来越多，如图片、文本、音频等等。虽然对待不同的数据格式处理方式也不尽相同，但是大数据的处理流程却基本一致。本文从相关文论中收集了一些资料，整理后总结出如下 4 个处理流程：

1. 数据收集
2. 数据预处理
3. 数据分析与挖掘
4. 数据解释

由于不同的数据源产生的数据结构不尽相同，如结构数据、半结构和非结构数据等等，因此需要将这些数据处理成统一的数据结构，然后才能使用相关的数据挖掘算法进行处理并将最终结果展示出来。

一、数据收集

在大数据时代，数据量变得越来越大，数据的形式也日益复杂，如何利用各种方法获取数据变得越来越重要。在大数据处理过程中第一步便是如何收集数据，目前常用的数据收集途径有数据库、互联网、传感器等等，另外由于智能手机、平板电脑等移动设备的普及，出现了大量的移动设备软件，增加了数据的收集方式。

二、数据预处理

数据预处理主要是对收集到的数据进行合理的处理，常用的处理方法有清洗、降噪、合并、转化等等。

众所周知，大数据有数据量巨大、数据类型繁多的特征，因此首先要对数据进行格式统一化处理，否则无法进行后续的分析和挖掘。通常情况下把众多类型的数据转化成一种方便后续处理的格式。另外，大量数据中参杂着很多没用的信息，甚至是错误的信息，因此接下来就要进行数据清洗，保证数据的质量，否则大量的噪声会影响到后续的分析工作。目前常用的数据清洗方法有聚类和关联分析等等，通过这些算法将一些错误的、无用的信息剔除掉。

数据分析的过程通常会碰到数据需要离散化的问题，有些数据是连续的，无法满足分析算法的要求，因此需要对连续数据进行离散化处理。另外还有数据的合并、转化等处理不再一一赘述。

三、数据挖掘与分析

数据挖掘与分析是所有处理步骤中最重要的一步，因为数据价值的发现就在这个步骤中。数据经过上述的预处理后便可以进行数据分析了，目前传统的数据分析方法有分类、聚类、关联规则等等。但是随着大数据的发展，这些传统的算法面临着重大的挑战。

传统的数据挖掘算法面对的是常规数据集，因此这些算法的首要问题是如何得到精确的结果，但是在大数据时代，当这些算法面对海量数据时，时间复杂度往往令人难以接受，因此，大数据时代数据挖掘算法的重心开始向效率方向发展，即在时间和空间复杂度可以接受的前提下提高算法的准确度。目前解决算法效率的方法有很多，例如优化传统算法，开发分布式算法等等。

四、数据解释

上述的数据挖掘与分析是大数据处理流程中最重要的一步，但是对于用户来说，他们其实并不关心数据挖掘与分析的过程，而是更加在乎数据分析的结果，因此，对于一个完善的大数据挖掘分析流程中，数据的解释尤为重要，甚至可以这么说，如果没有一个好的结果解释，那么前面的处理步骤就可能徒劳无功。

传统的数据分析结果的展示是以文本的形式，但是这种展示的方式只适合专业的数据分析人员查看，对于客户来讲，这些文本内容太过于专业，无法理解，因此传统的数据展示方式显然已经无法满足要求。为了提升数据的展示能力，如今大部分数据公司都采用了数据可视化技术，利用图片、表格、PPT 甚至动画等手段将数据结果展示给用户，更加方便用户的理解。

第五节　大数据的应用

随着大数据时代的来临，大数据相关产业不仅改变了科技公司的生产经营模式，提升了企业的竞争力，更是推动了世界经济的发展。收集大量的原始数据，通过数据挖掘、机器学习等尖端技术分析获取数据中的潜在知识，用来预测未来即将发生的事件，这将帮助决策者做出正确的判断，提高劳动效率获取更大的利益。

在商业领域，大数据的应用十分普遍，沃尔玛等跨国巨头对收集到的客户购物行为数据进行分析，掌握不同客户的购物习惯，进而合理安排超市里商品的摆放以及有针对的商品推荐，取得了巨大的经济利益，‘啤酒和尿布’的故事就是经典的商业案例之一。国内的商业平台如阿里巴巴、京东等互联网公司也早已经开始应用大数据来分析挖掘潜在价值，例如从客户的历史交易信息，包括购买、收藏、点击、加购物车哪些商品的信息来预测该客户未来可能会购买哪些商品，并及时将可能的商品推荐给该客户。

在金融领域，大数据也扮演着至关重要的作用。例如某金融公司通过分析股票的历史数据、股民的情绪、股吧等社交网络的留言等信息，得出某只股票的涨跌预测，从而决定是否卖出或买入股票，股票分析就是利用了大数据中的时间序列分析。另外，金融领域的信用风险评估也用到了大数据分析，通过对用户的行为、征信报告等信息分析，给出用户的信用评级，从而决定是否给该用户放贷等等。

在医疗领域，大数据的应用也在不断的扩大，例如医院的 B 超、病情分析、CT 扫描每天产生着令人难以置信的海量数据，从这些数据中可以分析得出十分

重要的信息。因此，2010 年我国的“十二五”规划中就提出要重点建设卫生信息平台来存储病人的相关数据。

在制造业领域，随着企业中 ERP 等信息化系统的普及，管理方式逐步的精细化，企业在对业务进行相关操作时产生了大量的数据信息，很多企业已经越来越重视这些产生的数据，通过对这些数据的分析往往能够得到很多隐藏的问题以及被人忽略的价值。因此，很多企业已经开始从原先的流程建设为主向流程的质量为主转变，所以建立以产品为核心的相关数据结构，并从收集的数据中分析潜在价值变得越来越迫切。

第六节　小结

本章主要介绍了大数据的相关基本概念，分别讲述了大数据的产生、发展及应用。另外着重介绍了大数据时代，传统的数据分析方法的改变，介绍了分类、聚类算法以及数据特征抽取在大数据时代下的变化。本章还介绍了大数据处理的流程等。

第三章　贝叶斯网络基本概念

贝叶斯网络又被称为信念网络，是一种概率网络，用来描述多个变量之间的统计关系[39]。贝叶斯网络是建立在概率论基础上的图形化网络，它为多个变量之间的复杂关系提供了一个直观的表示。贝叶斯网络最初被提出用来解决不定性问题，由于它能将大规模复杂的多元关系用简单直白的方法表达出来，因此受到人们的青睐。目前贝叶斯网络已在数据挖掘、人工智能以及模式识别等多个领域受到研究人员的关注。下面将对贝叶斯网络做一个简单的介绍。

第一节　贝叶斯网络模型

在介绍贝叶斯网络之前，首先介绍一下图论中的几个基本概念。设有一个有向无环图为 G，如果 G 中节点 *i* 有一条指向节点 *j* 的边，那么我们称节点 *i* 是节点 *j* 的父节点，节点 *j* 是节点 *i* 的子节点，节点 *i* 的父节点集合由 *pa()*来表示，子节点集合由 *ch()*来表示。连接节点 *i* 和 *j* 的边又称为弦。凡是有边和节点 *i* 连接的节点都称为节点 *i* 的邻居节点，用 *nb()*表示。若一个节点没有父节点则称为根节点，若一个节点不存在子节点则称为叶子节点。若节点 *i* 指向节点 *j*, 节点 *j* 又指向节点 *k*，那么我们称节点 *i* 为节点 *k* 的祖先节点，反过来节点 *k* 是节点 *i* 的子孙节点，祖先节点集合由 *an()*来表示。若有向图中存在某个节点的祖先节点是它自己则称该图中存在环，否则称该图为有向无环图。

定义 3.1 设贝叶斯网络 BN=(G,P),其中 G 为贝叶斯网络中的有向无环图。G=（V,E），其中 *V* 表示图中所有节点的集合，*E* 为反映变量之间关系的集合，若存在有向边从节点 *i* 指向节点 *j* 则表示节点 *i* 和 *j* 之间存在直接因果关系，若节点 *i* 存在祖先节点 *j*，则称 *i* 和 *j* 之间存在间接因果关系，*V* 和 *E* 中的关系一一对应。P 为各节点之间概率分布的集合，用来量化的表示各节点之间因果关系，每个节点都有各自的概率表。

图 3.1 所示为某个案例的贝叶斯网络建模实例。该贝叶斯网络拥有 5 个节点，每个节点都表示一个随机事件，从图中可以看到，当事件 H 发生时，可能会导致事件 B 和 L 的发生，而 B 和 L 又会导致事件 F 的发生，与此同时事件 L 还会影响到事件 C 的发生。由于不知道每个事件发生的确切概率,因此这里用‘是’和‘否’(yes 和 no)来表示，各事件对应的条件概率详见图中所示，该图直观的表达了各事件之间存在的因果关系。

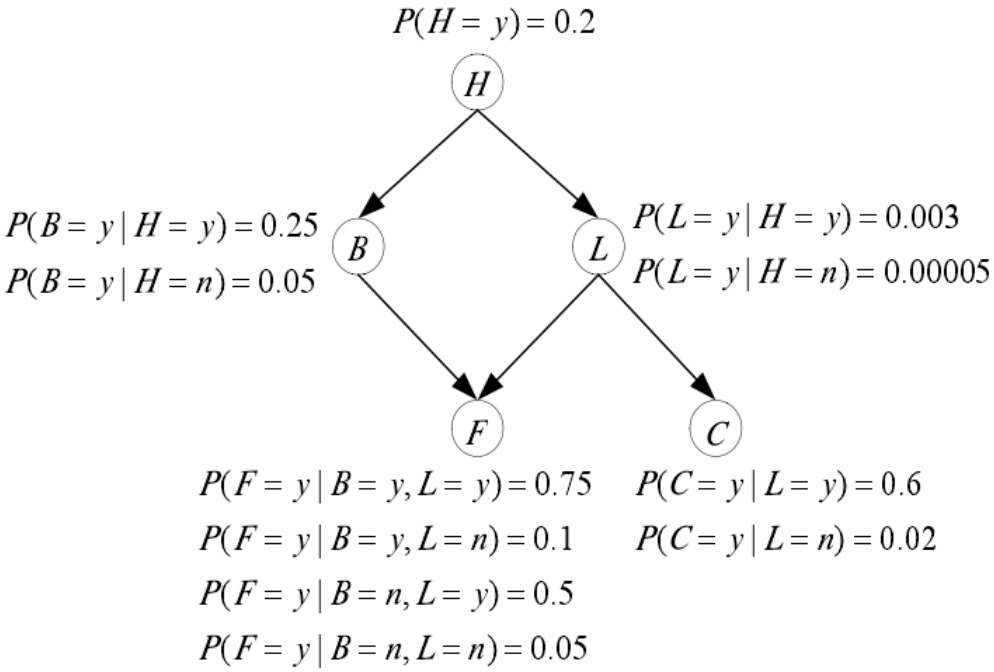

图 3.1 贝叶斯网络实例

从上图中可以看出贝叶斯网络分为两个部分，一个是贝叶斯网络包含了一个有向无环图，该图直观的表达出了各个事件之间的因果关系；另一个是贝叶斯网络用图表示事件之间的关系的同时还用概率分布定量的表示出事件之间概率的具体值。

