朴素贝叶斯分类模型的研究与应用

摘　要

朴素贝叶斯分类器是一种简单而高效的分类器，基于朴素贝叶斯技术的分类是当前数据挖掘领域的一个研究热点。本文以突破朴素贝叶斯分类模型属性间独立性假设限制为研究内容，从两个方面对朴素贝叶斯分类模型进行了深入的研究，并将朴素贝叶斯分类模型应用于指导学生选择专业方向。

本文主要工作如下：

(1) 基于属性相关性分析是改进朴素贝叶斯分类模型的结构。通过分析属性相关性度量和进行属性约简，得到满意的属性约简子集。在此基础上提出一种基于属性相关性度量的朴素贝叶斯分类模型 EANBC。实验结果表明，与朴素贝叶斯分类模型相比，EANBC 分类模型具有较高的分类正确率。

(2) 基于强属性限定是对朴素贝叶斯分类模型的结构进行了扩展。通过分析贝叶斯定理的变形公式和属性相关性度量，提出一种基于强属性限定的贝叶斯分类模型 SANBC。实验结果表明，与朴素贝叶斯分类模型相比，SANBC 分类模型具有较高的分类正确率。

(3) 将朴素贝叶斯分类模型应用于指导学生选择专业方向。通过建立专业方向选择的朴素贝叶斯分类模型，充分利用以往各届学生选择专业方向的先验知识，指导学生根据自己的专业知识结构以及专业知识的掌握程度科学合理地选择专业方向。

关键词： 贝叶斯定理　朴素贝叶斯　分类模型　属性相关性　属性约简

Research and Application of Naive Bayesian Classification Mode

Abstract

Naïve Bayes classifier is a simple and effective classification method. Classifying based on Bayes Technology has got more and more attentions in the field of data mining. In order to get rid of the limit of the assumption of independence among attributes of Naive Bayesian Classifier, this thesis makes a study of two Bayesian classifying model, meanwhile, Naive Bayesian Classification Model is applied to help students in selecting specialties direction. The major work of this thesis is described as follow.

Based on the evaluation of condition attribute with correlation improves structure of Naive Bayesian Classification Model. On the basis of analyzing the evaluation of condition attribute with correlation and attribute reduction, satisfied attributes reduction set has been given. According to this method, EANBC is proposed. Compared with Naive Bayesian Classification Model, experimental results show EANBC has higher accuracy.

Restricted Bayesian Classification Model Based on Strong Attributes extends the structure of Naive Bayesian Classifier. On the basis of analyzing a variant of Bayes theorem and the evaluation of condition attribute with correlation, SANBC is proposed. Compared with Bayesian Classification Model, experimental results show SANBC has higher accuracy.

The Bayesian Classification Model is designed to help the students with their selection of appropriate specialties. By constructing the Bayesian Classification Model and using the experience gained by the students in the past in their selection of specialties, students can base their selection of appropriate specialties on their personal knowledge framework, mastery of knowledge in their fields.

Key words: Bayes theorem naive Bayes classification model attributes with correlation attributes reduction

插 图 清 单

表 格 清 单

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标志和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 合肥工业大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签字：王骏 签字日期：2006年6月15日

学位论文版权使用授权书

本学位论文作者完全了解 合肥工业大学 有关保留、使用学位论文的规定，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，允许论文被查阅或借阅。本人授权 合肥工业大学 可以将学位论文的全部或部分论文内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文者签名：王骏 导师签名：

签字日期：2006年6月15日 签字日期：06年6月15日

学位论文作者毕业后去向：
工作单位：淮南师范学院 电话：0554-6673049
通讯地址：淮南师范学院教务处 邮编：232001

致　谢

在论文完成之际，首先感谢合肥工业大学给我这个学习和提高的机会，特别是我的导师胡学钢教授对我的悉心指导，借此机会向他们表示衷心的感谢！

论文是在我的导师胡学钢教授的悉心指导下完成的。胡教授严谨的治学态度、渊博的专业知识、敏锐的学术洞察力将对我以后的工作、学习产生深远的影响。论文的字里行间浸透了胡教授的心血。胡教授在学术上带给我启迪，拓宽了我的思路，引导了我的学术思维。胡老师在学术上对我要求严格，生活中是我的良师益友。

我要感谢合肥工业大学计算机与信息学院的老师们给我的教诲、关心和帮助。

我还要感谢胡春玲同学给予我的帮助。

在整个学习阶段，我的工作单位给予了我很大的支持和鼓励，对此表示诚挚的谢意。

作者：王峻

2006 年 5 月 20 日

第一章 绪 论

随着人类迈入崭新的信息时代，信息资源与日俱增，尤其近年来随着计算机科学技术的迅猛发展、数据库和计算机网络的广泛应用，使数据库领域的新内容、新应用、新技术层出不穷，产生大量的数据信息。数据的丰富带来了对强有力的数据分析工具的需求，数据挖掘就是在这种背景下产生并成为当前人智能领域的研究热点。本章概述了数据挖掘的研究和发展概况，重点介绍了其中的分类问题，此外还给出了本文的内容组织。

1.1 数据挖掘概述

1.1.1 什么是数据挖掘

数据挖掘[1]（Data Mining），也称数据库中的知识发现（KDD[2]，Knowledge Discovery in Databases），是指从大型数据库或数据仓库中提取人们感兴趣的知识的过程，这些知识是隐含的、事先未知的、潜在有用的信息。提取的知识一般可表示为概念（Concepts）、规则(Rules)、 规律(Regulation)、模式(Patterns)等形式。用数据库系统来存储数据，用机器学习的方法分析数据，挖掘大量数据背后蕴含的知识，这两者的结合促成了数据挖掘技术的产生。数据挖掘作为一门交叉学科，涉及到机器学习、模式识别、归纳推理、统计学、智能数据库、数据可视化、专家系统、高性能计算等多个领域。

1.1.2 数据挖掘技术产生的背景

一、 数据挖掘技术的应用需求分析

数据挖掘之所以吸引专家学者的研究兴趣和广泛关注，主要是近年来随着计算机科学技术的迅猛发展、数据库和计算机网络的广泛应用以及将数据转换成有用知识的迫切需求。目前，各种新技术与数据库技术的有机结合，使数据库领域的新内容、新应用、新技术层出不穷，形成庞大的数据库家庭，产生海量的数据信息。然而传统的数据分析手段很难对数据进行深层次的处理并获取有价值的知识，导致了“数据爆炸但知识贫乏”现象的产生。因此人们希望能够智能、自动地对数据进行更高层次的分析，挖掘数据背后蕴含的许多重要的信息以便充分地利用这些数据。新的需求推动新技术的诞生，这就是数据挖掘技术产生的应用需求背景。

二、数据挖掘技术产生的技术背景分析

数据挖掘技术的提出和广泛的接受是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。归纳数据挖掘产生的技术背景，下面一些相关技术的发展起到了决定性的作用：

1、数据库、数据仓库和Internet等信息技术的发展；

2、计算机性能的提高和先进的体系结构的发展；

3、统计学和人工智能等方法在数据分析中的研究和应用。

计算机芯片技术的发展使计算机的处理和存储能力日益提高；计算机的体系结构，特别是并行处理技术的日趋成熟和普遍应用，已成为支持大型数据库处理应用的基础；计算机性能的提高和先进的体系结构的发展使数据挖掘技术的研究和应用成为可能。

历经几十年的发展，包括基于统计学、人工智能等在内的理论与技术成果已经被成功地应用到数据的处理和分析中，这些应用从某种意义上为数据挖掘技术的提出和发展起到了极大的推动作用。正是由于实际的需求和相关技术的发展，数据挖掘技术才逐步发展起来。目前数据挖掘技术已成为国际上数据库和信息决策领域的最前沿的研究方向之一。

1.2 数据挖掘的现实意义

数据挖掘技术无疑在科学研究方面具有重大意义[1]。在信息量极为庞大的天文、气象、生物技术等领域，大量的实验和观测数据靠传统的数据分析工具难以处理，借助数据挖掘技术分析这些海量数据，可以极大地提高科学家发现知识的效率。目前在这方面已获得一些重要的应用成果，例如，美国加州理工学院喷气推进实验室与天文学家合作开发的SKICAT系统通过对几百万个天体进行分类，帮助天文学家发现了16个新的类星体；专家系统DENDRAL根据质谱仪提供的数据，能够发现已知或未知的高分子化合物的分子结构；机器学习系统BACON根据已有实验和观测数据，重新发现欧姆定律、凯普勒定律，当然也可以从新的实验和观测数据中发现新的物理和天体定律。

数据挖掘技术在其它方面也具有同样重要的作用。例如，在金融投资方面，在进行投资决策之前，需要对各种投资方向的有关数据进行分析，以选择最佳的投资方向。数据挖掘可以通过对已有数据进行处理，利用学习得到的模式进行市场预测。在保险业方面，保险是一项风险业务，保险公司可以利用数据挖掘技术进行风险分析，在保险公司建立的保单及索赔信息库的基础上，寻找保单中风险较大的领域，从而得出一些实用的控制风险的策略，指导保险公司工作。在制造业方面，利用数据挖掘技术进行零件故障诊断、资源优化、生产过程分析等，通过对生产过程进行分析，发现容易产生质量问题的工序及相关故障因素。

1.3 数据挖掘技术研究现状和发展趋势

目前，对数据挖掘的研究主要体现在以下几个方面：一是对知识发现方法的研究进一步发展，如近年来注重对Bayes（贝叶斯）方法以及Boosting方法的研究；二是传统的统计学回归法在数据挖掘中的应用；三是数据挖掘技术与数据库的结合越来越紧密。在应用方面，数据挖掘商业软件工具不断产生和完善，

国外很多计算机公司非常重视数据挖掘系统的开发应用，比较典型的有SAS公司的Enterprise Miner，IBM公司的Intelligent miner, SGI公司的SetlIiner，SPSS公司的Clementine等。

