

**概率论与数理统计课程论文**

题目：使**用贝叶斯网络改良数据挖掘算法**

|  |  |
| --- | --- |
| **班 号** |  |
| **学 号** |  |
| **姓 名** |  |
| **日 期** | **2022.11.3** |
| **成 绩** |  |

**摘 要**

在当下，互联网逐渐变的越来越发达。人类能够获取的数据量以指数的速度暴涨，信息查找技术亟待进一步的提升。高效的处理数据也成为了世界共同关注的难题。在概率论与数理统计和人工智能等学科知识的交叉融合下，数据挖掘技术诞生了。这一技术的主要目的是，从大量的数据中提取出有效的信息，它具有广泛的开发前景。

贝叶斯网络是概率论和图论知识结合的产物，在提供不确定性的环境下，通过结构学习和参数学习等手段，它可以完成决策、诊断、分类等任务。因其良好的可理解性和逻辑性成为了数据挖掘的主要手段。本文以数据挖掘技术作为主要研究对象，研究贝叶斯网络能够在数据挖掘领域得到的应用。通过三阶段结构学习的原理实现贝叶斯网络的学习。

本文以贝叶斯网络在数据挖掘领域如何应用为切入点，重点讨论了贝叶斯网络与数据挖掘的结合，贝叶斯网络的算法等问题。本文的主要成果及结论为：分析了贝叶斯网络与数据挖掘的关系；从贝叶斯网络的定义出发介绍了三阶段学习。

关键词：贝叶斯网络；数据挖掘；结构学习

**1. 了解数据挖掘**

**1.1 数据挖掘的产生和发展**

随着网络技术的发展，社会的信息化程度快速增长，人们生产数据的能力大幅度提高，然而信息处理技术却没有跟上脚步。如今，数据规模已经达到万亿字节的水平。如何从庞大的数据中寻找到潜在有用的信息是人们迫切需要解决的问题。数据挖掘技术就是在这样的背景下应运而生的[1]。

数据挖掘最终要达成的目的是，从海量的、人力无法完成的检索任务中，找到目标数据。它是一门综合性的技术领域，主要涉及数据库、人工智能和数理统计等三个技术领域。

数据挖掘技术多应用于统计，数据分析和信息管理系统等领域。近年来数据挖掘技术在人工智能和机器学习领域也得到了广泛的应用。

**1.2 数据挖掘的内容和任务**

数据挖掘有两个主要的目标：预测和描述。预测是基于使用历史数据对数据类型和值进行预测。描述是对数据间的关系进行识别，提供一种探索被分析数据性质的方法。

数据挖掘主要具备以下八项任务：

（1）分类

（2）回归

（3）时间序列分析

（4）预测

（5）聚类

（6）汇总

（7）关联规则

（8）序列发现

**1.3 数据挖掘使用的技术**

数据挖掘任务有很多实现方法，这些实现方法需要特定的算法，下面介绍一些常用的数据挖掘技术：

（1）数学的统计方法：利用数理统计原理建立数据的统计模型，然后从模型中提取感兴趣的信息。这些统计原理包括：点估计、基于汇总的模型、贝叶斯定理、假设检验、回归和相关。本文所讨论的Bayes网络就属于此法。

（2）神经网络方法：神经网络方法模拟了人脑神经元结构，以MP结构和Hebb学习规则为基础建立多种神经网络模型：前馈式网络、反馈式网络和自组织网络。前馈式网络可用于预测、模式识别等方面；反馈式网络可用于联想记忆和优化计算；自组织网络则用于聚类。神经网络像决策树一样，也需要建立代表模型的图结构，然后将这个结构应用于数据。