第二节　参数学习

根据第一节的介绍，我们知道贝叶斯网络由网络结构图和各事件之间的概率分布表示。其中各事件之间的概率分布又称为贝叶斯网络参数，这些参数往往要通过学习才能得到。在进行贝叶斯网络参数学习之前首先要给定一个已经确定的贝叶斯网络结构图以及对应的训练集，然后统计相关变量的先验概率，通过这些先验知识来计算得到贝叶斯网络中各节点的条件概率。我们知道，事件的先验分布往往服从一些常见的概率分布，比如正态分布、几何分布、多项式分布等等，利用某种方法结合已知条件就能估算这些分布的参数。

目前利用贝叶斯网络建模的案例有很多，在实际应用中往往会碰到很多问题，其中有一个很普遍的问题就是如何处理连续变量。由于传统的贝叶斯网络主要处理离散变量，因此碰到连续变量时首先要进行离散化处理。目前贝叶斯网络主要

应用在自然语言预处理、图像分类、社交网络分析等领域中，在这些应用中常见的离散变量分布有贝塔分布和多项式分布。下面将分别介绍这两种分布。假设某个变量的概率分布存在一个离散的样本空间，若这个变量只有两种状态则称他服从贝塔分布；若该变量具有两个以上的状态则称它服从多项狄利克雷分布。

目前，贝叶斯网络的参数学习方法主要有两种：极大似然估计(MLE)和贝叶斯估计(BE)。使用这两种方法进行参数学习的前提是：首先训练集要完整；其次训练集样本必须满足独立同分布的条件。即给定一个数据集 $D=(d_i,d_2,...,d_m)$，其中 $d_i=(d_{i1},d_{i2},...,d_{in})$，$i=1,2,3,...,m$，该数据集有 m 个样本，每个样本含有 n 个变量，则当数据集 D 中样本均满足如下条件时我们称之为独立同分布

(1) 在给定参数 θ 时，数据集 D 中样本相互独立，即 $P(D|\theta)=\prod_{i=1}^{m}P(d_i|\theta)$。

(2) 每个样本的条件概率分布 $P(d_i|\theta)$ 均相同。

一、极大似然估计

极大似然估计(MLE)的思想来源于统计学[40]，该方法通过样本和参数之间的似然度来进行参数学习，即：已知 $L(\theta|D)=P(D|\theta)$ 为 θ 的似然函数，则 $\theta^*=\arg\max L(\theta|D)$。假设有一个贝叶斯网络 A，A 由 n 个变量组成，$X=(x_1,x_2,...,x_n)$，其中某个变量 x_i 存在 k 个不同的取值，该变量所有可能的父节点的个数(可能存在多个父节点)为 q,若无父节点则规定 q=1，那么 $\theta=\{\theta_{ijk}|i=1,\cdots,n;j=1,2,\cdots,q;k=1,\cdots,r\}$，其中 $\theta_{ijk}=P(x_i=k|pa(x_i)=j)$。对 $L(\theta|D)$ 两边都取对数得：

$$l(\theta|D)=\log L(\theta|D)=\log\prod_{i=1}^{m}P(d_i|\theta)=\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}m_{ijk}\log\theta_{ijk} \qquad (3\text{-}1)$$

公式 3-1 中的 m_{ijk} 为数据集 D 中所有满足 $x_i=k$ 和 $pa(x_i)=j$ 的样本的个数，使用拉格朗日乘子法来求解公式 3-1 即可得到参数的 θ 极大似然值。

二、贝叶斯估计

传统的统计学认为概率就是频率的无限逼近，而贝叶斯估计则认为概率是人们对事物认知的程度并且这种认知程度的高低是由人的主观知识和现实中观测到的事物一同决定的。因此贝叶斯估计方法把需要估计的参数值当做其中的一个随机变量，在进行参数学习时，充分考虑到了参数本身的先验知识对参数本身的影响，所以贝叶斯估计从理论上要比极大似然估计来的更加合理[41]。下面将对贝叶斯估计做出详细的介绍。

介绍贝叶斯估计之前，首先说明一下后面出现的一些符号的含义：

$\theta=\{\theta_{ijk}|i=1,\cdots,n;j=1,2,\cdots,q;k=1,\cdots,r\}$，用 θ_{ij} 来表示 $\theta_{ij1},\theta_{ij2},...,\theta_{ijr_i}$，用 θ_i 来表示 $\theta_{i1},\theta_{i2},...\theta_{iq_i}$，$\theta_i$ 表示 $P(x_i \mid pa(x_i))$ 的参数，θ_{ij} 表示 $P(x_i \mid pa(x_i)=j)$ 的参数，其中 $P(x_i \mid pa(x_i))$ 和 $P(x_i \mid pa(x_i)=j)$ 分别是 x_i 的条件概率分布和所有的分布。假设参数 θ 的分布满足以下三个条件[42]：

(1) 不同变量的参数之间互相独立，即 $P(\theta)=\prod_{i=1}^{n}P(\theta_{i..})$。

(2) 对于任意变量 x，若 x 有 k 个不同的取值，那么这 k 个值的参数之间互相独立，即 $P(\theta_{i..})=\prod_{j=1}^{q_i}P(\theta_{ij})$。

(3) θ_{ij} 的分布服从狄利克雷分布。

从上面的三个条件可以得到 θ 的先验概率分布为 $\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1}$。

已知数据集 Data，由贝叶斯公式可得出以下公式：

$$P(\theta \mid D)=\lambda\cdot\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r}\theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1} \qquad (3\text{-}2)$$

公式 3-2 中 $P(\theta \mid D)$ 即是 θ 的后验概率分布，同时 $P(\theta \mid D)$ 也服从狄利克雷分布，见公式 3-3：

$$D[m_{ij1}+\alpha_{ij1}m_{ij2}+\alpha_{ij2},...,m_{ijr_i}+\alpha_{ijr_i}] \qquad (3\text{-}3)$$

该公式也满足条件 1 和条件 2 中的独立条件，因此，利用贝叶斯网络的可拆分性并通过公式 3-4 即可求出参数 θ 的估计值：

$$\theta_{ijk}^{*}=\int P(x_i=k \mid pa(x_i)=j,\theta_{ijk})P(\theta_{ijk} \mid D)d\theta_{ijk}=\int\theta_{ijk}P(\theta_{ijk} \mid D)d\theta_{ijk} \qquad (3\text{-}4)$$

另外，由于 $P(\theta|D)$ 服从狄利克雷分布 $D[m_{ij1}+\alpha_{ij1}m_{ij2}+\alpha_{ij2},...,m_{ijr_i}+\alpha_{ijr_i}]$，因此还能得到公式 3-5：

$$\theta_{ijk}^{*}=(m_{ijk}+\alpha_{ijk})/\sum_{k=1}^{r_i}\alpha_{ijk} \quad (3\text{-}5)$$

第三节　结构学习

BN 结构学习的思想是：已知一个训练样本集 D，找出所有符合该样本集的 BN 网络结构，然后从中选出最符合的一个作为最优的贝叶斯网络。这种全局搜索的思想虽然简单，但这其实是一个 NP 难问题，现实中无法直接使用。因此，关于如何在时间和空间复杂度可接受的情况下找出最优 BN 结构一直是研究的重点。

贝叶斯网络经过几十年的发展取得很多重大突破，其中在 BN 结构学习方面，国内外研究学者提出了众多的方法，其中使用最广泛，认知度最高的方法主要有两种：一种是基于搜索的方法，另一种是基于条件独立测试的方法。下面将主要介绍这两种方法。

一、基于评分的方法

基于评分方法的思想是把 BN 结构学习看成是一个优化问题[43]，首先定义一个评分函数，然后通过评分函数对数据空间中的各个变量进行打分，最后使用搜索算法找出打分最高的网络结构。评分最高说明该网络与原始数据集拟合最好。

通过上述基于评分搜索方法的思想我们可以把 BN 结构学习看做一个优化模型：M =（G,C,F）。该优化模型中 G 代表的是整个网络空间，网络空间中是数据集中所有节点之间的因果关系；C 为所有的约束条件；F 为指定的评分函数，用来评价贝叶斯网络的拟合程度，评分越高拟合度越好。

公式 3-6 给出了拥有 n 个变量的数据集可能含有的 BN 结构个数：

$$f(n)=\sum_{i=1}^{n}(-1)^{i+1}\cdot(n!/i!(n-i)!)\cdot 2^{i(n-i)}\cdot f(n-i) \quad (3\text{-}6)$$

上式中 f(n)为可能的结构个数，由此可见 BN 结构的个数随着节点个数的增加而成指数级增长，所以如何选定评分函数和搜索方法是打分算法的核心。

目前常用的评分函数主要有基于贝叶斯统计和基于信息论两种[44]。基于贝叶斯统计的打分算法主要有 K2 算法和 BD 算法，基于信息论的打分算法主要有 AIC 算法、BIC 算法和 MDL 算法。

（一）基于贝叶斯统计的评分算法

根据上一节介绍可知在贝叶斯统计中将未知变量作为随机变量来看待，同理，在 BN 结构学习中结构G和参数θ_G被当做两个随机变量。假设存在某个数据集$D=(d_1,d_2,...,d_m)$，该数据集中的样本含有 n 个变量，即$d_i=\{d_{i1},d_{i2},...,d_{in}\}$，结构 G 包含所有可能的有向无环结构，$G$和$\theta_G$的值一一对应。用概率分布$P(G)$来表示结构的先验分布；用概率分布$P(\theta_G)$来表示参数$\theta_G$的先验分布。