数据挖掘的研究趋势体现在以下几个方面[3][4]：

(1)**挖掘方法和用户交互问题**:这反映在所挖掘的知识类型、在多粒度上挖掘知识的能力、领域知识的使用、特定的挖掘和知识显示；

(2)**性能问题**:包括数据挖掘算法的有效性、可伸缩性和并行处理以及分布式和增量挖掘算法；

(3)**挖掘中的可视化方法**:使得知识发现的过程能够被用户理解，也便于在知识发现过程中的人机交互；

(4)**加强对各种非结构化数据的挖掘**:如文本数据、图形图象数据、多媒体数据；

(5)**研究在网络环境下的数据挖掘技术**:特别是在Internet上建立DM Server与数据库服务器配合，实现数据挖掘。

1.4 数据挖掘的主要任务

数据挖掘的主要任务是从大量数据中提取出可信的、新颖的、有效的并能被人们理解的模式，挖掘数据背后蕴含的许多重要的信息以便充分地利用这些数据。数据挖掘的两个高层次目标是预测和描述。前者是根据一些变量或数据库的若干已知字段预测其他感兴趣的变量或字段的未知的或未来的值；后者是找到描述数据的可理解模式。根据发现的知识的不同，我们可以将数据挖掘任务归纳为以下几类[3][6][7]：

(1)**特征规则**：特征规则挖掘是把所有数据满足的概念特征化。特征规则挖掘能够总结并发现用户指定的数据集的一般特征，从而从与学习任务相关的一级数据中提取出关于这些数据的特征式，这些特征式表达了该数据集总体特征。例如可以从某种疾病的症状中提取关于该疾病的特征规则等。

(2)**序列模式**:是指在多个数据序列中发现共同的行为模式。序列模式发现算法的框架与发现关联规则相同。例如，对于某顾客，在序列数据库中，序列模式发现问题就是在该数据库中寻找所有的频繁序列或所有的最长频繁序列。R. Agranal称最长频繁序列为序列模式。

(3)**分类**:在数据挖掘的各种方法中，分类是一种主要的分析手段，旨在生成一个分类函数或分类模型，由该模型把数据库中的数据项映射到某一给定类别中，从而实现对数据的分类。分类问题被广泛应用于疾病诊断、银行信贷等领域。目前研究的分类模型主要有决策树、 贝叶斯分类、神经网络、粗糙集、统计方法、遗传算法等。本文的研究是基于朴素贝叶斯分类模型的改进。

(4)**关联规则**：关联规则挖掘是数据挖掘研究的一个重要的、高度活跃的领域。关联性用来发现一组项目之间的关联关系和相关关系，它们经常被表达为

规则形式。发现关联规则的任务就是从数据库中发现那些确信（Confidence）和支持度(Support)都大于给定值的强规则。近几年对关联规则研究颇多，研究工作已经从单一概念层次关联规则的发现发展到多个概念层的关联规则的发现，随着概念层次的不断深入，使发现的关联规则能提供更具体的信息。关联性分析广泛应用于交易数据分析，通过分析结果来知道销售、目录设计及其他市场决策的制定。例如，在分析美国加州某连锁店的销售记录时，发现下班以后购买婴儿尿布的男性顾客往往同时也会购买啤酒。关联性问题是数据挖掘中比较成熟的问题。

(5)**聚类**：聚类是一种常见的描述性工作，搜索并识别一个有限的种类集合或簇集合，从而描述数据。简单地说，就是识别出一组聚类规则，将数据分成若干类。聚类的目的是使属于同一类别的个体之间的距离尽可能小而不同类别的个体间的距离尽可能的大，也就是说，聚类使类内个体间的相似性最大，而类别间的相似性最小。

(6)**预测**：预测是构造和使用模型评估无标号样本类，或评估给定样本类可能具有的属性值或值区间。预测的两类问题是分类和回归，其中分类是预测离散或标称值，而回归是用于预测连续或有序值。预测主要是根据已知数据项和预测模型，预测该数据项特定属性的值，预测也包含基于可用数据的分布趋势识别，连续性的预测可以用回归统计技术建模。

(7)**变化和偏差分析**：变化和偏差分析是探测数据现状、历史记录或标准之间的显著变化和偏离，偏差包括很大一类潜在的有趣知识。如观测结果与期望的偏离、分类中的反常实例、模式的例外等。

1.5 数据挖掘中的分类问题

分类[7]是数据挖掘和机器学习中的一个重要研究课题。它旨在生成一个分类函数或分类模型，对由属性集描述的实例指定最适合的类标签，从而实现对数据的分类。数据分类一般分为两个步骤：

第一步：建立分类模型，描述预定的数据类集或概念集。通过分析有属性描述的数据库元组来构造模型。通常分类器用分类规则、判定树或数学公式的形式提供。常用的分类器模型有决策树、决策表、贝叶斯方法、神经网络、遗传算法等。

第二步：使用建立的分类模型对新的数据集进行划分，主要考虑分类规则的准确性、矛盾划分的取舍等。一个好的分类规则集合应该是对新的数据集而言具有很高的准确性、尽可能少的矛盾划分和较少的规则集。

1.5.1 常用的几种分类模型

一、决策树

决策树是常用的分类模型之一，它利用树中从根到叶子节点的路经表示分

类规则，其中每个内部节点表示在一个属性上的测试，每个分枝代表一个属性上的测试，每个树节点代表类或类分布，树的最顶层节点是根节点。决策树方法的优点是可理解性较强，比较直观，缺点是处理复杂数据时，受噪音数据等因素的影响而导致出现过多碎片。ID3算法是较早也是最著名的决策树算法，在ID3算法的基础上，又演化出ID3增强版及C4.5、CART、CHAID，后期的改进算法有QUEST和PUBLIC，目前推出一些可伸缩性的决策树算法，如SLIQ、SPRINT、“雨林”和BOAT算法等。

二、贝叶斯方法

贝叶斯分类模型是一种典型的基于统计方法的分类模型。贝叶斯定理是贝叶斯理论中最重要的一个公式，是贝叶斯学习方法的理论基础，它将事件的先验概率与后验概率巧妙地联系起来，充分利用先验信息和样本数据信息确定事件的后验概率。贝叶斯分类器分为两种：

一种是朴素贝叶斯分类器，它是贝叶斯分类模型中一种最简单、有效的而且在实际使用中很成功的分类器，其性能可以与神经网络、决策树相媲美[6]。朴素贝叶斯分类模型基于假定特征向量的各分量间相对于决策变量是相对独立的，即条件独立性假设。尽管这一假定在一定程度上限制了朴素贝叶斯分类模型的适用范围，但在实际应用中，降低了贝叶斯网络构建的复杂性。朴素贝叶斯分类模型已成功地应用到聚类、分类等数据挖掘的任务中。当然朴素贝叶斯分类模型仍有需要改进的地方，条件独立性假设在一定程度上限制了朴素贝叶斯分类模型的使用范围，因此人们开始研究放松独立性条件的限制，以提高朴素贝叶斯分类器的分类性能。为了突破朴素贝叶斯分类器的独立性假设条件的限制，人们通过改变其结构假设的方式来达到目的。例如半朴素贝叶斯分类器SNBC(Semi-Naive Bayesian Classifier)、树扩张型贝叶斯分类器TAN(Tree—Augmented Bayesian Classifier)及增强型贝叶斯分类器BAN(Bayesian Network Augmented Naive Bayes)等。

另一种是贝叶斯网络分类器，贝叶斯网络又称为信念网络，它是基于后验概念的贝叶斯定理。贝叶斯网络是一个有向无环图，其中结点代表论域中的变量，有向弧代表变量的关系，变量之间的关系强弱由结点与其父结点之间的条件概率来表示，通过贝叶斯网络可以准确地反映实际应用中变量之间的依赖关系。贝叶斯网络可用于分类、聚类、人工神经网络、预测和因果关系分析等。贝叶斯网络分类器具有很强的学习、推理能力，能很好地利用先验知识。

三、神经网络

神经网络[7][8]是大量的简单神经元按一定规则连接构成的网络系统。它通过模拟人类大脑的结构和功能，采用某种学习算法从训练样本中学习，并将获取的知识存储在网络各单元之间的连接权中。神经网络的分类过程分为训练和分类两个阶段。在训练阶段，首先定义网络的拓朴结构，再对训练样本中的每

个属性的值进行规范化预处理，然后用神经网络对已预处理的输入进行学习，训练完毕后，用训练好的神经网络对标识样本进行分类。

目前神经网络模型很多，反向传播模型（BP模型）是最典型的神经网络。反向传播算法是在多层前馈神经网络上学习，在学习时，权值向量任意给出，反向传播通过迭代处理一组训练样本，将每个样本的网络预测与实际已知道的类标号进行比较、学习，通过修改权，使网络预测与实际类之间的均方差最小。神经网络的优点是抗干扰能力强，对未经训练的数据有较好的预测分类能力。神经网络的缺点是用加权链连结单元的网络表示的知识很难被人们理解。

四、遗传算法

遗传算法[10]是模拟生物进化过程的全局优化方法，通过繁殖、交叉、变异，在求解空间按一定的随机规则迭代搜索，直到求得问题的最优解。根据适者生存的原则，形成当前群体中最适合的规则组成新的群体，以及这些规则的后代。规则的适合度是用它对训练样本集的分类准确度来进行评估。后代通过使用交叉和变异等遗传操作来创建。在交叉操作中，来自规则对的子串交换，形成新的规则对；在变异操作中，规则串随机选择的位被反转，由先前的规则群体产生新的规则群体的过程继续，直到群体P进化，P中的每个规则满足预先指定的适合度阈值。遗传算法易于并行，并且已用于分类和其他问题的优化，但遗传算法过于复杂。

五、粗糙集合

粗糙集理论[8][11]可以用于分类，发现不准确数据或噪声数据内在的结构联系，它用于离散值属性。粗糙集理论基于给定训练数据内部的等价类的建立，形成等价类的所有数据样本是不加区分的。粗糙集理论是用元素的成员关系函数、概念的上近似和下近似等来刻划知识处理的方法。由不可区分关系确定给定论域的等价类，使用粗糙集合相应的公式计算条件属性和决策属性的依赖性，通过数据约简在保持分类一致的约束下简化样本数据，从而削减冗余对象和属性，寻求属性提取最小子集以确保产生满意的近似分类，由此得出知识的相对约简和相对核以及种类的相对约简和相对核等目标数据，通过对目标数据的分析，使用很少的逻辑规则就能描述分类规则。