（3）遗传算法：遗传算法是模拟生物进化过程的算法。通过种群遗传产生优良的后代，这些后代需要满足适应值，经过若干代的遗传，将得到满足要求的后代(问题的解)。

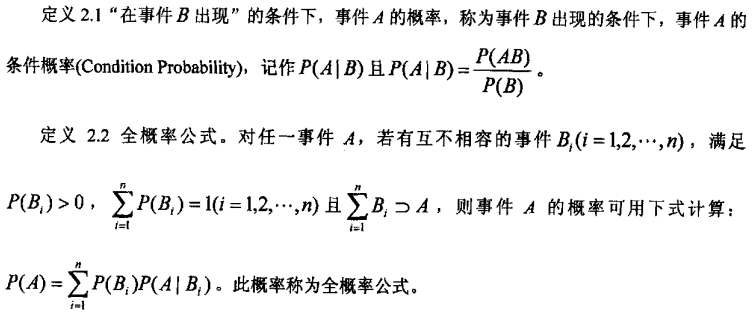
本文将重点介绍使用数学统计的方法进行数据挖掘。

**2. 数据挖掘和贝叶斯网络**

**2.1 贝叶斯网络的基本观点**

贝叶斯统计分析起源于英国数学家贝叶斯（R.T.Bayes)撰写的一篇论文：“An Essay Towards Solving a Problem in the Doctrine of Chances(机遇理论中一个问题的解) 。文中给出了著名的贝叶斯公式，后来的学者们在此基础上将其发展成一整个数理统计的体系，这就是贝叶斯统计[5]。

贝叶斯统计学已广泛应用于几乎所有的学科，并取得了显著的成果。



上图为贝叶斯网络用到的概率论公式

贝叶斯方法是以贝叶斯理论为基础，它的一个显著特点是可以通过分析结果来了解假设。也就是说，在先验条件模糊不清的情况下，贝叶斯方法具有其它方法无可替代的优点。而数据挖掘技术最重要的一个特点就是挖掘未知的知识，数据挖掘与传统的数据分析的本质区别之一是数据挖掘是在没有明确假设的前提下挖掘信息。0

总结贝叶斯网络的特点来看，它是一种非常适合数据挖掘的概率统计方法。

**2.2 贝叶斯网络用于数据挖掘的好处**

用贝叶斯网络的方法进行数据挖掘的好处如下：

（1） 贝叶斯网络描述了变量之间的因果联系。

（2）能挖掘出知识的隐含性。

（3）贝叶斯网络具有良好的可理解性和逻辑性。

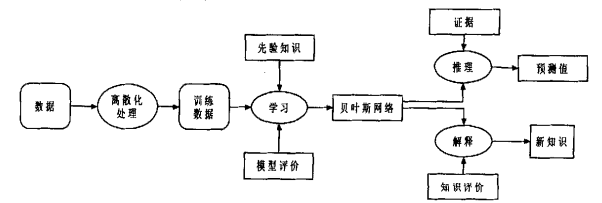
（4）贝叶斯网络结合先验知识，并且用图模型描述了数据间的相互关系，便于进行预测分析。