贝叶斯评分的核心思想是：已知训练样本，先计算出先验概率，然后再计算后验概率，从众多网络结构中选出后验概率最大的那个，即

$$G^*=\arg\max_G P(G|D) \tag{3-6}$$

由于$P(G|D)=P(D|G)P(G)P(D)$，$P(G)$和$P(D)$独立，因此公式 3-6 两边取对数后：

$$\log P(G,D)=\log P(D|G)+\log P(G) \tag{3-7}$$

公式 3-7 中$\log P(G,D)$为数据集 D 的某个结构的贝叶斯评分。$P(G)$为先验概率分布，通常情况下假设$P(G)$为均匀分布。将公式 3-7 变换后得到公式 3-8：

$$P(D|G)=\int P(D|G,\theta_G)p(\theta_G|G)d\theta_G \tag{3-8}$$

公式 3-8 中的$P(D|G)$为边缘似然函数。

K2 算法和 DB 算法均为基于贝叶斯统计的评分函数，这两个算法的区别在于参数服从的分布不同，下面将详细介绍这两种算法的区别。设存在一个数据集 D，D 中的任意一个样本均可由$\{X_1,X_2,...,X_n\}$表示，n 个变量均独立同分布，若存在 n 个变量的 BN 结构 G，而且$p(\theta_G|G)$为均匀分布，则 K2 评分函数见公式 3-9：

$$F_{K2}(G|D)=\log P(G)+\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log((r_i-1)!/(m_{ij^*}+r_i-1)!)+\sum_{k=1}^{r_i}\log(m_{ijk}!)\right] \quad (3\text{-}9)$$

若 $p(\theta_G|G)$ 服从狄利克雷分布（见公式 3-10）

$$p(\theta_G|G)\propto\prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}-1} \quad (3\text{-}10)$$

那么便可得到 BD（Bayesian Dirichlet）评分函数（见公式 3-11）：

$$F_{BD}(G|D)=\log(P(G))+\sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log(\Gamma(\alpha_{ij^*})/\Gamma(\alpha_{ij^*}+m_{ij^*}))+\sum_{k=1}^{r_i}\log(\Gamma(\alpha_{ij^*}+m_{ij^*})/\Gamma(\alpha_{ij^*}))\right] \quad (3\text{-}11)$$

公式3-11中α_{ijk}为Dirichlet分布中的超参值，对比公式3-11和3-9不难看出，当α_{ijk}等于 1 时 BD 评分函数将变成 K2 评分函数，所以 K2 评分函数是 BD 评分函数的一种特殊情况。

（二）基于信息论的评分算法

在信息论中有一种度量方法是最小描述长度法（MDL），该方法的基本原理来自于数据的存储。假设存在某个数据集，现在需要将其保存，但是由于该数据集占用空间大，为了节约存储空间首先要对其进行压缩，比如使用某种压缩算法，然后再进行存储。此外，由于数据进行算法压缩后，数据结构已然改变，当我们需要该数据时需要将被压缩的数据还原，因此在存储压缩后数据的同时还需要一同保存压缩时使用的算法，另外还需保存用来描述恢复数据所需的存储空间，简称算法描述长度，所以真正所需的存储空间是压缩后的数据空间加上描述长度，总称描述长度。而 MDL 方法的原理就是要从众多的压缩算法中找出描述长度最短的算法。

基于信息论的评分算法就是按照 MDL 的原理进行评分的。在 BN 结构学习中[45]，MDL 算法通过寻找具有最小描述长度的结构来确定最优的贝叶斯网络结构，但是这种方法显然存在着缺点：偏向于寻找结构简单的网络结构。因此需要在这个最小长度上添加一个惩罚项，用来平衡结构复杂度和准确度，保持总体的合理性。在传统的 MDL 算法中，惩罚项取决于参数的个数，惩罚函数见公式 3-12：

$$C(G)=\frac{1}{2}\log m\sum_{i=1}^{n}(r_i-1)q_i \quad (3\text{-}12)$$

上式中 m 为数据集中的样本数；$\sum_{i=1}^{n}(r_i-1)q_i$ 为网络结构中参数的个数。公式 3-13 表示使用编码压缩后的数据长度，

$$LL_D(G)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}}) \quad (3\text{-}13)$$

根据公式 3-12 和 3-13 便可得到 MDL 评分函数，见公式 3-14：

$$F_{MDL}(G\mid D)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}})-\frac{1}{2}\log m\sum_{i=1}^{n}(r_i-1)q_i \quad (3\text{-}14)$$

MDL 打分方法的特点是该方法不需要变量的先验知识，而且当数据量足够并且相互独立时，通过 MDL 打分算法得出的最优网络结构理论上能够逼近任何样本分布。

当数据集服从多项式分布时 MDL 打分函数和 BIC 打分函数相等。

对公式 3-14 做适当简化便可得到 AIC 评分函数，见公式 3-15：

$$F_{AIC}(G\mid D)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}})-\sum_{i=1}^{n}(r_i-1)q_i \quad (3\text{-}15)$$

另外基本 BN 具有可分解的特征，人们提出了一种新的评分函数 MIT，该方法和 MDL 的区别在于 MIT 并不是将网络中所有的参数和数据用作惩罚项，而是根据互信息、卡方分布选取其中的一部分作为惩罚项，评分函数见公式 3-16：

$$F_{MIT}(G\mid D)=\sum_{i=1,Pa(X_i)\neq\phi}^{n}\left(2mMI_D(X_i,Pa(X_i))-\max_{\sigma_i}\sum_{j=1}^{s_i}\chi_{\alpha,l_{i\sigma_i(j)}}\right) \quad (3\text{-}16)$$

上式中 $MI_D(X_i,Pa(X_i))$ 表示子节点与其其父节点之间的互信息；S_i 表示节点 i 的父节点个数；$\chi_{\alpha,l_{i\sigma_i(j)}}$ 为卡方分布值，其中 $l_{i\sigma_i(j)}$ 为自由度，α 为置信度。

（三）搜索策略

选定了评分函数后的下一步工作便是如何制定搜索方法。在 n(n>1)个节点中寻找最优结构是个典型的 NP 难问题[46]，因此无法采用全局搜索的方式。对于这种传统的 NP 难问题，常用的解决办法有贪婪搜索法，首先人工选定一个初始的 BN 结构，利用评分函数进行打分，然后随机调整 BN 结构，对调整好的结构进行打分，比较两个分数的高低，如果调整后评分高于不调整则更新结构，否则不做更新继续进行随机调整直到出现符合更新条件的结构为止。这种搜索策略原理简单易懂并且容易实现，但是贪婪搜索的缺点显而易见，得到的结构往往是局部最优的，并不是全局最优。由于该问题属于一个典型的最优化问题，因此研究人员

参考了传统的组合优化问题并将一些优化算法应用到 BN 的结构学习中，比如有爬山法、模拟退火算法、遗传算法等等。相关算法详见表 3.1：

年份	算法	搜索空间	评分函数	搜索策略
1968	Chow-Liu	Tree	Entropy	--
1992	K2	DAG	K2	Hill Climbing
1994	Lam-Bacchus	DAG	MDL	Heuristic Search
1995	GHC	DAG	BD	Hill Climbing
1996	GA-O	Ordering	K2	Genetic Algorithm
1996	BA-B	DAG	K2	Genetic Algorithm
1996	Suzuki	DAG	MDL	Branch and Bound
2002	GES	PDAG	BDeu	Greedy Search
2002	ACO-B	DAG	K2	Ant Colony Optimization
2003	EDA-B	DAG	Entropy	Estimation of Distribution
2004	HEA	DAG	MDL	Evolutionary Algorithm
2009	ACO-E	PDAG	BIC	Ant Colony Optimization

表 3.1 基于评分的 BN 结构学习

二、基于条件独立检验方法

在概率论中，若存在变量 A 和 B 在条件 X 下相互独立，即 $P(AB|X)=P(A|X)\cdot P(B|X)$，则称 A 和 B 关于 X 条件独立。条件独立性在多元统计中有着非常重要的意义[47]。基于条件独立检验（CIT）方法的思想是将贝叶斯网络看成是表达节点之间关系图的模型[48]。假设在贝叶斯网络中存在任意节点 A 和 B,若节点 A 和 B 之间存在有向边,那么我们称节点 A 和 B 之间有关系；若节点 A 和 B 之间不存在有向边，那么称节点 A 和 B 之间条件独立。CIT 方法在原理上更加贴近于贝叶斯网络,一般采用卡方分布值或者互信息来衡量变量之间的关系。CIT 方法的思想是给定一个无向完全图，根据相关算法来判断节点之间的是否存在边，若存在则保留否则删除该边，下面将详细介绍 CIT 方法。

定理 3.3 存在数据集 D， $X=\{X_1,X_2,...,X_n\}$ 为 D 中的变量，如果 $Ind(X_i,X_j\,|\,X_k)$ 成立，那么存在等式：

$$U_{ij|k}^2 = 2\sum_{a,b,c} N_{ijk}^{abc} \log\left[\frac{N_{ijk}^{abc} N_k^c}{N_{ik}^{ac} N_{jk}^{bc}}\right] \tag{3-17}$$

等式 3-17 中，$U_{ij|k}^2$ 为数据集的统计信息，N_{ijk}^{abc} 是指在数据集中满足 $X_i = a, X_j = b, X_k = c$ 的样本个数，r_i 为变量 i 不同值的数量。通过比较 $U_{ij|k}^2$ 和 χ_α^2 的大小即可得到节点 i 和 j 是否独立。

综上所述，独立检验方法运用的是统计理论中的知识，但是这种方法的缺点有两个：第一，该方法通过假设检验的方法来判断独立性，因此具有一定的误差；第二，在大数据环境下，数据巨大并且数据质量不高时该方法的效率低下并且准确度也难以保证。目前在 BN 结构学习中如何综合运用独立检验法和随机搜索法已经成为一个新的研究方向。

第四节　贝叶斯网络推理

在传统的统计学中，推理方法通常用于对抽取的特征变量进行相关性分析，从而得到变量之间的关联关系。常见的推理方法有回归分析等，该方法可以根据特征变量之间的相关性通过极大似然函数来预测未来事件发生的概率，另外回归分析还可以及时更新数据源。

BN 推理是指在给定相关节点（变量）、节点之间的网络结构和节点之间的条件概率的前提下，求解目标节点的边缘概率及预测子节点等问题。BN 推理比传统的推理方法更加的科学合理。BN 推理模型统一了多个变量之间的因果关系，这比只进行简单的相关性分析来的更加的重要。

BN 推理是贝叶斯网络研究领域中的一个重要分支。BN 推理是 BN 应用中的核心问题，因为应用中所需的结果就是 BN 推理的结果，另外，在很多的应用中推理算法是很重要的一个中间环节。目前，BN 推理已在多个领域有着成功的应用，比如生物学、信息学等等。由于贝叶斯网络推理面临着一个重要的难题：NP 难问题，因此多年来研究人员在 BN 推理的准确性方面发明了众多的方法。