六、关联规则

关联规则挖掘是数据挖掘研究的一个重要的、高度活跃的领域。关联性用来发现一组项目之间的关联关系和相关关系，它们经常被表达为规则形式。关联规则[7]是KDD研究中的一个重要的研究课题，关联规则广泛地应用于各个领域。目前关联规则的挖掘已经取得了令人瞩目的成果，根据研究方向的不同可分为:多循环方式的挖掘算法、并行挖掘算法、增量式更新算法、基于约束条件的关联规则挖掘和挖掘多值属性的关联规则。近年来，数据挖掘已将关联规则

挖掘用于分类，并已取得很好的效果。

ARCS是基于聚类挖掘关联规则，然后使用规则进行分类；CBA是关联分类，它将分类规则挖掘与关联规则结合在一起；CAEP使用项集支持度挖掘显露模式，EP用于构造分类；基于多维关联规则的分类算法CMAR是利用FP－Growth算法挖掘关联规则，建立类关联分布树FP树。

1.5.2 分类模型的评价

分类模型可以从以下几个方面进行评价[11]：

(1)**预测准确度**：预测准确度是评价分类模型的最广泛的一种比较尺度，用于评价一个分类模型对于预测将来数据的准确度。常见的两种方法是保持和K次N交叉验证法。

(2)**计算复杂度**：计算复杂度依赖于具体的实现细节和硬件环境。在数据挖掘中，由于操作对象通常是大型的数据库，因此空间和时间的复杂度问题将是非常重要的一个环节。

(3)**模型描述的简洁度和可解释性**：对于描述型的分类任务，模型描述越简洁，越具可解释性，越受欢迎。

(4)**健壮性**：是分类模型抗干扰能力的度量，这涉及对于数据集中噪声数据或空缺数据的处理，它反应在有噪声数据或空缺数据的情况下模型是否有正确分类的能力。

(5)**可伸缩性**：大部分的分类算法通常假定数据量很小，算法的可伸缩性意味着对于大量数据能否具有有效的构造模型的能力。

1.6 本文的内容组织

本文由六章组成：

第一章主要介绍了数据挖掘及其研究对象的发展现状和未来趋势，详细阐述了数据挖掘中各种分类问题的定义、方法以及分类模型评价的标准等。最后简要给出了文章的组织结构。

第二章系统介绍了贝叶斯分类的基本理论，详细阐述了几种常用的贝叶斯分类模型：朴素贝叶斯分类模型、贝叶斯网络模型及增量贝叶斯分类模型等，并分析了各种分类模型的优缺点。

第三章是本文的核心内容。本章介绍了属性相关性的度量方法，并运用此方法对条件属性进行约简，从而改善了条件属性间的依赖程度，弱化独立性条件假设，并在此基础上将属性约简与朴素贝叶斯分类两种计算方法相结合，提出了一种基于属性相关性度量的朴素贝叶斯分类模型EANBC，选择UCI机器学习数据库提供的典型数据库实例，通过实验对EANBC算法和NBC算法进行了比较，实验表明，EANBC算法分类的准确率优于NBC算法。

第四章是本文的核心内容。通过分析贝叶斯定理的变形公式和属性相关性

度量，介绍了强属性的选择方法，并提出一种基于强属性限定的贝叶斯分类模型SANBC。基于强属性限定的贝叶斯分类模型是对朴素贝叶斯分类模型的结构进行了扩展，其目的是为了突破朴素贝叶斯分类模型特征属性间独立性假设限制，提高分类性能。文中给出构造SANBC的算法，选择UCI机器学习数据库提供的典型数据库实例，通过实验对SANBC算法和NBC算法进行了比较，实验表明，SANBC算法分类的正确率优于NBC算法。

第五章介绍朴素贝叶斯分类模型在指导学生选择专业方向中的应用，通过建立专业方向选择的朴素贝叶分类模型，指导学生根据自己的专业知识结构合理地选择专业方向，为学生的学习起一个科学导向作用。

第六章对已做的工作进行总结，并对下一步的工作进行了展望。

第二章　贝叶斯理论和贝叶斯分类模型

2.1 引言

分类[12][13]是根据数据的不同特征将其划分为不同的类别。在数据挖掘中，构造分类模型可以使用许多不同的方法，如决策树、贝叶斯分类法、神经网络分类法等。通过对分类算法的比较研究发现，贝叶斯分类算法可以与决策树算法和神经网络算法相媲美[8]。对于大型数据库，朴素贝叶斯分类法也已表现出高准确率与高速度。贝叶斯分类是统计学分类方法，是建立在经典的贝叶斯概率理论基础上的基于统计方法的分类模型，本章主要介绍贝叶斯基本理论和贝叶斯分类模型。

2.2 数理统计基础理论

2.2.1 条件概率和乘法定理

在事件 A 已经发生的条件下，事件 B 发生的概率，称为事件 B 在给定事件 A 的条件概率（也称为后验概率），记作 $P(B|A)$。相应地，$P(A)$ 称为无条件概率（也称为先验概率）。条件概率可以由下式进行计算：

$$P(B \mid A) = \frac{P(A \cdot B)}{P(A)}$$

由条件概率可求得概率的乘法定理：

$P(A \cdot B) = P(B|A)P(A)$

对于 n 个事件 $A_1, A_2, \ldots A_n, n \geq 2$，则有：

$P(A_1, A_2, \ldots, A_n) = P(A_n|A_1 \cdot A_2 \cdots A_{n-1})P(A_{n-1}|A_1 \cdot A_2 \cdots A_{n-2}) \cdots P(A_2|A_1)P(A_1)$

2.2.2 全概率公式和贝叶斯定理

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

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots + P(A|B_n)P(B_n)$$

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

上式称为全概率公式。

设试验 E 的样本空间为 S，A 为 E 的事件，$B_1, B_2, \ldots B_n$ 为 S 的一个划分，且 $P(A) \succ 0, P(B_i) > 0 (i = 1,2,\ldots n)$，则由条件概率的定义和全概率公式：

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

上式称作贝叶斯定理[14]。

2.2.3 极大后验假设与极大似然假设

贝叶斯定理提供了一种计算假设概率的方法，它基于假设的先验概率、给定假设下观察到不同数据的概率以及观察到的数据本身。

我们用P(h)表示没有训练数据前假设h拥用的初始概率。P(h)被称为h的先验概率（prior probability），表示所拥有的关于h是一正确假设的概率的背景知识。用P(D)表示将要观察的训练数据D的先验概率（在没有确定某一假设成立时D的概率）。P(D|h)表示假设h成立的情况下数据D的概率，则由贝叶斯公式得出计算后验概率P(h|D)的方法：

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

在许多学习任务中，需要考虑候选假设集合H并在其中寻找给定的数据D时可能性最大的假设h∈H。任何这样具有最大可能性的假设被称作极大后验假设[11]（maximum a posteriori,MAP），记作h_{MAP}：

$$h_{MAP}=\arg\max(P(h|D)=\arg\max(P(D)P(h)/P(D))$$

$$=\arg\max P(D|h)P(h)$$

由于P(D)是不依赖于h的常量，所以在最后一步去掉P(D)，上式就是一个原始的分类模型。贝叶斯分类就是根据上述MAP假设找出新实例最可能的分类。所有对贝叶斯分类模型的研究工作都是以此为前提的。

在某些情况下，可假设H中每个假设有相同的先验概率（即对H中任意的hi和hj，P(hi)= P(hj)），此时可进一步简化，只考虑P(D|h)来寻找极大可能假设。P(D|h)常被称作极大似然假设（maximum likelihood,ML）记为h_{ML}，

$$h_{ML}=\arg\max P(D|h)$$

在分类过程中，上式常被用来在启发式搜索时进行模型检测。

2.2.4 事件的独立性

设 A,B 是试验 E 的两个事件，一般 A 的发生对 B 发生的概率是有影响的，这时 $P(B|A)\neq P(B)$，只有在这种影响存在时才会有 $P(B|A)=P(B)$，这时有：

$$P(A\cdot B)=P(B|A)P(A)=P(A)P(B)$$

则称 A,B 为相互独立事件。

同样，对于 n 个事件 $A_1,A_2,...A_n$，如果有

$$P(A_1\cdot A_2\cdots A_n)=P(A_1)\cdot P(A_2)\cdots P(A_n)$$

则称 $A_1,A_2,...,A_n$ 为相互独立事件。

2.3 贝叶斯分类模型

分类有规则分类（查询）和非规则分类（有指导学习）。贝叶斯分类是非规则

分类，它通过训练集（已分类的例子集）训练而归纳出分类器（被预测变量是离散的称为分类，连续的称为回归），并利用分类器对没有分类的数据进行分类。贝叶斯分类器中有代表性的分类器有朴素贝叶斯分类器 、贝叶斯网络分类器和树扩展的朴素贝叶斯分类模型TAN分类器等。贝叶斯分类具有如下特点：

（1）贝叶斯分类并不把一个对象绝对地指派给某一类，而是通过计算得出属于某一类的概率，具有最大概率的类便是该对象所属的类；

（2）一般情况下在贝叶斯分类中所有的属性都潜在地起作用，即并不是一个或几个属性决定分类，而是所有的属性都参与分类；

（3）贝叶斯分类对象的属性可以是离散的、连续的，也可以是混合的。

根据给定的训练集归纳出分类器是数据挖掘的一项重要而基本的任务，在众多的分类器中（决策树、决策表、神经网络和粗糙集分类器等），朴素贝叶斯分类器以简单的结构和良好的性能受到人们的关注，在理论上它在满足其限定条件下是最优的，针对其较强的限定条件，可以尝试着减弱独立条件以扩大最优范围，从而产生更好的分类器。

2.3.1 朴素贝叶斯分类模型

朴素贝叶斯分类器（Naive Bayes Classifier,NBC）是贝叶斯分类模型中一种最简单、有效的而且在实际使用中很成功的分类器[15][16][17]，其性能可以与神经网络、决策树相媲美，甚至在某些场合优于其它分类器。

朴素贝叶斯分类模型描述如图2-1所示，设有变量集$U=\{A_1,A_2,...A_n,C\}$，其中$A_1,A_2,...A_n$是实例的属性变量，C是取m个值的类变量。假设所有的属性都条件独立于类变量C，即每一个属性变量都以类变量作为唯一的父节点，就得到朴素贝叶斯分类模型。