**2.3 贝叶斯网络用于数据挖掘的不足**

（1）贝叶斯方法最大问题是先验信息的使用。先验信息来源于经验，众所周知在科研任务中没有经过实验验证的数据是不可靠的，因此在很多方面都有争议。

（2）处理数据复杂性高，因此时间和空间消耗也比较大。贝叶斯方法要进行后验概率的计算、区间估计、假设检验等，大量的计算是不可避免的。

**2.4 将贝叶斯网络应用于数据挖掘**

****

上图为基于贝叶斯网络的数据挖掘框架

贝叶斯网络不但可以表示不确定知识，而且可以进行概率推理，贝叶斯网络的学习算法还能从大量的数据中自动构造贝叶斯网络，这使得贝叶斯网络非常适合于不确定性知识发现。

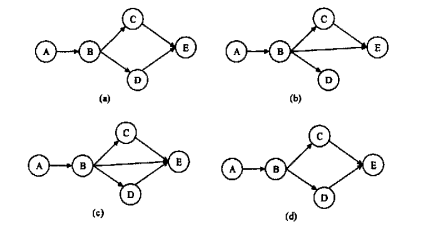
**2.5 贝叶斯网络的结构学习**

**2.5.1贝叶斯网络与结构学习算法**

讨论贝叶斯结构学习，必须从贝叶斯定义出发。贝叶斯网络，本质上就是一个建立局部条件概率分布的有向概率图模型。为了得到这个有向无环图，我们需要用到结构学习算法。结构学习的定义是：学习出最优网络结构，也就是各节点之间的依赖关系。传统的结构学习方法分为两类：基于评分搜索的方法和基于依赖分析的方法。经过比较，本文决定主要以结合参数学习的.方式进行贝叶斯学习。这种将获取参数的过程分成三部分的方法称为“三阶段学习”。

**2.5.2 给定节点次序的贝叶斯网络学习算法**

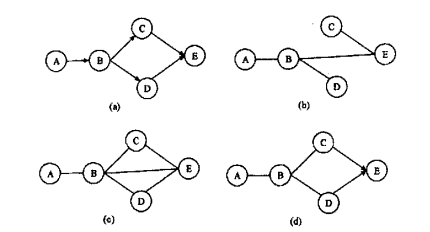
三阶段算法的所谓的三个阶段分别是：起草(drafting)，增厚过程(thickening)和减薄过程 (thinning)。第一阶段要计算每对节点相互间的信息，基于这些信息建立网络草图；第二阶段，在 成对的节点在某种条件下不是条件独立时，在草图上添加边。假设隐藏的模型是DAG的P-Map，第二阶段的结果得到概率模型的I-map；第三阶段，对I-map的每条边进行条件独立性的测试，如果两节点是条件独立的，则移除它们之间的边，第三阶段的结果是隐藏模型的P-Map[6]。



贝叶斯网络的学习过程

**2.5.3 未定节点次序的贝叶斯网络学习算法**

事先未定节点次序的贝叶斯网络学习算法是对给定节点次序的算法的改进，这种算法降低了对节点顺序的要求，同样也是基于三个阶段的算法。算法也是以训练数据集作为输入，构建的贝叶斯网络作为输出。因为事先并不知道节点的顺序，所以算法必须处理两个问题：如何确定两个节点是否条件独立和如何确定连接两个节点边的方向[6]。



贝叶斯网络的学习过程

**3. 总结**

作为数据挖掘重要手段之一的贝叶斯网络．是研究的热点。它不仅有着坚实的概率论理论基础，同时又能够很好地得到学者的逻辑方面的表达，将数据间因果关系很好的进行表达，并且能定量的给出数据间的依赖程度，挖掘出数据集中的有效信息。随着对贝叶斯网络学习算法研究的不断深入，算法的原理越来越简容易理解和应用，计算复杂度也会越来越低。贝叶斯网络的应用前景非常广泛，会在更多的领域显示出强大的特性，挖掘出庞大数据中隐藏的重要信息，为人们的工作和生活提供帮助。

**参考文献**

[1] 王春梅. 基于神经网络的数据挖掘算法研究[J]. 现代电子技术,2017,40(11):111-114. DOI:10.16652/j.issn.1004-373x.2017.11.029.

[2] 李刚. 知识发现的图模型方法．【博士学位论文】[D]，上海大学．2001，

[3] 陈文伟，黄金才．数据仓库与数据挖掘[M]．人民邮电出版社2004．

[4] Fayyad，U．M．，et al。From Data Mining Knowledge Discovery：An overview．In：Advance in Knowledge Discovery and Data Mining【J】．AAAIT Press，1996，1-34．

[5] 毛国君．段立娟．王实. 数据挖掘原理与算法，清华大学出版杜【M】2005．

[6] 沈海峰．基于贝叶斯网络数据挖掘技术理论及算法的研究．【硕士学位论文】[D]．合肥工业大学．2002