一、变量消元算法

变量消元法(VE)通过变量之间的相关性将原来的计算过程进行分解，从而达到降低计算的复杂度，该方法的思想来源于动态规划问题[49]。变量消元法利用链式乘积法以及变量之间的条件独立性等特征把原来的计算过程拆分成带参数的

条件概率的乘积，并对拆分后的等式进行相应的转换，改变了计算的顺序，消去某些变量从而达到减少计算复杂度的目的。VE 算法巧妙了利用了公式中的数学特性，在解决复杂 BN 等推理问题时有着良好的表现。

由于变量消元法的简单易懂而且在很多问题上具有通用性，近年来一直是研究的热点。到目前为止，研究人员提出了桶消元法等一系列基于消元思想的算法，这些算法之间的最大区别在于如何找出最好的消元次序。最优消元次序问题是一个 NP 难的问题，目前寻找最优次序的算法主要是最小度法和最小缺陷法。

最小度法是找出网络中出度入度之和最小的节点，将该节点设置为最后一个消去的节点，然后在剩下的节点中继续寻找出度入度之和最小的节点，如此循环直至网络结构中没有节点为止。最小缺陷法的思想类似于最小度法，只不过在最小缺陷法中寻找的不是度最小的节点而是寻找缺边数最少的那个节点。

二、联接树算法

在众多 BN 精确推理算法中，联接树算法由于其效率高、能处理单连通和多连通问题等优点而被广泛的使用。联接树方法（JT）的思想是将 BN 转化成一棵树，树中节点由对应的无向边连接，然后设定树中进行消息传递的规则，根据设定的规则来进行相关概率计算。

消息传递总共有两个阶段，第一个阶段是信息收集，首先定义一个根节点，从离根节点最远的节点开始往根节点方向传递信息，沿途经过的节点接收到信息时对自身的概率值进行更新，直到消息传递到根节点为止，如此反复直到根节点接收到所有信息；第二个阶段是扩散，该阶段的运行方式正好和第一阶段相反，将根节点的信息依次扩散出去，直到所有节点都收到根节点的信息为止。图 3.4 中，信息通过上述两个阶段后树中所有节点的信息均保持同步。

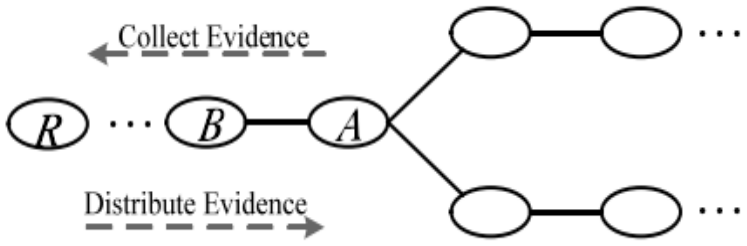

图 3.4 联接树消息传递

目前寻找最优联接树是一个 NP 难问题，主要的解决办法类似于最优规划问题：采用启发式算法。当 BN 中节点数较多时往往牺牲精确度来提高计算的效率。

三、随机抽样算法

随机抽样算法[50]的基本思想是从大的数据样本中按照某种指定的概率分布来进行样本收取，将抽取的样本组成一个新的小样本集，然后在小样本集上进行计算。随机抽样算法不考虑原始样本的分布以及样本之间是否独立，计算得出的结果是原始数据结果的近似值，因此该算法属于近似算法中的一种。随机抽样算法虽然得到的结果并不精确，但是通过对该算法进行改进往往能够得到非常近似的结果。

目前随机抽样算法主要有两种：一种是重要性抽样，得到的抽样样本之间独立；另一种是 Markov 抽样，得到的抽样样本之间不独立。由于随机抽样算法能够显著降低计算的时间复杂度并且得到的结果也能非常节点真实结果，因此该算法成为了今年来研究的热点。

四、变分近似算法

变分近似算法[51]是将概率推理问题通过变分法转化后变成变分优化问题。将概率推理问题转化成变分优化问题后难度其实并没有降低，但是变分优化问题往往能够对优化条件进行简化，这样就能通过放松某些条件从而降低问题复杂度，进一步便能得到概率推理问题的近似解。目前常用的变分近似算法有朴素平均场算法和循环传播法。

朴素平均场算法来源于统计力学，是一种常用的近似算法。该算法的主要思想是在拥有多个节点的贝叶斯网络中，每个节点均由和该节点的临节点的“平均场”来代替该节点的“场”，然后在平均后的“场”上进行计算。这种算法能平滑复杂的贝叶斯网络，从而达到简化的目的，得到相应的近似结构。

人们在多连通贝叶斯网络中使用信息传递算法时常常发现由于信息一直在节点中循环传播而无法收敛的问题，因此 Mur 等人提出了信息传递算法的改进方法：循环传播法。该算法的主要思想是迭代使用信息传递算法，利用得到的信息来计算各节点的后验概率分布的近似值，若某个节点一直在循环传递信息而无法收敛时便形成震荡。

第五节　小结

本章主要介绍了贝叶斯网络的基本概念，分别从贝叶斯网络的基本模型、参数学习、结构学习以及贝叶斯网络推理等几个方面做了详细的介绍。在介绍的过

程中不仅详细的阐述了基本原理而且简单的概括了各个领域多年来发展的情况以及现状。

目前，在不确定知识表示和推理问题研究等领域，贝叶斯网络因其丰富的语义结构、深厚的概率统计理论基础以及独特的推理优势成为最有效的模型之一，并成为了近几年来的研究人员关注的热点。贝叶斯网络已经在医疗、生物、社交网络等多个领域成功应用。

第四章 时间序列数据获取与预处理

时间序列[52]指的是由同一个现象在不同时间点上的连续观察值排列而成的一组数字序列，例如某一只股票的收盘价格、高铁客流量、某商业街的人流量、某个月的降水量、水流量等等，都形成了一个个不同的时间序列。

从统计学的角度上来讲，时间序列指的是将某一个维度在不同时间点上的不同数值，按照时间的先后顺序排列而成的数据[53]。因此，时间序列常常受到各种不确定因素的干扰而表现出一定的随机性，数据之间往往存在一定的相关性。

从数学的角度上来讲，随机序列指的是由一系列随机变量组成的数组，如$\{X_1, X_2, ..., X_n\}$，我们用x_t来表示，其中 t=1,2,3,…,n。随机序列还有另外一种定义，即在多维空间中的一个随机向量。时间序列是随机序列的一种特殊情况。时间序列是按照时间的顺序来排列的，因此上面的表达式中 t 为时间的整数变量，用来表示等间隔的增长，比如第 t 时间点、第 t 月、第 t 个等等，我们用x_t来表示，其中 t=1,2,3,…,n，这里的 t 表示时间的顺序。另外还有一点不同的是在时间序列中变量 t 既可以为正数也可以为负数，这是由于时间序列都是以当前的时间为基准，若 t 为负数则说明该数据发生在当前时间点之前，若 t 为正数则说明该数据发生在当前时间点之后，但是 t 的值必须为整数。

事实上，时间序列不一定需要按时间顺序排列，也可以按照其他物理量顺序排列的随机数据。这里的“时间”指的是广义上横坐标的值。

综上所述，时间序列一般具有如下特点:首先，序列的数据或者数据点的位置必须依赖于时间，即数据的取值必须随着时间的变化而变化，但不一定是时间 t 的严格函数。其次，每一个“时间”点上的取值或数据点的位置具有一定的随机性，不可能完全准确地拟合历史值。最后，不同时刻的数值或数据点的位置有一定的相关性，这种相关性称为系统的动态规律性。

第一节 期货时间序列数据的来源

期货交易数据都是按照时间顺序观察收集得到的期货合约交易的价格数据，例如开盘价，收盘价，最低价，最高价等等。从期货交易软件上连续获得的期货

时间序列数据称作期货时间序列数据流[54]。时间序列数据有连续的，也有离散的。期货时间序列数据是一组随时间变化而观察得到的价格数据，该数据是离散的。我们假设

$$S=\{(x^1,y^1),(x^2,y^2),...,(x^n,y^n)\}$$

为时间序列，n 表示数据点的个数，x^i 表示买卖期货合约时的价格数据,其中 i∈[1,n]，表示买卖期货合约的时间点;给定一个维度，规定该时间序列是某种价格数据，例如开盘价，收盘价，最高价等等。通常情况下，使用自动的期货交易平台进行交易的时候，只需要分析众多维度中的一种期货时间序列数据即可，即固定属性，例如期货收盘价时间序列数据。但是在实际应用中，时间点往往由特殊的方法来标识，在这里用一个一维数组来表示某个期货时间序列：$S(n)=\{x_1,x_2,...,x_n\}$，时间点用数组下标 i 标识。

本文采用的期货时间序列数据来自 UCI 网站上的开源数据，选取了其中 3 份期货交易时间序列。每一份时间序列拥有 1000 万左右的数据样本。每个样本拥有交易日期、该期货的 ID、更新时间、实时价格、开盘价、收盘价、最高价、最低价等维度，其中更新时间从几秒至几分钟不等。下表给出了部分数据样本及部分维度信息：

表 1 某期货数据样本

交易日期	期货 ID	更新时间	实时价格	收盘价
20130930	y1408	15:00:17	6978	6978
20130930	y1408	14:59:37	6978	6978
20130930	y1408	14:59:03	6967	6978
20130930	y1408	14:58:43	6963	6978
20130930	y1408	14:58:14	6965	6978
20130930	y1408	14:58:02	6951	6978
20130930	y1408	14:57:49	6945	6978

第二节　编程语言及相关包介绍

一、Python 简介

Python 是一门面向对象的计算机编程语言，1989 年由 Guido van Rossum 发明，发展至今已有 20 多年历史。Pyhton 是一种解释型语言，无需编译。Python 包含有强大的标准库，语法简单、人性化而且成熟稳定，能够完成很方便的完成常见的作业。

Python 的语法是其亮点，语法简单明了，非常的人性化，初学者很容易理解其语义。Pyhton 采用缩进的方式来规定语句，这一点和其他语言均不相同，另外它还有高效的顶层结构，能够快速而又简单地实现编程。

Python 因其语法简单易懂、支持动态输入及解释型语言等特性使其成为一门很受欢迎的脚本语言，尤其适合用于快速的应用程序开发。Python 能够支持的编程范式有命令式、面向对象式、函数式、面向切面式、泛型式等等，它和其他经典的动态语言一样具有垃圾自动回收功能而且能够自动管理内存。