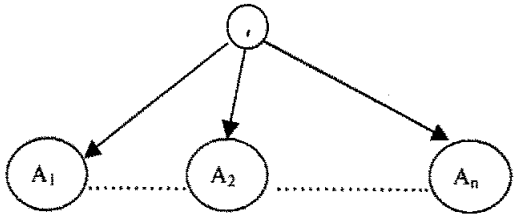

图 2-1 朴素贝叶斯分类模型结构示意图

朴素贝叶斯分类模型假定特征向量的各分量间相对于决策变量是相对独立的，也就是说各个变量独立地作用于决策变量，尽管这一假定在一定程度上限制了朴素贝叶斯分类模型的适用范围，但在实际应用中，大大降低了贝叶斯网络构建的复杂性。朴素贝叶斯分类模型已成功地应用到聚类、分类等数据挖掘的任务中。

一、朴素贝叶斯分类的工作过程

（1）每个数据样本用一个n维特征向量$X=\{x_1,x_2,...,x_n\}$表示，分别描述对n个属性$A_1,A_2,...,A_n$样本的n个度量。

(2)假定有m个类$C_1,C_2,...C_m$。给定一个未知的数据样本X(即没有类标号)，分类法将预测X属于具有最高后验概率(条件X下)的类。朴素贝叶斯分类将未知的样木分配给类Ci，当且仅当:

$$P(C_i|X)>P(C_j|X),1\le i,j\le m,j\ne i$$

这样，最大化$P(C_i|X)$，$P(C_i|X)$最大的类C_i称为最大后验假定，根据贝叶斯定理

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

(3)由于$P(X)$对于所有类为常数，只需要$P(X|C_i)P(C_i)$最大即可。如果类的先验概率未知，则通常假设这些类是等概率的，即$P(C_1)=P(C_2)=...=P(C_m)$，并据此对$P(C_i|X)$最大化。类的先验概率可以用$P(C_i)=s_i/s$计算，其中s_i是类C_i中的训练样本数，而s是训练样本总数。

(4)给定具有许多属性的数据集，计算$P(X|C_i)$的开销可能非常大。为降低计算$P(X|C_i)$的开销，可以做类条件独立的朴素假定。对于给定样本的类标号，假定属性值相互条件独立，即在属性间不存在依赖关系。这样，

$$P(X|C_i)=\prod_{k=1}^{n}p(x_k|c_i)$$

概率$P(x_1|C_i),P(x_2|C_i),...P(x_n|C_i)$可以由训练样本估计，其中

(a)如果A_k是分离散属性，则$P(x_k|C_i)=s_{ik}/s_i$，其中s_{ik}是在属性A_k上具有值x_k的类C_i的训练样本数，而s_i是C_i中的训练样本数。

(b)如果A_k是连续值属性，则通常假定该属性服从高斯分布。因而，

$$P(x_k|C_i)=g(x_k,u_{ci},\sigma_{ci})=\frac{1}{\sqrt{2\pi}\sigma_{ci}}e^{\frac{(x_k-u_{ci})^2}{2\sigma_{ci}^2}}$$

其中给定类C_i的训练样本属性A_k的值，$g(x_k,u_{ci},\sigma_{ci})$是属性$A_k$的高斯密度函数，而$u_{ci},\sigma_{ci}$分别为平均值和标准差。

(5)为对未知样本X分类，对每个类C_i，计算$P(X|C_i)P(C_i)$。样本X被指派到类C_i，当且仅当

$$P(X|C_i)P(C_i)>P(X|C_j)P(C_j),1\le i,j\le m,j\ne i$$

换言之，X被指派到其$P(X|C_i)P(C_i)$最大的类C_i。

朴素贝叶斯分类模型的优点是:

(1)算法逻辑简单，易于实现;

(2)算法实施的时间、空间开销小;

(3)算法性能稳定，对于不同特点的数据其分类性能差别不大，即模型的健壮性比较好。

二、朴素贝叶斯分类模型的改进方法[18]

朴素贝叶斯分类器是基于一个简单的假定：在给定分类特征条件下属性值之间是相互条件独立的。在现实世界中，它的属性独立性假设使其无法表示实际应用中各属性之间的依赖关系，影响了它的分类性能。因此需要针对实际应用对朴素贝叶斯分类器模型进行改进，使之在属性独立性假设不满足的情况下依然具有较高的分类精确度。

通过分析，朴素贝叶斯分类器的本质是一种具有很强限制条件的贝叶斯网络分类器，但是它限制条件太强，不适于现实应用。然而完全无限制条件的贝叶斯网络也是不现实的，因为学习这样的网络非常耗时，其时间复杂度为属性变量的指数级，并且空间复杂度也非常高。因此，研究朴素贝叶斯分类器的改进模型，只能从这两者之间来考察，即研究具有较宽松条件限制的贝叶斯网络分类器。

(1)**属性删除技术**：适用于存在冗余属性的情况。Langley和Sage提出了一种基于属性删除的选择性贝叶斯分类器。当存在一些属性依赖于其他属性，特别是存在冗余属性时，属性删除方法确实能够改善朴素贝叶斯分类器的预测精确度。

(2)**构造新属性或概率调整技术**：适用于某些属性依赖于其他属性时。Pazzani等提出了通过相互依赖的属性构造一个新属性，并用新属性取代原来相互依赖的那些属性的方法。这种方法也可以视作为事先的条件概率调整技术。Wang和Webb等提出了一种准懒惰式(Semi - Lazy)的限制性贝叶斯网络分类器的条件概率调整方法，在某些情况下可以减小误分类率。

(3)**局部朴素贝叶斯分类器**：适用于属性之间相互依赖的情形比较复杂的情况。这种方法是为属性变量的每一种取值(或某个范围)建立一个朴素贝叶斯分类器，也就是说，单一的全局朴素贝叶斯分类器被许多局部朴素贝叶斯分类器所代替，将属性独立性假设放宽到只要局部属性独立就可以了。

Kohavi将朴素贝叶斯分类器和决策树相结合，用一棵决策树来分割实例空间，在每个叶子结点上建立局部朴素贝叶斯分类器。

Zheng和Webb等利用懒惰式学习策略提出了一种懒惰式贝叶斯规则(Lazy Bayesian Rule)学习技术，该方法将懒惰式技术应用到局部朴素贝叶斯规则的归纳中。该算法虽然较大地提高了分类精确度，但是效率很低。

为了提高LBR的效率，Wang和Webb给出了一种启发式LBR算法和HLBR算法，可以有效地提高学习效率。LBR和HLBR是目前该方向上的最新研究成果之一。

(4)**树扩张型贝叶斯方法**：Friedman等提出了一种树扩张型贝叶斯方法。这种方法的基本思路是放宽朴素贝叶斯的独立性假设条件，扩展朴素贝叶斯的结构，使其能够容纳属性间存在具有某种特征的依赖关系。Friedman利用条件相互信息(Conditional MutualInformation)建立属性之间的依赖关系矩阵，构造一棵最大权生成树作为一个分类器。由于限制每个属性结点最多有一个非类变量

(类标)的父结点，也就是说每个属性结点最多仅依赖于一个非类标结点，使其表示依赖关系的能力受到限制。

(5)限定性双层贝叶斯分类模型DLBAN[19]：石洪波等提出了一种限定性的双层贝叶斯分类模型DLBAN ，这种方法的出发点是通过对属性空间的搜索，找出一些对其他属性有较强影响的属性，那么所有其他的属性仅通过与这些属性的关联就可以将重要的依赖关系表示出来。

2.3.2 半朴素贝叶斯分类模型

为了突破朴素贝叶斯分类器的独立性假设条件的限制，可以通过改变其结构假设的方式来达到目的，为此有人提出了半朴素贝叶斯分类[4] SNBC(Semi Naive Bayesian Classifier)的构想。从名称可以看出，SNBC依然属于朴素贝叶斯分类的范畴。SNBC的结构比NBC紧凑，在SNBC的模型构建过程中，依照一定的标准将关联程度较大的基本属性(即NBC中的特征属性)合并在一起构成“组合属性”(也称之为“大属性”)。逻辑上，SNBC中的组合属性与NBC中的基本属性没有根本性差别，SNBC的各个组合属性之间也是相对于类别属性相互独立的。图2-2是SNBC分类模型结构示意图：

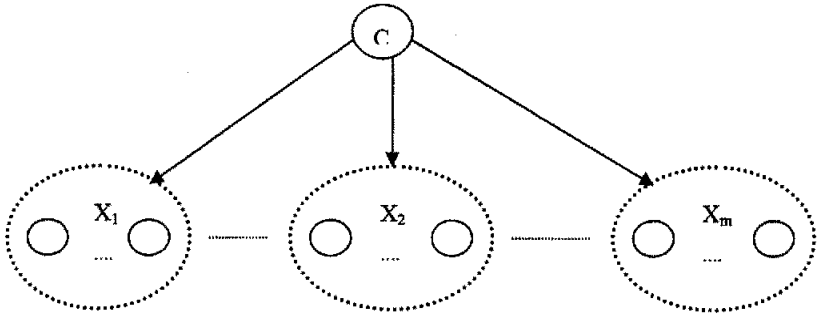

图 2-2 SNBC 分类模型结构示意图：

2.3.3 提升的朴素贝叶斯分类模型

对朴素贝叶斯分类模型进行“提升”(Boosting)[20][21]是在不改变独立性假设的前提下提高分类性能的一种方法。提升的方法是由Freund和Schapire于1995年提出，其主要思想是从训练实例中学习一系列的分类器。每一个分类器根据前一个分类器错误分类的实例，对训练例的权重进行修正，再学习新的分类器。例如，学习得到分类器Hk后，增加了由Hk导致分类错误的训练实例的权值，并且通过重新对训练实例计算权值，再学习下一个分类器Hk＋1。这个过程重复T次，从这个系列的分类器中可以综合得出最终的分类器。

Freund和Scbapire给出的AdaBoost算法实现了提升算法对分类问题的处理，具体算法如下[20]：

Input:N个训练实例：D=$\langle(x^1,c^1),...,(x^N,c^N)\rangle$以及待分类实例优易包联十
N个训练实例上的分布D：w,w为训练实例的权向量。

T：训练重复的趟数

Output: $h(x)=\arg\max\sum_{t=1}^{T}\left(\log\frac{1}{\beta^{(t)}}\right)I(h^{(t)}(x)=c)$,