Python 是一门开源的语言，它的功能库均有开发者开发维护，目前 Python 的官方功能库由 Python 管理协会管理。

二. Pandas 和 NumPy 包简介

Pandas 是 Python 的一个功能库，专门用于数据分析，是一个开源的库，目前由 PyData 开发和维护。最初研究人员开发出 Pandas 库是为了方便做金融领域的数据分析，因此一开始 Pandas 只为时间序列数据分析提供了相应的分析功能，但是随着 Python 的普及，其他领域的数据分析需求日益扩大，因此 Pandas 的功能也一直被扩展和完善，目前 Pandas 库已经能够实现大部分常用的数据分析工作。

NumPy 也是 Python 的一个功能库，专门用来进行数学计算，它可以存储和处理大规模矩阵，比 Python 自带的列表来的更加高效。NumPy 可以用来进行严格的数字处理工作，很多大型金融公司用 NumPy 来处理相关的数字计算。

由于本文研究的时间序列数据量较大，传统的数据分析工具已经无法满足效率要求，所以决定采用 Python 中的 Pandas 和 NumPy 包来进行相关的数据处理，并使用 Python 语言来实现模型。

第三节　原始数据的降维、缺失值处理及离散化

一、数据降维

本文分析的期货时间序列数据来自三份拥有 1000 万个样本的期货数据，每个样本记录的维度有开盘价，收盘价，最高价，当前价格等等。本文研究的目标是该期货在不同的时间点上价格变化的因果关系，如果将原始数据全部读入计算机进行计算显然是不科学的，会造成大量计算资源的浪费，并导致效率低下，所以处理原始数据的第一步工作就是对原始数据集进行降维，只保留当前价格维度。

本文采用的降维工具是 Python 中的 Pandas 包，利用该包可以实现将原始数据集读入计算机并生成一个数据框（DataFrame），然后使用 Pandas 中的命令选取当前价格维度的数据。

二、数据缺失值填充

在处理原始数据集时发现数据样本中部分数据因某些原因导致丢失，我们采用平均值填充的方法来填充缺失值。设某时间序列为 $S=\{x_1,x_2,x_3,...\}$，如果其中 x_2 由于某种原因缺失，那我们采用缺失值前后非空数值的平均值来填充 x_2。考虑到单个数据缺失和连续多个数据缺失的情况，算法伪代码如下所示：

算法 1：缺失值填充算法

INPUT:降维后的数据集 dataSet

OUTPUT:缺失值填充后的数据集 dataSet_Filled

BEGIN

4 **FOR** $\forall x_i \in dataSet$:

5 **IF** X_i != NULL **THEN** CONTINUE

6 **ELSE IF** X_i = = NULL **AND** X_{i+1} != NULL

7 **THEN** $X_i = (X_{i-1} + X_{i+1})/2$

8 **ELSE** 查看下一个数据是否为空，直至找到不为空的数值 X_k

9 **THEN** $X_i = (X_{i-1} + X_k)/2$

10 **RETUEN**

11 **END**

三、数据离散化

经过上述预处理之后的时间序列是一组连续的数据，代表了某期货在不同时间点的价格，由于我们研究的目标是变量之间上升和下降之间的因果关系，而并不关心变量在某一时刻的具体值，所以需要将连续变量进行离散化处理。

定义 4.3 设原始时间序列为 $s_1 = \{x_1, x_2, x_3, ..., x_n\}$，则经过离散化之后的时间序列为：

$$X = \{x_1, x_2, x_3, ..., x_i, ..., x_n\}$$

其中

$$x_i = \begin{cases} 1 & x_i > x_{i+1} \\ 2 & if \quad x_i < x_{i+1} \\ 3 & x_i = x_{i+1} \end{cases}$$

经过上述数据离散化处理后，原始数据集将变成全部由 1,2,3 三个离散数字表示的数据集，如下图所示：

离散化之后的实时价格时间序列：								
实时价格	1	2	2	3	……	1	3	3……

图 4.3.3 离散化之后的时间序列

四、贝叶斯网络数据集的构建

原始的期货数据集经过上述的维度清洗和离散化处理后变成了由 1000 万个离散数据组成的时间序列，但是这样的大时间序列是无法直接进行贝叶斯网络训练的，因为要进行贝叶斯网络训练的数据必须是矩阵的形式，所以接下来就要进行数据集的构建。本文的贝叶斯网络分为一阶和二阶，两种贝叶斯网络的数据集构建方式不一样，本节主要叙述如何构建一阶贝叶斯网络的数据集。

在经过离散化处理后的时间序列上随机（初始点不放回）截取 5000 个连续的时间序列片段，每个片段拥有 5000 个节点。随机截取时间序列的算法伪代码如下所示：

算法 2：时间序列随机截取算法

```
1   INPUT:离散化处理后的时间序列 seriesData
2   OUTPUT:5000*5000 的矩阵（每一行为一个时间序列）
3   BEGIN
4   SET record = []          //record 为一个列表，用来存放随机生成的数
5   SET matrix = []          //matrix 用来存放随机截取的时间序列矩阵
6   FOR ∀i ∈ (0,5000)
7   WHILE  x ∉ record
8                   x = random(0,length(seriesData)-5000)
9               将 x 添加到 record 中
10              从 seriesData 的第 x 个位置开始连续获取 5000 个数值作为矩阵
```

```
11    的一行存入 matrix 中
12    RETURN matrix
13    END
```

经过上面的随机截取后获得了 5000 个长度为 5000 的时间序列，下面将对每一个时间序列进行数据集的转化。

由于考虑到二阶贝叶斯网络训练时计算机的运行能力，所以在构建一阶贝叶斯网络数据集时不宜选取过多的节点，在经过多次试验后决定选取 5 个节点，从时间复杂度的角度出发选取 5 个节点比较合适，方便实验。

设初始时间序列为 $S=\{x_1,x_2,x_3,x_4,x_5,x_6,...\}^T$，将该时间序列转化成拥有 5 个变量的数据集，转化过程如下图所示：

$$\{x_1,x_2,x_3,x_4,x_5,x_6,...\}^T \Longrightarrow \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ x_2 & x_3 & x_4 & x_5 & x_6 \\ x_3 & x_4 & x_5 & x_6 & x_7 \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{matrix}$$

全部转化完成后得到 5000 个数据集，每个数据集都为 4996 行，5 列。

第四节　小结

采用贝叶斯网络进行模型训练对输入的数据集有着较高的要求，本文的原始数据集来自期货时间序列，该数据集中有较多的缺陷，例如数据不完整，冗余数据过多等问题，所以试验的第一步就是进行数据预处理，保证后期实现的高效和准确。另外，由于本文采用了一种全新的贝叶斯网络训练方法，需要大量的时间序列，所以需要对原始时间序列进行随机截取来产生足够多的小时间序列，保证模型的实现。最后，贝叶斯网络训练需要的是矩阵形式的数据集，而单变量的时间序列是一个数组，需要考虑如何将数组转化成符合贝叶斯网络训练的数据集。

第五章 基于贝叶斯网络的因果关系挖掘

传统的一阶贝叶斯网络是在小数据集上挖掘节点与节点之间的因果关系，因此，发现的因果关系仅限于节点与节点之间，具有一定的局限性。但是当数据量大到一定程度时，研究边与边之间的因果关系变得十分有意义。因此本文提出了对大时间序列构建二阶贝斯网络模型来挖掘边与边之间的因果关系，并着重研究如何实现二阶贝叶斯网络模型，并对模型得出的因果关系进行解释。

在基于贝叶斯网络理论学习的基础上，为了更好的挖掘时间序列中隐藏的因果关系，本文主要分两大模块进行模型的构建与仿真，即基于贝叶斯网络最小描述长度得分搜索学习算法的单时间序列模型构建与仿真和多时间序列模型构建与仿真。模型学习的框架如图 5.1 所示:

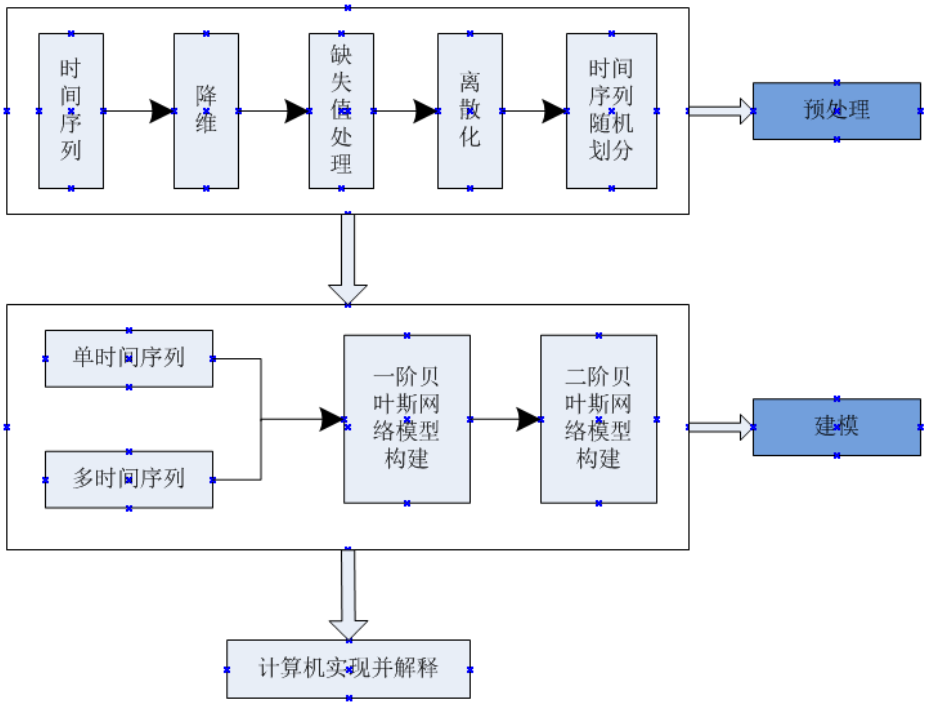

图 5.1 模型学习框架图

由于本文研究的重点是实现二阶贝叶斯网络模型并研究边与边之间的因果关系，在对一阶贝叶斯网络训练时做了适当的简化。在实现一阶贝叶斯网络模型时规定只寻找节点 5 的父节点，这样在进行二阶模型训练后就可以方便的找出边与边之间的因果关系，大大简化了实验的复杂度。

第一节 MDL 打分算法

在信息论中有一种度量的方法是最小描述长度法（MDL），该方法的基本原理来自于数据的存储。假设存在某个数据集，现在需要将其保存，但是由于该数据集占用空间大，为了节约存储空间首先要对其进行压缩，比如使用某种压缩算法，然后再进行存储。此外，由于数据进行算法压缩后，数据结构已然改变，当我们需要该数据时需要将被压缩的数据还原，因此在存储压缩后数据的同时还需要一同保存压缩时使用的算法，另外还需保存用来描述恢复数据所需的存储空间，简称算法描述长度，所以真正所需的存储空间是压缩后的数据空间加上描述长度，总称描述长度。而 MDL 方法的原理就是要从众多的压缩算法中找出描述长度最短的算法。

基于信息论的评分算法就是按照 MDL 的原理进行评分的。在 BN 结构学习中，MDL 算法通过寻找具有最小描述长度的结构来确定最优的贝叶斯网络结构，但是这种方法存在着明显的缺点：偏向于寻找结构简单的网络结构。因此需要在这个最小长度上添加一个惩罚项，用来平衡结构复杂度和准确度，保持总体的合理性。在传统的 MDL 算法中，惩罚项取决于参数的个数，惩罚函数见公式 5-1：

$$C(G)=\frac{1}{2}\log m\sum_{i=1}^{n}(r_i-1)q_i \tag{5-1}$$

上式中 m 为数据集中的样本数；$\sum_{i=1}^{n}(r_i-1)q_i$ 为网络结构中参数的个数。公式 5-2 表示使用编码压缩后的数据长度，

$$LL_D(G)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}}) \tag{5-2}$$

根据公式 5-1 和 5-2 便可得到 MDL 评分函数，见公式 5-3：

$$F_{MDL}(G\mid D)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}})-\frac{1}{2}\log m\sum_{i=1}^{n}(r_i-1)q_i \tag{5-3}$$

MDL 打分方法的特点是该方法不需要变量的先验知识，而且当数据量足够时，通过 MDL 打分算法得出的最优网络结构理论上能够逼近任何样本分布。

当数据集服从多项式分布时 MDL 打分函数和 BIC 打分函数等价。

对公式 3-14 做适当简化便可得到 AIC 评分函数，见公式 5-4：

$$F_{AIC}(G\mid D)=\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{r}m_{ijk}\log(\frac{m_{ijk}}{m_{ij^*}})-\sum_{i=1}^{n}(r_i-1)q_i \tag{5-4}$$

另外基于 BN 具有可分解的特征，人们提出了一种新的评分函数 MIT，该方法和 MDL 的区别在于 MIT 并不是将网络中所有的参数和数据用作惩罚项，而是根据互信息、卡方分布选取其中的一部分作为惩罚项，评分函数见公式 5-5：

$$F_{MIT}(G \mid D) = \sum_{i=1, Pa(X_i)\neq\phi}^{n} \left(2mMI_D(X_i, Pa(X_i)) - \max_{\sigma_i} \sum_{j=1}^{s_i} \chi_{\alpha, l_{i\sigma_i(j)}} \right) \quad (5\text{-}5)$$

上式中 $MI_D(X_i, Pa(X_i))$ 表示子节点与其其父节点之间的互信息； S_i 表示节点 i 的父节点个数； $\chi_{\alpha, l_{i\sigma_i(j)}}$ 为卡方分布值，其中 $l_{i\sigma_i(j)}$ 为自由度， α 为置信度。

第二节　构建贝叶斯网络模型

在贝叶斯网络中，对期货时间序列进行贝叶斯网络的推理，即贝叶斯网络的因果关系挖掘，其主要思想是:在给定某些证据变量取值的条件下，求解给定变量和目标变量之间的因果关系。

现阶段有关贝叶斯网络推理问题主要分为三种方法:(1)后验概率问题;(2)最大后验假设问题;(3)最大可能解释问题。一般而言，贝叶斯网络推理问题是 NP 难问题，但在现实应用中，需根据贝叶斯网的结构特点，仍可采用有效地推理算法进行推理。在本节中，我们利用最小描述长度（MDL）来对单时间序列（期货）进行节点与节点之间，边与边之间的因果关系挖掘，并通过实验给出具体的图结构和相应的解释。下图为模型训练框架图：

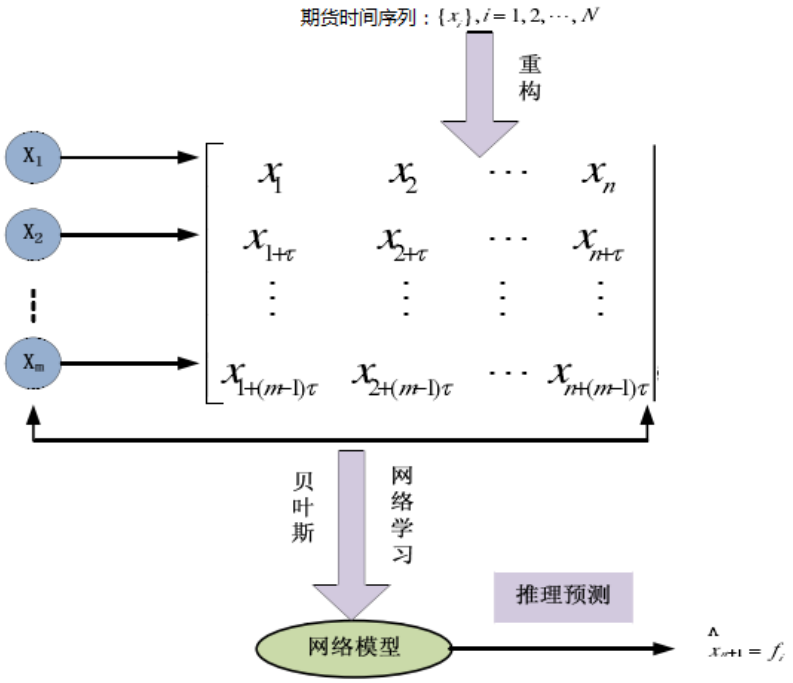

图 5.2 模型训练框架

一、一阶贝叶斯网络模型

经过第四章的数据预处理和数据集构建后得到 5000 个数据集。对于每一个数据集都需要采用 MDL 打分标准来找出节点的父节点。由于 5000 个数据集的模型训练方式都一样，只要循环训练 5000 次便可得到相应的因果关系，所以下面只拿其中的一个数据集来进行研究。

由于无论是一阶还是二阶贝叶斯网络训练中都需要使用 MDL 打分准则，所以下面给出了 MDL 打分算法实现的步骤及伪代码：

MDL 打分算法实现的核心步骤：

（1）得到一个可能的父节点并计算该与该父节点有关的所有可能的先验概率。

（2）计算该父节点的惩罚项的值。

（3）计算机该父节点的似然项。

（4）对（2）和（3）中的值进行计算得到该父节点的 MDL 值并保存。

（5）根据步骤（2）-（4）计算机子节点的 MDL 值。

（6）将第（4）步中的值与子节点的 MDL 值进行比较，若小于子节点 MDL 值则该

MDL 值对应的节点便是子节点的父节点。

MDL 打分算法伪代码如下：

算法 3：MDL 打分算法

INPUT:可能的父节点集合 fatherNode；子节点集合 childrenList

所有可能的先验概率 countList；数据集 dataSet

OUTFUT:fatherNode

BEGIN

SET $C(G)$= 0

SET $\forall x_i \in fatherNode$

π_i = $\prod_{i=1}^{k} |y_i|$ //$|y_i|$ 表示子节点维数

//其中 $y_i \in childrenList$,k 为子节点个数

$C(G)$ = $\frac{1}{2}\log N \cdot |x_i| \cdot \pi_i$

SET $LL_D(G)$= 0

FOR $\forall x_i \in dataSet$

计算 $P(x_k|\pi_k, G)$ 联合概率的值

$LL_D(G)$ = $LL_D(G)$ + $\log P(x_k|\pi_k, G)$

MDL = $C(G)$ - $LL_D(G)$

RETURN MDL

END

有了 MDL 打分算法的实现后，我们便可以进行一阶贝叶斯网络的训练，利

用 MDL 打分算法求出各个数据集中节点 5 的父节点。下面将给出一阶贝叶斯网络训练核心步骤：

（1） 统计各个节点的先验概率和维数。

（2） 由近到远的计算子节点之前的节点是否为其父节点。若是，则进入步骤（3），否则进入步骤（4）。

（3） 保存该父节点，将该父节点和子节点合并，形成新的子节点，重复步骤（2）。

（4） 节点往前移一个，进入步骤（2）。

（5） 所有可能的父节点都计算完成后结束，画出节点之间的关系图。

一阶贝叶斯网络训练伪代码如下：

算法 4：一阶贝叶斯网络算法

```
1   INPUT：数据集 dataSet
2   OUTPUT：父节点
3   BEGIN
4   Dim   fatherNodeList = [4,3,2,1]
5   Dim   childNodeList = [5]
6   Dim   realFatherList = []
7       计算出各种可能情况的条件概率的值并保存到 countList 中
8   FUNCTION   Father(fatherNodeList, childNodeList, countList, dataSet){
9   FOR  ∀x ∈ fatherNodeList
10                  //调用 MDL 打分算法，传入相应的参数
11                  mdl = Function   MDL(x,children,dataset,countList)
12                  保存 mdl 值到集合 mdlList 中
```

```
13              minMdl = MIN(mdlList)          //从集合中找出最小的 mdl 值
14              //计算无父节点时的 MDL 值
15              MDL = Function    MDL(children,children,dataset,countList)
16  IF    minMdl < MDL
17                      将该节点保存到 realFatherList 中
18                      从 fatherNodeList 中移除该节点
19                      将该节点加入 childNodeList 中
20  ELSE
21                      从 fatherNodeList 中移除该节点
22              Father(fatherNodeList, childNodeList, countList, dataSet)    //递归
23  Return    realFatherList
24      }
25      //调用函数
26      realFatherList    = Father(fatherNodeList, childNodeList, countList, dataSet)
27  END
```