其中$I(\omega)$是示意函数，当$\omega=T$时$I(\omega)=1$，否则$I(\omega)=0$

步骤：

初始化训练实例的权向量，$W_i=1/N,i\in(1...N)$;

for t=1 to T

给定权值$W_i^{(t)}$得到一个假设$H^{(t)}:X\to C$

估计假设$H^{(t)}$的总体误差，$e^{(t)}=\sum_{i=1}^{N}w_i^{(t)}I(c^i\neq h^{(t)}(x^i))$

计算$\beta^{(t)}=e^{(t)}/(1-e^{(t)})$

计算下一轮样本的权值$w_i^{(t+1)}=w_i^{(t)}(\beta^{(t)})^{1-I(C^i=h^{(t)}(x^i))}$

正规化$w_i^{(t+1)}$，使其总和为1

End for

假设每一个分类器都是实际有用的，$e^{(t)}<0.5$，也就是说，在每一次分类的结果中，正确分类的样本个数始终大于错误分类的样本个数。可以看出，此时$\beta^{(t)}<1$，那么当对某个训练实例x^i分类结果不正确时，示意函数$I(\omega)=0$，导致$w_i^{(t+1)}$增加，因此满足了提升的思想。

上述提升朴素贝叶斯分类器的时间复杂度是0(Tnf)，其中f是每个样本的属性的个数。在一般情况下，提升后的分类性能有了较大的提高，但是这种提升方法也存在以下的不足：

一是不能捕捉属性间的相关性，也就是说没有突破条件独立性假设的限制；二是当训练集中存在噪音数据时，提升方法会把噪音数据当成有用的信息通过权值而放大，从而降低提升的性能。

2.3.4 树扩展的朴素贝叶斯分类模型

NBC需要一个很强的属性之间条件独立性的假设，而这些假设在许多问题中并不成立，如果忽视这一点儿会引起分类的误差。由于在NBC中，所有的属性结点都属于类结点的马尔可夫覆盖，所以可以对NBC进行增强，保留其结构特点，放松它的独立性假设，使属性变量之间存在一定的简单依赖关系，这类分类器称为树扩展的贝叶斯分类器TAN[22](Tree Augmented Naive Bayesian Classifier)分类器。

TAN贝叶斯网络要求属性结点除类结点为父结点外最多只能有一个属性父结点，其中$\{X,,\{X_1,X_2,...,X_n\}\}$构成一棵树，TAN贝叶斯网络被这棵树所唯一确

定，而树可被函数 $\pi:\{1,...,n\}\mapsto\{0,1,...,n\}$（使 $\pi(i)=0$ 的结点为父结点，不存在序列 $i_1,...,i_k$ 使得 $\pi(i_j)=i_{j+1}$（其中 $i\le j<k$ 而且 $\pi(i_k)=i_1$）所确定，当 $\pi(i)>0$ 时，$\prod_{x_i}=\{X_{\pi(i)}\}$；当 $\pi(i)=0$ 时，$\prod_{x_i}=\phi$，因此，函数 π 就定义了TAN贝叶斯网络。给定属性结点之间的条件互信息函数

$$Ip(X,Y|Z)=\sum_{X,Y,Z}p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}$$

具有最大似然的TAN贝叶斯网络构造过程如下：

(1)通过训练集计算属性对之间的条件互信息 $I_{pp}^{\wedge}(A_i,A_j|C)$。

(2)建立一个以 $I_{pp}^{\wedge}(A_i,A_j|C)$ 为弧的权重的加权完全无向图。

(3)建立一个最大权重跨度树。

(4)选择一个根结点，设置所有边的方向是由根结点向外，把无向树转换为有向树。

(5)增加一个类变量结点及类变量结点与属性结点之间的弧。

建立最大权重跨度树的方法是：首先把边按权重由大到小排序，之后遵照被选择的边不能构成回路的原则，按照边的权重由大到小的顺序选择边，这样由所选择的边构成的树便是最大权重跨度树。

贝叶斯网络分类器TAN是对朴素贝叶斯网络分类器进行有效改进的分类器，它既有朴素贝叶斯分类器的简单性，又有比朴素贝叶斯分类器更好的分类性能，当然，还可以对贝叶斯网络分类器TAN进行有效的改进。

2.3.5 贝叶斯网络分类模型

贝叶斯网络[11][23][24][25]是一个有向无环图，其中结点代表论域中的变量，有向弧代表变量的关系，变量之间的关系强弱由结点与其父结点之间的条件概率来表示。通过贝叶斯网络可以准确地反映实际应用中变量之间的依赖关系。当把网络中代表类别变量的节点作为根节点，其余变量作为它的子节点时，贝叶斯网络就成为了贝叶斯网络分类器。

贝叶斯网络又称为信念网络，是一种图型化的模型，能够图形化地表示一组变量间的联合概率分布函数。一个贝叶斯网络包括了一个结构模型和与之相关的一组条件概率分布函数。结构模型是一个有向无环图，其中的节点表示了随机变量，是对于过程、事件、状态等实体的某一特性的描述；边表示变量间的概率依赖关系，图中的每个节点都有一个给定其父节点情况下该节点的条件概率分布函数。这样一个贝叶斯网络就用图形化的形式表示了如何将与一系列节点相关的条件概率函数组合成为一个整体的联合概率分布函数。

一、增强型朴素贝叶斯分类模型 BAN

增强朴素分类模型的方式是使属性间关系不只局限于树形，而是任意的贝叶斯网络，这样的分类器称作BAN (Bayesian Network Augmented Naive Bayes)。

BAN结构进一步扩展了TAN的结构，允许属性之间形成任意的有向图（BAN分类模型结构示意图如图2-3所示）。构建一个BAN 分类器分为三步：

(1)针对属性变量,按贝叶斯网络的建立方法,构建一个贝叶斯网络；

(2)加入分类变量,将它作为所有属性变量节点的父节点；

(3)学习分类器的条件概率分布。

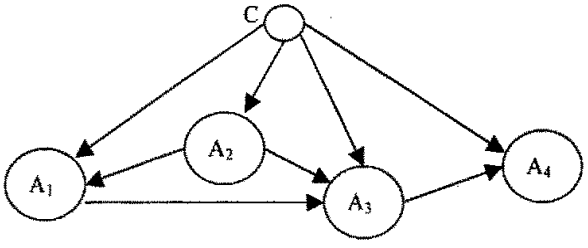

图2-3 BAN分类模型结构示意图

二、通用贝叶斯网络分类器GBN

该分类器是将类节点和属性节点作为同等地位的网络节点,根据数据集中的数据训练出贝叶斯网络,直接作为分类器的。这种结构学习需要获得一个完整的贝叶斯网络，而分类问题可以当作一种特殊的推理过程或决策问题。贝叶斯网络的结构学习可以分为两种形式:一是找出最适合数据集的网络结构,称为模型选择(Model Selection);二是选出一组网络结构,代表所有的网络,称为模型平均(ModelAveraging)，一般采用前者作为最终网络的获取方法。模型选择可分为两类：一种是基于评分函数(ScoringFunction)的学习，另一种是基于独立性测试(Conditional Independence Test,CIT)的学习。

基于评分函数的学习是用一个预定义好的函数作为评分标准,对模型结构空间中所有的模型进行评分,选出分值最大者所对应的网络结构作为最终网络结构。常用的评分函数有基于贝叶斯统计的BDE、最小描述长度MDL 和贝叶斯信息标准BIC。当然在巨大的模型空间中作盲目搜索,要得到最终的网络结构是相当困难的,故而常采用贪心策略、模拟退火、最优最先等算法进行搜索。

基于独立性测试的学习是通过度量属性之间的独立性关系,确定贝叶斯网络结构的方法。建立通用贝叶斯网络分类器一般采用此方法 。

用贝叶斯网络分类器进行分类的过程,实际上就是将属性节点作为证据节点引入到贝叶斯网络中,求得类节点后验概率的过程。当后验概率最大时,类别节点相应的取值即作为分类的结果。

在贝叶斯网络中,把某节点的父节点、子节点及子节点的父节点称为该节点的马尔可夫覆盖。根据有向马尔可夫性质,某节点取值的概率只受其马尔可夫覆盖节点的影响,而与其余节点无关。这样一些可能对分类有重要意义的变量,由于被归于马尔可夫覆盖,而影响不到分类节点。

三、贝叶斯多网分类器 BMN

贝叶斯多网分类器实际上是BAN分类器的扩展，它是由多个子贝叶斯网络分类器组成。对于类别变量所有取值，BAN分类器使属性变量之间保持相同的关系，而BMN分类器属性变量之间的关系却随类变量取值的不同而不同。一个简单的贝叶斯多网分类器BMN分类模型结构示意图如图2-4所示。

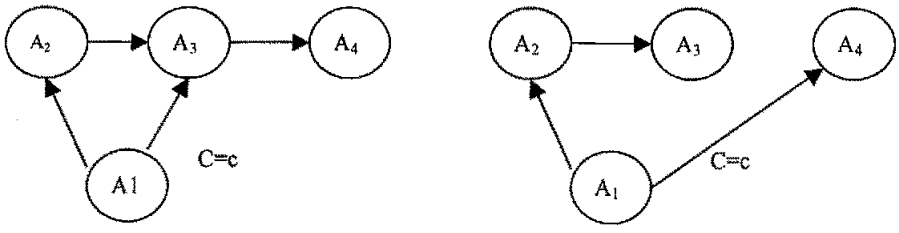

图 2-4　BMN 分类模型结构示意图

从结构上看，BMN结构更简洁，因为BMN中的每个子网内的局部比BAN简单。

建立一个BMN分类器的步骤如下：

(1)将训练数据集根据类变量取值的不同作相应的划分；

(2)对每个划分好的数据集，用三阶段法进行学习；

(3)学习每个局部网络的参数。

三阶段法分为Drafting、Thickening、Thinning 三个步骤。

对于给定节点顺序的训练集，具体算法如下：

(1)初始化空集S、R，计算每对属性变量的互信息 $I(v_i, v_j)$ ，对大于设定值 ε 的互信息进行排序，并按互信息由大到小的顺序，将对应的属性变量对放入集合S中；

(2)按S中的第一组变量添加弧，弧的指向由节点顺序决定，在S中删除该变量对；

(3)继续从S中取出排序变量对，在不给定任何节点取值的条件下，如果该变量对之间是d分割的，则添加相应的弧，否则将其添加到有序集合R 中，作为末尾节点对，在S中删除该变量对；