利用上述模型对 5000 个数据集进行循环训练，训练完成后将得到 5000 个贝叶斯网络，将其节点和节点之间的关系保存，以方便后面建立二阶贝叶斯网络时使用。

二、二阶贝叶斯网络模型

在进行二阶贝叶斯网络训练时首先要构建数据集，和一阶贝叶斯网络构建数据集的方式不同，二阶贝叶斯网络的数据集来自一阶贝叶斯网络图结构，一个一

阶贝叶斯网络可以转化成一条记录，5000 个贝叶斯网络图就可以转化成 5000 条记录。

5000 个数据集进行一阶贝叶斯网络训练后，每一个数据集都将得到一个节点之间的因果关系图，如图 4.4.2 所示：

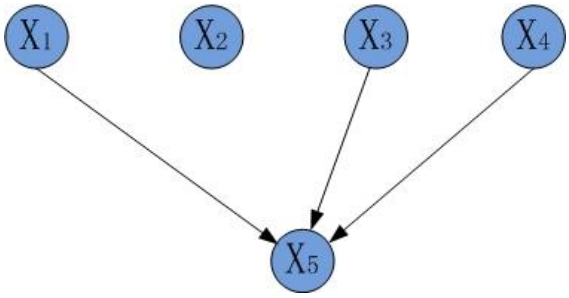

图 5.3 一阶贝叶斯网络结构图

从上图中可以看到节点 1、3、4 是节点 5 的父节点，为了进行二阶贝叶斯网络训练，需要将一阶贝叶斯网络结构图转化成一条记录。

定义 4.2　若节点 i 是节点 5 的父节点则记为 1，否则记为 2。

对 5000 个单层贝叶斯网络模型训练后得到了 5000 个网络拓补结构图，将这些图结构按定义 4.2 转化成数据集，这样 5000 个图结构就能转化为 5000 条记录。下图为图结构转化成一条记录的过程：

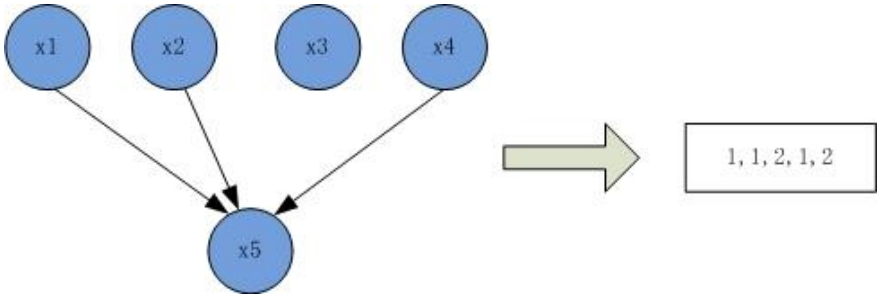

图 5.4 结构图转化为记录

按上述过程将 5000 个贝叶斯网络图转化成数据集之后便能进行二阶贝叶斯网络模型训练。二阶贝叶斯网络模型和一阶贝叶斯网络模型的不同之处在于：不再只找出节点 5 的父节点，而是要找出所有节点的父节点。下面是二阶贝叶斯网络模型实现伪代码：

算法 5：二阶贝叶斯网络模型

```
INPUT：数据集 dataSet
OUTPUT：每个节点的父节点 fatherNode
BEGIN
Dim fatherList = [1,2,3,4,5]
Dim childList = [1,2,3,4,5]
FOR ∀x ∈ childList
        从 fatherList 中删除 x
        调用算法 4（参数：x, fatherList, dataSet，countList）
        将返回值和对应的节点保存至 fatherNodeDict 中
Return fatherNodeDict
END
```

第三节　贝叶斯网络模型实验结果及解释

在实验阶段我们将实验分为两部分：一部分是单时间序列，数据集由一个期货时间序列构成；另一部分是多时间序列，数据由 3 个期货时间序列构成。

一、单时间序列

按第五章第二节的算法，我们对期货 A 的时间序列数据进行贝叶斯网络模型训练后得到如下因果关系图，整理后如下：

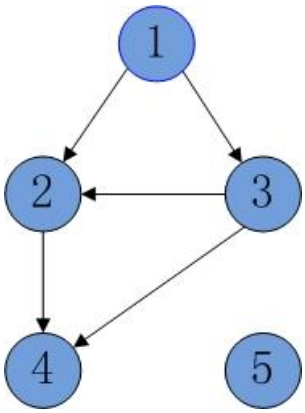

图 5.5 二阶贝叶斯网络因果关系图

从上图我们可以看到总共有 5 个节点，每个节点本身代表着在一阶贝叶斯网络中的一种因果关系。在一阶贝叶斯网络模型中，我们寻找的是节点之间的因果关系，对模型进行适当的简化后变成了寻找节 5 和其他节点之间的因果关系。一阶贝叶斯网络图反映的是节点 1，2，3，4 是否为节点 5 的父节点，当进入二阶贝叶斯网络数据集构建时，我们按照节点是否为节点 5 的父节点来进行转化：‘1’代表了该节点是节点 5 的父节点；‘2’代表了该节点不是节点 5 的父节点，所以，在二阶贝叶斯网络图中的节点表示的含义是：该节点和节点 5 之间的因果关系。

图 5.5 中节点 1 表示在原始时间序列中第一个节点和第五个节点之间的因果关系，我们用因果关系 1 来表示。同理，用因果关系 2 和 3 分别表示图 5.5 中的节点 2 和 3。从上图可以看到节点 1 是节点 2 和 3 的父节点，这说明因果关系 1 是因果关系 2 和 3 的父节点。由此可以得出以下结论：如果已知原始时间序列中节点 1 和节点 5 之间存在因果关系，那么节点 2 和节点 3 与节点 5 之间也存在着因果关系。

其他节点之间的因果关系推理也如上所述，这里不再赘述。另外，可以看到图 5.5 中有一个孤立的节点 5，那是因为我们在建立一阶贝叶斯网络模型时规定找出时间序列中节点 5 的父节点，由于节点 5 不可能是它本身的父节点，所以该因果关系无意义。

二、多时间序列

在多时间序列的贝叶斯网络模型中我们选取了 3 个期货时间序列，分别标记为 A,B,C。对期货 A,B,C 分别进行一阶贝叶斯网络模型训练后得到 3 组结果，每

组含有 5000 个贝叶斯网络图，对这些贝叶斯网络图进行数据集转化后按专家意见进行排序，然后拼接。拼接方式见下图：

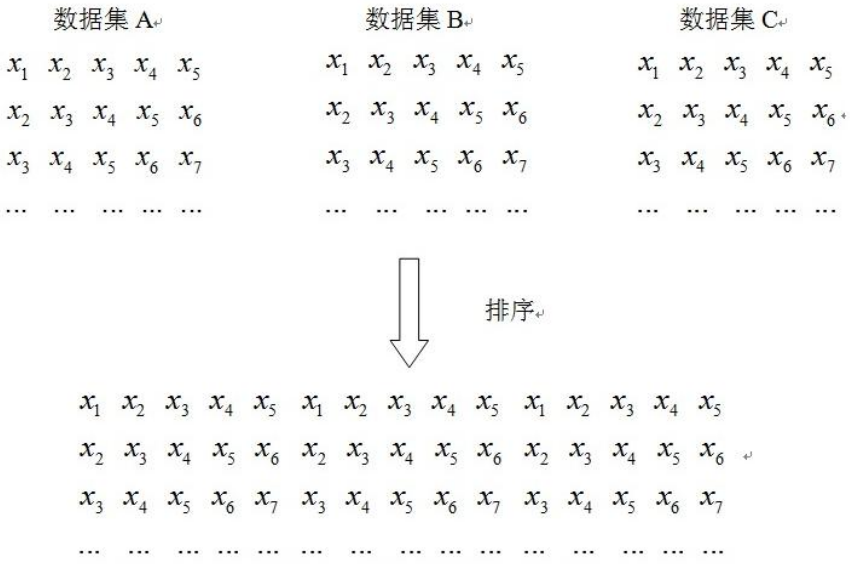

图 5.6 数据集拼接

根据上图进行数据集构建后，对多时间序列进行二阶贝叶斯网络模型训练，在训练的过程中设置了模型的深度，如果深度越大则模型复杂度越大。本文为了方便结果的解释，将模型深度设置成 1，最后得出的结果调整后如下图所示：

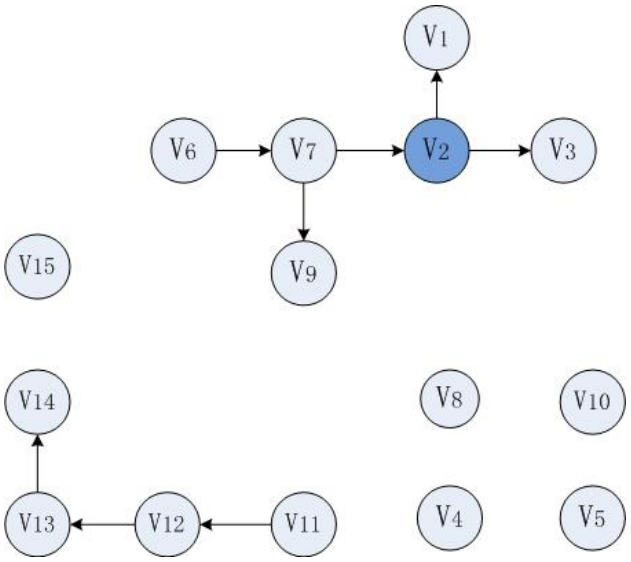

图 5.7 多时间序列二阶贝叶斯网络图

在图 5.7 中，节点 V1 到 V5 来自期货 A，V6 到 V10 来自期货 B，V11 到 V15 来自期货 C。从图中可以看到期货 C 和期货 A,B 之间并不存在任何因果关系，因此，期货 C 独立于其他两个期货。另外，由图中深色标示的节点 V2 可以看到该节点是 V1 和 V3 的父节点，同时它又是 V7 的子节点，这说明期货 A 和 B 之间存在着如下因果关系:如果已知原始期货 A 的时间序列中节点 2 是节点 5 的父节点，那么可以推断在原始期货 B 的时间序列中节点 2 也是节点 5 的父节点。

第四节　小结

本章主要讲述如何构建二阶贝叶斯网络模型，并对模型结果进行了解释。我们知道常规的贝叶斯网络研究节点与节点之间的因果关系，往往局限在一定的已知范围之内，无法挖掘多个已知范围之间的因果关系。而二阶贝叶斯网络研究的是因果关系之间的因果关系，能够通过一个范围之内的因果关系来推测另一个范围之内的因果关系。

在期货时间序列研究中，二阶贝叶斯网络可以用来研究多个期货之间的因果关系，分析期货数据之间一些隐藏的因果关系，因此，因果关系之间的因果关系研究有着重大的现实意义。

第六章 总结与展望

第一节 总结

随着大数据的快速发展，以概率统计为基础的机器学习在近年来受到工业界和学术界的极大关注，并在互联网、金融、自然语言、生物等领域获得很多重要的成功应用，其中贝叶斯网络在过去多年也得到了快速发展，成为非常重要的一类机器学习方法。贝叶斯网络是描述随机变量之间因果关系图的模型，其重要应用之一是随机变量之间的因果知识表示和推理。