(4)重复(3)，直到S为空集；

(5)按顺序取出集合R中的节点对，找出阻塞该对节点每条开放路径(open path)的最少数量的节点，组成阻塞集B；

(6)计算已知B中各节点取值的条件下该对节点的条件互信息，如果条件互信息大于 ε，则添加弧，并在R 中删除该变量对；

(7)重复(5)，直到R 集合为空；

(8)对于每对已添加弧的节点，如果它们之间还存在除该弧之外的其他开放路径，则暂时删除该弧，寻找阻塞集，计算在阻塞集节点取值已知条件下的互信息。根据条件互信息判断该对节点是否相互独立，若是，则永久删除该弧，否则恢复该弧。

在以上计算中，(1)～(4)是Drafting阶段，(6)～(7)是Thickening阶段，(8)是Thinning阶段。

其中，互信息和条件互信息的计算公式为：

$$I(x_i, x_j) = \sum_{xi,xj} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

$$I(x_i, x_j \mid c) = \sum_{xi,xj,c} P(xi, xj \mid c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)}$$

2.3.6 增量贝叶斯分类模型

增量分类[26]是一种动态分类过程。其特点是随着分类过程的推进，训练集的规模不断扩大，分类过后的实例逐一被纳入训练集合中，也就是说，分类器的参数随着新实例的加入而不断更新。这种分类模型的关键是在有众多候选实例的情况下，如何选择新的实例优先加入训练集，从而促进分类性能的提高，使得当所有的待分类实例都被分类结束后总体分类性能最高。

优先实例的选择有两种处理策略[27]：被动选择策略与主动选择策略。

在增量分类过程中，被动策略随机地选择新的实例加入训练集，然后在新的训练集的基础上更新分类器参数，尽管这种策略算法简单，但是它存在明显的不足：

一是顺序地选择新实例往往会使学习的分类器具有顺序相关性，对数据过分敏感；

二是遇到噪音样本时，会使这种噪音一直传播下去，影响分类精度；

三是缺乏综合未带类别标注样本信息的能力。由于未带类别标注的样本往往包含有助于分类的信息，在这种情况下，选择好的未带类别标注的样本，把它加入到当前的分类器中，是非常重要的。

主动选择策略对新样本的选择是主动的，它首先选择最有利于分类器性能的样本来训练分类器，属于更高层次的，具有潜意识的学习，主动选择优先实例增量分类过程如图 2-5 所示。

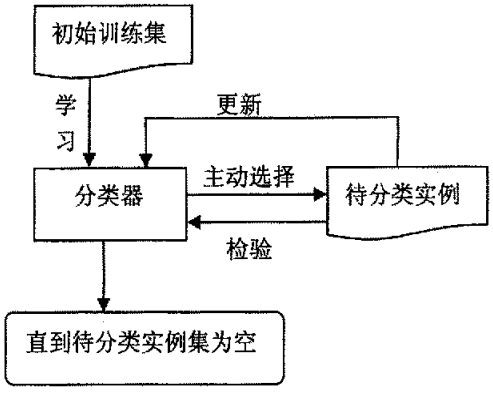

图 2-5　主动选择优先实例增量分类过程

2.4 本章小结

本章对贝叶斯分类技术的原理和方法作了系统的介绍：首先介绍贝叶斯分类的原理，即贝叶斯定理，贝叶斯定理是贝叶斯分类技术的理论基础，所有贝叶斯分类模型均是基于贝叶斯定理的。然后讨论了几种常见的贝叶斯分类模型，介绍了各种模型的分类思想，分析了它们各自的优缺点。

第三章 基于属性相关性分析的朴素贝叶斯分类模型

本章主要针对朴素贝叶斯分类模型的改进方法进行进一步的讨论，通过分析属性相关性度量和属性约简，提出一种基于属性相关性度量的朴素贝叶斯分类模型 EANBC, 并给出构造 EANBC 的算法和相应的实验结果。

3.1 问题的提出

朴素贝叶斯分类模型是基于各条件属性之间相对于类变量相对条件独立的假设，即各条件属性独立作用于类变量。根据朴素贝叶斯分类方法的条件独立性假设，只有在各属性变量与类变量相对独立，或各属性变量的相关性小到可以忽略的程度，朴素贝叶斯分类器才能得到最优的分类效果。尽管朴素贝叶斯的条件独立性假设看起来是合理的，而且对简化贝叶斯方法也是有效的，但是这个限制过于严格，在实际的应用中通常难以满足，各属性变量之间常常具有明显的依赖性。大量的研究表明，条件独立性假设在一定程度上限制了朴素贝叶斯分类模型的使用范围，所以人们开始研究放松独立性条件的限制。例如 Friedman,J.H 等人提出了 TAN 分类模型，就是利用在类变量和特征属性之间建立树形结构的方式放松了朴素贝叶斯分类方法的相对独立假设限制。本章以突破严格的独立性假设限制为出发点，通过分析属性相关性度量和属性约简，找出一组最近似独立的属性约简子集，这样不仅可以改善或弱化属性间的依赖关系，而且可以降低特征维数，减小求解问题的复杂度，提高朴素贝叶斯分类器的性能。

3.2 基于 x^2 统计的属性相关性度量

对于两个基本属性 A、B，分别有值 $a_i(i=1,2,...,m), b_j(j=1,2,...,n)$，其频数的列表如表 3-1 所示。

表 3-1 两个属性 A、B 的频度列表

	b_1	$b2$	…	b_n	SUM
a_1	f_{11}	f_{12}	…	f_{1n}	$A_1=\sum f_{1j}$
a_2	f_{21}	f_{22}	…	f_{2n}	$A_2=\sum f_{2j}$
…	…	…	…	…	…
a_m	f_{m1}	f_{m2}		f_{mn}	$A_m=\sum f_{mj}$
SUM	$B_1=\sum f_{i1}$	$B2=\sum f_{i2}$		$Bn=\sum f_{in}$	

为检验行列变量的相关性[28][29]，使用x^2统计量

$$x^2=\sum_{i,j}^{m,n}\frac{(f_{ij}-A_iB_j/f)^2}{A_iB_j/f}$$

其中f_{ij}表示ai,bj同时出现的频度，A_i表示a_i出现的频度，B_j表示b_j出现的频度，f为样本容量。

由x^2统计量，可以得到$m*n$列表数据中行列变量属性相关性的度量：

$$\psi=\begin{cases}\dfrac{f_{11}f_{22}-f_{12}f_{21}}{\sqrt{A_1A_2B_1B_2}} & \text{m=n=2}\\ \sqrt{x^2/f} & \text{其 它}\end{cases}$$

根据有关统计学理论，x^2提供有无关联性的证据，而ψ反映出关联性的强弱，ψ的绝对值越大，属性相关性越强，其绝对值接近于零时属性相关性较弱。该方法同样适用于各属性变量与类变量之间的相关性的度量。

3.3 基于属性相关性的属性约简方法

为了提高分类的正确性、有效性和可伸缩性，需要对分类所用的数据进行必要的预处理，属性约简[30]作为一个重要的数据预处理手段在数据挖掘的过程中得以频繁使用。属性约简的目的是在现有的属性集合中找出能保留信息系统所有分类信息的最优属性约简子集。通常情况下，属性约简对目标子集有两个基本要求：一是属性集的各个条件属性与决策属性的相关性越大越好，以保证属性集中没有无关属性；二是属性集的各个条件属性之间的相关性越小越好，以保证条件属性集中没有冗余属性。

针对第一个要求，现有的很多属性选择方法是以属性相关度为启发式信息，对属性集进行搜索来得到目标子集的，但是这些方法往往都忽略了属性选择的第二个要求，致使目标子集中的冗余信息无法过滤。

对于第二个要求，理论上可以用$\psi(A,B)$来度量各条件属性之间的相关性大小，据此过滤冗余属性，但是相对于考察各个属性跟决策属性之间的相关性，逐对考察各条件属性之间的相关性的计算量是相当大的。

通常情况下若属性 A 是某一属性强相关，在一定程度上属性 A 与其他属性之间也存在较强的相关性，因此可以将一个属性与决策属性的相关性作为衡量该属性与其他属性之间相关性的一个参考。此外，对于一对相关属性的取舍，可以参考对目标子集的第一个要求，将这两个属性与决策属性的相关性作为一个标准，与决策属性相关性强的优先考虑加入目标子集。

设有变量集$U=\{A_1,A_2,...A_n,C\}$，其中$A_1,A_2,...A_n$是实例的属性变量，C 是取 m 个值的类变量。

定义：如果$\psi(A_i,C)>\psi(A_i,C)$且$\psi(A_i,A_i)>\psi(A_i,C)$，则称$A_i$是$A_j$的冗余属性。

给定一组条件属性$A=\{A_1,A_2,...A_n\}$和决策属性 C，要从 A 中找出最优约简

子集 RedAttriSet，根据定义可以按以下步骤求得：

(1)先计算所有条件属性与决策属性之间的属性相关度$\psi(A_i,C)$；

(2)根据$\psi(A_i,C)$的绝对值将所有条件属性降序排列，得到条件属性序列 AttriSet;

(3)先选择 AttriSet 中的第一个属性 A_1（与决策属性相关度最强的属性），依次计算该属性与其他属性的相关度，若$\psi(A_1,A_i)>\psi(A_i,C)$，从 AttriSe 中删除 A_1的冗余属性 A_i；

(4)再选择 AttriSet 中的第二个属性 A_2，重复步骤 3，按上述方法依次删除 A_2冗余属性；

(5)得到约简子集 RedAttriSet 和约简后的数据集。

具体约简算法如下：

```
For i=1 to n(n=num(attribute))
    Do While not Eof()
      计算 fij
      Skip
    EndDo
ψi = ψ(Ai,C)
Next
AttriSet = DescendSorted(ψi)
For i=1 to n-1(n=num(attribute))
    j=i+1
    Do While not Eof()
      计算 fij
      Skip
    EndDo
ψij = ψ(Ai, Aj)
If  ψij > ψj
  AttriSet = RemoveElement(AttriSet, Aj)
EndIf
Next
RedAttriSet=AttriSet
```

算法复杂度分析

算法的第一部分是计算所有条件属性与决策属性的相关度，计算属性相关性的x^2统计量和ψ的时间复杂度为 $O(n^2)$，由复杂性理论可证明：当 n>2，$O(n^2)=O(n^2\log_2 n)$，排序的复杂度为 $O(n\log_2 n)$；算法的第二部分，每一次循环是计算条件属性 Ai 与其余条件属性的相关度，从而删除 Ai 的冗余属性，每次循

环的复杂度是 $O((AttriSet-1)^2)$，但在大多数情况下，由于算法首先删除许多无关属性，第一次循环常常能发现大量冗余属性，因此，实际复杂度远远小于理论计算值。

该算法不但能过滤掉属性集合中的无关属性，而且能有效地找到属性集合中的冗余属性，得到满意的属性约简子集，从而弱化了条件独立性假设的限制。

3.4 基于属性相关性的朴素贝叶斯分类模型 EANBC

3.4.1 EANBC 分类算法的理论依据

由第二章的理论介绍可以知道朴素贝叶斯分类的工作过程如下：

假定有 m 个类 $C_1,C_2,...C_m$，给定一个未知的数据样本 X(即没有类标号)，分类法将预测 X 属于具有最高后验概率(条件 X 下)的类，即朴素贝叶斯分类将未知的数据样本分配给类 Ci，当且仅当

$$P(C_i \mid X) > P(C_j \mid X), 1 < j < m, j \neq i$$

由此得到朴素贝叶斯分类的公式如下：$V_{NBC} = \arg\max P(C_i \mid X)$

其中，$P(C_i \mid X) = \dfrac{P(X \mid C_i)P(C_i)}{P(X)}$

由于 $P(X)$ 对于所有类为常数，只需计算最大 $P(X \mid C_i)P(C_i)$ 即可。计算 $P(C_i)$ 可以通过公式 $P(Ci) = Si/S$ 计算，其中 S_i 是类 C_i 中的训练样本数，S 是训练样本总数。

但是在实际应用中，对于给定具有许多条件属性的数据集，计算最大后验概率 $P(X \mid C_i)$，计算的开销可能非常大。为了降低计算的开销，朴素贝叶斯分类器作了条件独立假设，假定各属性相互条件独立，即在属性间不存在依赖关系，因此，

$$P(X \mid C_i) = \prod_{k=1}^{n} p(x_k \mid C_i)$$

概率 $P(x_k \mid C_i)$ 可以由训练样本计算，即 $P(x_k \mid C_i) = S_{ik}/S_i$，其中 S_{ik} 是在属性 A_k 上具有值 x_k 类 C_i 的训练样本数，而 S_i 是 C_i 中的训练样本数，由此得到 NBC 算法的分类公式：

$$V_{NBC} = \arg\max P(C_i)\prod_{k=1}^{n} P(x_k \mid C_i)$$

为测试未知样本 X 的分类，对于每个类 C_i，计算 $P(X \mid C_i)P(C_i)$，样本 X 被指派到 $P(X \mid C_i)P(C_i)$ 最大的类 C_i，即

$$P(X \mid C_i)P(C_i) > P(X \mid C_j)P(C_j), 1 < j < m, j \neq i$$

与其他分类算法相比，NBC 算法理论上具有最小的误分类率，但朴素贝叶斯分类器是基于一个简单的假定：即在给定分类特征条件下各属性之间是相互

条件独立的。在现实世界中，这种独立性假设经常是不能满足的，因此需要针对实际应用对朴素贝叶斯分类器模型进行改进，本文提出一种基于属性相关性度量的朴素贝叶斯分类算法 EANBC，就是对 NBC 算法进行改进，实现对 NBC 算法分类性能的优化。

3.4.2 EANBC 算法描述

基于属性相关性分析的朴素贝叶斯分类模型是对朴素贝叶斯分类模型的改进，EANBC 算法是将属性约简的方法与朴素贝叶斯的分类算法相结合而得到的。

算法实现要用到 3 个列表：

AttriSet：用于存放属性约简子集，AttriSet 中的各项代表一个条件属性，列表初始化为所有条件属性；

XTr：属性相关度列表，存放条件属性与决策属性的属性相关度 ψ_i，初始化为空；

CPTr：条件概率列表，用于存放 AttriSet 当前各属性相对于类别属性的条件概率。

算法主要步骤如下：

输入：建立数据库文件*. DBF，$D=\{X_1,X_2,...X_n\}, x_i=(A_1',A_2',...A_n',C_i)$，其中字段名 $A_1,A_2,...A_n$ 为原始属性集，字段名 $C=\{C_1,C_2,...C_m\}$ 为类别属性

输出：样本 X 的类标号；

第一步：计算条件属性与决策属性的相关度 ψ_i，结果放入 XTr；

第二步：根据 XTr 中值的绝对值的大小降序排列 AttriSet ；

第三步：依次计算各条件属性的相关度，得到属性约简子集 AttriSet：$Ai(i=1,..l,l<n)$；

第四步：计算后验条件概率[1] $P(C_i|X)$；

for $(j=1;j<m;j++)$ {

计算 $P(C_j)$；

while$(1<i<l)$ {

for $(k=1;k\le S_i;k++)$ 计算 $p(x_k|C_j)=S_{ik}/S_i$；

$i++$;}}

计算 $P(C_i|X)=P(C_i)\prod_{k=1}^{n}p(x_k|C_i)$，将结果存入 CPTr；

第五步：分配类标号；

for $(i=1;i<m;i++)$ {

if Max$(P(C_i|X))$

then 样本 X 类标号为 C_i；}

第六步：结束，输出类标号 C_i。

3.5 EANBC 实验结果及分析

3.5.1 实验数据

本实验所用的数据来自 UCI 机器学习数据库，该数据库是由加州大学欧文分校(University of California, Irvine，简称 UCI)，计算机科学与信息学院(Dept. of Information and Computer Sciences)提供的，(网址：http://www.ics.uci.edu/~mlearn/databases/)。我们从中选 4 个数据集分别为：Vote，Tic-Tac-Toe, Zoo ,Postoperative-Patient。下面对每个数据集作简要介绍：

(1) Vote

这是 1984 年美国国会选举记录的数据，数据最初来源于 Jeff Schlimmer 的论文，总共有 435 个实例，16 个条件属性和 1 个类别属性，所有属性均为离散值，类别属性有 2 种不同的取值，属性信息如下：

@relation Vote

@attribute Class Name{democrat,republican}

@attribute handicapped-infants{y,n}

@attribute water-project-cost-sharing{y,n}

@attribute adoption-of-the-budget-resolution{y,n}

@attribute physician-fee-freeze{y,n}

@attribute el-salvador-aid{y,n}

@attribute religious-groups-in-schools{y,n}

@attribute anti-satellite-test-ban{y,n}

@attribute aid-to-nicaraguan-contras{y,n}

@attribute mx-missile{y,n}

@attribute immigration{y,n}

@attribute synfuels-corporation-cutback{y,n}

@attribute education-spending {y,n}

@attribute superfund-right-to-sue {y,n}

@attribute crime{y,n}

@attribute duty-free-exports{y,n}

@attribute export-administration-act-south-africa{y,n}

(2) Tic-Tac-Toe

该数据来源于 David W. Aha，总共有 958 个实例，9 个条件属性和 1 个类别属性，所有属性均为离散值，类别属性有 2 种不同的取值，属性信息如下：

@relation Tic-Tac-Toe

@attribute top-left-square{x,o,b}

@attribute top-middle-square{x,o,b}

@attribute top-right-square{x,o,b}

@attribute middle-left-square{x,o,b}

@attribute middle-middle-square{x,o,b}

@attribute middle-right-square{x,o,b}

@attribute bottom-left-square{x,o,b}

@attribute bottom-middle-square{x,o,b}

@attribute bottom-right-square{x,o,b}

@attribute {positive,n ative}

(3) Zoo

该数据来源于 Richard S. Forsyth，总共有 101 个实例，17 个条件属性和 1 个类别属性，所有属性均为离散值，属性信息如下：

@relation zoo

@attributeanimal-name{aardvark,antelope,bass,bear,boar,buffalo,calf,carp,catfish,cavy,cheetah,chicken,chub,clam,crab,crayfish,crow,deer,dogfish,dolphin,dove,duck,elephant,flamingo,flea,frog,fruitbat,giraffe,girl,gnat,goat,gorilla,gull,haddock,hamster,hare,hawk,herring,honeybee,housefly,kiwi,ladybird,lark,leopard,lion,lobster,lynx,mink,mole,mongoose,moth,newt,octopus,opossum,oryx,ostrich,parakeet,penguin,pheasant,pike,piranha,pitviper,platypus,polecat,pony,porpoise,puma,pussycat,raccoon,reindeer,rhea,scorpion,seahorse,seal,sealion,seasnake,seawasp,skimmer,skua,slowworm,slug,sole,sparrow,squirrel,starfish,stingray,swan,termite,toad,tortoise,tuatara,tuna,vampire,vole,vulture,wallaby, wasp,wolf,worm,wren}

@attribute hair{0,1}

@attribute feathers{0,1}

@attribute eggs{0,1}

@attribute milk{0,1}

@attribute airborne{0,1}

@attribute aquatic {0,1}

@attribute predator{0,1}

@attribute toothed {0,1}

@attribute backbone{0,1}

@attribute breathes{0,1}

@attribute venomous{0,1}

@attribute fins{0,1}

@attribute legs{0,2,4,5,6,8}

@attribute tail{0,1}

@attribute domestic{0,1}

@attribute catsize {0,1}

@attribute type{1,2,3,4,5,6,7}

(4) Postoperative-Patient

该数据来源于 Jerzy W. Grzymala-Busse，总共有 90 个实例，8 个条件属性和 1 个类别属性，所有属性均为离散值，属性信息如下：

@relation Postoperative-Patient

@attribute L-CORE{high,mid,low}

@attribute L-SURF{high,mid,low}

@attribute L-O2{excellent,good,poor}

@attribute L-BP{ high , mid , low }

@attribute SURF-STBL{stable, mod-stable, unstable}

@attribute CORE-STBL{stable, mod-stable, unstable}

@attribute BP-STBL{stable, mod-stable, unstable}

@attribute COMFORT{05,07,10,15}

@attribute decision ADM-DECS{I ,S ,A }

3.5.2 实验平台及实验结果

一、实验平台

硬件环境：celeron(R)1. 70GHZ, 内存 256M

软件环境：操作系统 Windows XP Professional

实验工具：Visual Foxpro6.0,Weka[31]