目前因果关系研究主要使用格兰杰方法来研究因果关系，但是该方法存在众多的不足。另外常规的因果关系研究局限在节点与节点之间，无法对于跨范围的节点之间的因果关系进行研究。因此，针对这些缺点，本文提出了使用二阶贝叶斯网络模型和 MDL 打分算法来研究因果关系之间的因果关系。本文的主要工作和成果如下：

首先，介绍了大数据相关知识并简述了贝叶斯网络相关理论，分析对比了现有的因果关系挖掘模型和贝叶斯网络结构学习方法，明确研究方向，选择了基于 MDL 打分原理的贝叶斯网络模型作为本文的研究方法。

其次，提出了一种新型贝叶斯网络模型：二阶贝叶斯网络模型。设计出了新型模型构建的方法，并基于期货时间序列重构后的数据，实现了采用 MDL 打分算法的贝叶斯网络模型。

最后，利用二阶贝叶斯网络推理模型对期货时间序列进行仿真实验，实验不仅得到了单个期货时间序列内部节点之间的因果关系，而且得到了多个时间序列边与边之间的因果关系。

第二节 展望

基于二阶贝叶斯网络模型的期货时间序列因果关系挖掘比现存的多种模型都有着明显的优势，然而仍有需要改进的地方。从本文的实验过程来看，我们还需要在以下几个方面进行更加深入的研究:

（1）当网络节点的个数较多并且数据量较大时所需的存储空间比较大，进行贝叶斯网络学习比较困难，如何优化实现模型的算法以及保证模型有效性的前提下简化模型是进一步需要解决的问题。

(2)本文的研究是基于比较完备数据集的情况，因此数据预处理算法较为简单，但在实际生活中我们得到的时间序列可能会发生大量缺失，当数据大量缺失时必然导致贝叶斯网络模型的内部网络信息的丢失，从而影响到模型的准确性，因此，如何处理有大量缺失的期货时间序列也是需要进一步解决的问题。

（3）本文在进行二阶贝叶斯网络训练时得出的贝叶斯网络图过于复杂，如何简化因果关系需要进一步研究。

参考文献

[1]. Lee P M. Bayesian statistics: An introduction [M]. New York: Jone Wiley&Sons, 2012, 115-119.

[2]. Gelman A, Carlin J, Stern H et al. Bayesian data analysis [M]. Boca Raton:CRC Press, 2013, 67-74.

[3]. Tsamardinos I, Brown L E, Aliferis C F. The max-min hill-climbing Bayesian network structure learning algorithm[J]. Machine Learning, 2010, 65(1): 31-79.

[4]. 王德禄, 李尚, 王智勇等. 大数据:现状与展望[J]. 经营与管理, 2015, 5: 213-216.

[5]. 张兰廷.大数据的社会价值与战略选择[D]. 中共中央党校, 2014.

[6]. Markowetz F, Grossmann S, Spang R. Probabilistic soft interventions in conditional Gaussian networks[R]. In 10th AI/Stats, 2014,89-97.

[7]. Yin J, Zhou Y, Wang C. Partial orientation and local structural learning of causal networks for prediction[C]. Hong Kong:JMLR W&CP, WCCI2008 workshop on causality, 2014,93-104.

[8]. Eaton D, Murphy K. Exact Bayesian structure learning from uncertaininterventions[R]. In: AI & Statistics, 2011,107–114.

[9]. Li GL, Leong TY. Active learning for causal Bayesian network structurewith non-symmetrical entropy[C]. In: Proceedings of 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009,290–301.

[10].Korb K, Hope L, Nicholson A. Varieties of causalintervention[C]. In Pacific Rim Conference on AI, 2013.

[11].Li G J, Cheng X Q. Research status and scientific thinking of big data[R]. Bulletin of Chinese Acadenmy of Sciences, 2012, 27(6): 647-657.

[12].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.

[13].王双成,林士敏,陆玉昌.贝叶斯网络结构学习分析[J].计算机科学,2000,27(10): 77-79.

[14].Li B H, Gao J H. A note on minimal d-separationtrees for structural learning[J], Artificial Intelligence, 2010,174,442-448.

[15].Friedman N, Murphy K, Russell S. Learning the structure of dynamic probabilistic networks[C]. 14th Conf.on Uncertainty in Artificial Intelligence, 1998.

[16].王双成．贝叶斯网络学习、推理与应用[M]．立信会计出版社, 2009,9-11.

[17].Binder J, Koller D, Russell S．Adaptive probabilistic net-works with hidden variables[J]．Machine Learning, 1997,29(2):213-244.

[18].Ram R, Chetty M, Dix T. Causal modeling of gene regulatory network[C]. In: The 6th IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. Toronto, 2006,1-8.

[19].Yavari F, Towhidkhah F. Gene regulatory network modeling using Bayesian networks and cross correlation[C]. In Biomedical Engineering Conference. Cairo, 2008,1-4.

[20]. Pearl J. Causality: models, reasoning and inference[M]. Cambridge Univ, 2013.

[21].Cooper, G.F, Yoo. Causal discovery from a mixture of experimental andobservational data[C]. In: Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 2012,116-125.

[22].Li GL, Leong TY. Active learning for causal Bayesian network structurewith non-symmetrical entropy[C]. In: Proceedings of 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining , 2009,290–301.
[23].Tong S, Koller D. Active learning for structure in Bayesian networks[J]. In Intl. Joint Conf. on AI, 2014.
[24].Madigan D, York J. Bayesian graphical models for discrete data. Intl[J].Statistical Review, 2015,215–232.
[25].Markowetz F, Grossmann S, Spang R. Probabilistic soft interventions in conditional Gaussian networks[J]. In 10th AI/Stats, 2005.
[26].Yin J, Zhou Y, Wang C. Partial orientation and local structural learning of causal networks for prediction[C]. In: JMLR W&CP, WCCI2008 workshop on causality, 2008,93-104.
[27].Korb K, Hope L, Nicholson A. Varieties of causal intervention[J]. In Pacific Rim Conference on AI, 2014.
[28].光大证券股份有限公司研究所. 计算机行业-大数据 Big Data)专题报告[R].上海, 2011.
[29].The McKinscy Global Institute.Big data: The next frontier for innovation,competition,and productivity[R].USA, 2011.
[30].Pearl J. Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo, CA, 1988.
[31]. Lauritzen. Graphical models, clarendon press[J]. Oxford, 2006,197-235.
[32].Cowell G, David A P, Lauritzen S L. Probabilistic networks and expert systems[J], Springer Publications, New York, 2013,465-481.
[33].Melanc G, Bousquet M. Random generation of dags for graph drawing[C]. Technical Report INS-R0005, Centre for Mathematics and Computer Sciences, Amsterdam, 2012.
[34].Robinson R W. Counting labeled acyclic digraphs[R]. New Directions in the Theory of Graphs. Academic Press, 1999.
[35].Madigan D, York J. Bayesian graphical models for discrete data[J]. Intl.Statistical Review, 2005.
[36].Friedman N, Koller D. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks[J]. Machine Learning, 2013, 50(1): 95-126.
[37].Koivisto M. Advances in exact Bayesian structure discovery in Bayesian networks[J]. In UAI, 2006.
[38].Binghui liu,Jianhua Guo. A note on minimal d-separation trees for structural learning[C].Artifical Intelligence, 2010.
[39].朱明敏，刘蔚，杨有龙.基于全条件独立的贝叶斯网络 MPD-JT 构造算法[J].系统工程与电子技术. 2010, 32(6): 8-11.
[40].Ruiz C. Illustration of the K2 algorithm for Learning Bayes net structures computer science department[J].WPI, 2009.
[41].王越，谭署秋，刘亚辉.基于互信息的贝叶斯网络结构学习算法[J].计算机工程, 2011,37(7): 62-64.
[42].王飞，刘大有，王淞听. 基于遗传算法的贝叶斯网结构增量学习的研究[J]. 计算机研究与发展, 2005, 42 (9): 1461-1466.
[43].陶雪娇. 大数据研究综述[R]. 第八届全国仿真器学术年会, 2013.
[44].Paul C, Chris Eaton, Dirk R. Understanding big data [M]. USA: The McGraw-Hill

Companies, 2012,231-289.
[45].徐子沛. 大数据[M]. 广西: 师范出版社, 2012,57-64.
[46].符健. 解读大数据[Z]. 证券研究报告, 2011.
[47].涂兰敬. 大数据与海量数据的区别[J]. 网络与信息, 2011, 25(12): 37-38.
[48].李国杰. 大数据研究的科学价值[J]. 中国计算机学会通信, 2012, 8(9): 8-15.
[49].员巧云, 程刚. 近年来我国数据挖掘研究综述[J]. 情报学报, 2005,5:27-34.
[50].甘晓, 李国杰. 大数据成为信息科技新关注点[J]. 中国科学报, 2012, 7: 43-46.
[51].人大经济论坛, 传统分析与大数据分析的对比[DB/OL]. 2012.
[52].王珊, 王会举, 覃雄派等. 架构大数据: 挑战、现状与展望[J]. 计算机学报, 2011, 37(10): 1472-1484.

致谢

三年光阴一晃而过，回顾三年的研究生生活，有许多的快乐，也有许多的烦恼。在这里要对给我提供帮助的每一位老师和每一位同学说声感谢。

本文是在导师王双成教授的悉心指导下完成的。王老师一丝不苟的工作态度，为人师表的优秀品质将对我今后的工作与生活产生巨大影响。从课题的选题研究到论文的撰写过程，都离不开恩师的悉心教诲，值此论文完成之际，谨向王老师表示崇高的敬意和衷心的感谢！

感谢曾志勇导师，他是我的学业导师，引领我进入数据挖掘的大门。

感谢张旭洁老师，感谢她在学习和生活等各方面给予我的巨大关心和帮助！

感谢我的父母，是你们多年来一直坚定的支持我，鼓励我，没有你们我将无法实现自己的梦想！

最后，由衷的感谢在百忙之中审阅我论文以及参加答辩的各位老师们。

研究生生涯的结束只是另一段生涯的开始，以后的路还很长，愿与每一位帮助过我的人共勉。

在学期间研究成果和已发表的论文

硕士期间论文发表情况：

[1].姚衡，王双成，高瑞. 基于贝叶斯网络分类器的财务信息失真识别研究. 新会计，2015, 6:37-40.

[2].姚衡，王双成. 基于贝叶斯网络的企业风险因果关系发现. 新会计, 2016, 3:14-18.

硕士期间参与科研项目情况：

[1].上海市自然科学基金(15ZR1429700)，时间序列数据挖掘的贝叶斯网络方法研究。