实验过程：首先使用 EANBC 分类模型对实验数据集进行属性约简，对属性约简后的新的数据集，使用 Weka 系统中朴素贝叶斯分类工具求出分类的正确率，在测试过程中采用 10 次交叉验证法；同样使用 Weka 系统中朴素贝叶斯分类工具对未经属性约简处理的数据集进行分类，求出分类的正确率，在测试过程中同样采用 10 次交叉验证法；最后将得到的两种实验结果进行比较，完成本实验。

二、实验结果

对四个实验数据集属性约简的结果如下：

Vote：原条件属性 16 个，约简后条件属性 6 个

Tic-Tac-Toe：原条件属性 9 个，约简后条件属性 5 个

Zoo：原条件属性 17 个，约简后条件属性 13 个

Postoperative-Patient：原条件属性 8 个，约简后条件属性 3 个

EANBC 算法与 NBC 算法分类准确率比较见表 3-2

表 3-2　两种分类算法正确率比较

数据集	实例数	条件属性个数	约简后条件属性个数	分类正确率	
				NBC	EANBC
Vote	435	16	6	75.2%	88.3871%
Tic-Tac-Toe	958	9	5	69.3835%	73.2497%
Zoo	101	17	13	93.0693%	97.0297%
Postoperative-Patient	90	8	3	67.7778%	68.9655%

三、实验结果分析

实验的主要目的是对 EANBC 分类模型与 NBC 分类模型在每个数据集上的分类正确率进行比较，每个分类正确率是在测试集上成功预测的实例占总实例的百分比。实验结果明显表明，EANBC 分类模型在每个实验数据集上均取得了较好的分类性能，其分类性能明显优于 NBC 分类模型。通过对四个实验数据集进行跟踪实验，在属性约简的过程中，与决策属性相关性最大的属性，在进行第一轮属性约简后即得到属性约简子集，后面的属性约简计算均未删除冗余属性，这样属性约简计算的复杂度会降低，但对其它数据集能否适用有待于进一步探讨。

3.6 本章小结

本章介绍了基于 x^2 统计的属性相关性分析和属性约简的方法，提出一种基于属性相关性的 EANBC 分类算法模型，着重介绍构造 EANBC 分类模型的算法。通过在 UCI 机器学习数据集上对 EANBC 分类算法和 NBC 分类算法进行比较，在分类的正确率方面，EANBC 分类算法优于 NBC 分类算法。

第四　基于强属性限定的贝叶斯分类模型

本章对朴素贝叶斯分类模型的改进方法进行进一步的探讨，提出了一种基于强属性限定的贝叶斯分类模型 SANBC，给出了构造 SANBC 分类模型的算法和相应的实验结果。

4.1 问题的提出

朴素贝叶斯分类模型是一种简单而有效的分类方法，但它的条件独立性假设使其无法完全表达条件属性间存在的依赖关系，影响了它的分类性能。因此，人们采取放松朴素贝叶斯的独立性假设来改善朴素贝叶斯分类器的性能，而属性间的依赖关系与属性本身的特性紧密相关，有些属性本身所具有的特性决定了其他属性必然会依赖于它。因此，我们通过对由属性结点组成的属性空间进行搜索，找出一些对其他属性具有较强影响的属性，即强属性，所有其他属性通过与这些强属性的关联就可以将属性的依赖关系表达出来。

4.2 强属性的选择方法

令 S_1 和 S_2 是条件属性集 $\{A_1, A_2, \ldots\ldots A_n\}$ 的一个划分，S_1 是强属性集，S_2 是弱属性集。按照第三章介绍的属性相关性的分析方法，分别计算出每个条件属性与其它条件属性的相关性度量，然后计算每个条件属性与其它条件属性相关性度量的平均值，并按平均值的大小对条件属性进行排序，平均值大的条件属性显然对其他属性具有较强影响，可以作为强属性处理。具体算法描述如下：

(1) 令强属性集 S_1 为空，弱属性集 $S_2 = \{A_1, A_2, \ldots\ldots A_n\}$；

(2) 设最大强属性个数为 m；

(3) 计算所有条件属性之间的属性相关度 $\psi(A_i, A_j)$；

(4) 计算每个条件属性的属性相关度的平均值 $E\psi(A_i)$；

(5) 根据平均值 $E\psi(A_i)$ 将所有条件属性降序排列；

(6) 根据设定的 m 值，将强属性放入强属性集 S_1，并从弱属性集 S_2 中删除。

算法复杂度分析

算法的第一部分是计算所有条件属性之间的相关度，每次循环的复杂度是 $O((AttriSet-1)^2)$，算法的第二部分排序的复杂度为 $O(n\log_2 n)$。

4.3 基于强属性限定的贝叶斯分类模型 SANBC

4.3.1 贝叶斯定理的变形公式

令 S_1 和 S_2 是条件属性集 $\{A_1, A_2, \ldots\ldots A_n\}$ 的一个划分，s_1 和 s_2 分别是条件属性集 S_1 和 S_2 的取值，实例 $\{a_1, a_2, \ldots a_n\}$（或表示为 $(s_1\ s_2)$）属于类 c_j 的概率，可由贝叶斯定理的变形公式[19][32][33]表示为：

$$P(c_j \mid s_1, s_2) = \frac{P(s_2 \mid c_j, s_1)}{P(s_2 \mid s_1)} \cdot P(c_j \mid s_1)$$

$$= \beta \cdot P(s_2 \mid c_j, s_1) \cdot P(c_j \mid s_1) \quad (1)$$

其中 β 是一个正则化因子，假设 $s_1 = \{a_{k_1}, a_{k_2}, \ldots a_{k_m}\}$, $s_2 = \{a_{l_1}, a_{l_2}, \ldots\ldots a_{l_{n-m}}\}$，并且在给定 c_j 和 s_1 时，s_2 中的各属性是条件独立的，则式(1)可表示为

$$P(c_j \mid s_1, s_2) = \beta \cdot P(c_j \mid a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m}) \cdot P(s_2 \mid c_j, , a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m})$$

$$= \beta \cdot P(c_j \mid a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m}) \cdot \prod_{i=1}^{n-m} P(a_{l_i} \mid c_j, a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m}) \quad (2)$$

式(2)是假设在较少的条件属性(即 $a_{l_1}, a_{l_2}, \ldots\ldots a_{l_{n-m}}$)之间是条件独立的，该假设比朴素贝叶斯独立性假设要弱，从而通过贝叶斯变形公式弱化了朴素贝叶斯独立性假设。这个假设的强弱取决于强属性集 $S_1 = \{A_{k_1}, A_{k_2}, \ldots\ldots A_{k_m}\}$ 中属性的个数，强属性集 S_1 中属性的个数越多，条件独立性假设就越弱。又由于

$$P(c_j \mid a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m}) = \gamma \cdot P(c_j) \cdot \prod_{i=1}^{m} P(a_{k_i} \mid c_j, a_{k_1}, a_{k_2}, \ldots\ldots a_{k_{i-1}}) \quad (3)$$

其中 γ 是一个正则化因子，将式(3)代入式(2)等号的右侧，得

$$P(c_j \mid s_1, s_2) = \beta \cdot \gamma \cdot \prod_{i=1}^{m} P(a_{k_i} \mid c_j, a_{k_1}, a_{k_2}, \ldots\ldots a_{k_{i-1}}) \cdot \prod_{z=1}^{n-m} P(a_{l_z} \mid c_j, a_{k_1}, a_{k_2}, \ldots\ldots a_{k_m})$$

$$= \beta \cdot \gamma \cdot P(c_j) \cdot \prod_{i=1}^{m} P(a_{k_i} \mid c_j, K(a_{k_i})) \cdot \prod_{Z=1}^{n-m} P(a_{l_z} \mid c_j, K(a_{l_z}))$$

$$= \beta \cdot \gamma \cdot P(c_j) \cdot \prod_{i=1}^{n} P(a_i \mid c_j, K(a_i))$$

$$\propto P(c_j) \cdot \prod_{i=1}^{n} P(a_i \mid c_j, K(a_i))$$

其中 $K(a_i)$ 是 A_i 的非父类结点集 $K(A_i)$ 的取值，若 $A_i \in S_1$，则 $K(A_i) = (A_{k_1}, A_{k_2}, \ldots\ldots A_{k_m})$；若 $A_i \in S_2$，则 $K(A_i) = (A_{l_1}, A_{l_2}, \ldots\ldots A_{l_{n-m}})$。

4.3.2 基于强属性限定的贝叶斯分类模型 SANBC

基于强属性限定的贝叶斯分类模型是对朴素贝叶斯分类模型的结构进行了扩展。构造 SANBC 模型的关键是确定强属性集 $S_1 = \{A_{k_1}, A_{k_2}, \ldots\ldots A_{k_m}\}$，$S_1$ 中的任意两个属性可以有依赖关系，给定 S_1 和 C，$S2$ 中任意两个属性是条件独立的，S_1 中的属性可以 $S2$ 中每个属性的非类父结点，因此可以在属性之间添加增强弧以弱化朴素贝叶斯的独立性假设，从而构造出 SANBC 模型。

对于给定某一实例$\{a_1, a_2, \ldots\ldots a_n\}$，计算

$V_{SANBC} = \arg\max P(c_j)\prod_{i=1}^{n} P(a_i \mid c_j, K(a_i))$的最大类$c_j$作为该实例的类标签。算法描述如下：

输入：$D = \{X_1, X_2, \ldots X_n\}, x_i = (A_1', A_2', \ldots A_n', C_i)$，其中$A_1, A_2, \ldots A_n$为原始属性集，$C = \{C_1, C_2, \ldots C_m\}$为类别属性；

输出：样本 X 的类标号；

第一步：令强属性集S_1为空，弱属性集$S_2 = \{A_1, A_2, \ldots\ldots A_n\}$；

第二步：设最大强属性个数为 m；

第三步：计算所有条件属性之间的属性相关度$\psi(A_i, A_j)$；

第四步：计算每个条件属性的属性相关度的平均值$E\psi(A_i)$；

第五步：根据平均值$E\psi(A_i)$将所有条件属性降序排列；

第六步：根据设定的 m 值，将强属性放入强属性集S_1，并从弱属性集S_2中删除；

第七步：计算后验条件概率$V_{SANBC} = \arg\max P(c_j)\prod_{i=1}^{n} P(a_i \mid c_j, K(a_i))$；

第八步：分配类标号；

第九步：结束，输出类标号c_i。

由于第七、八、九步的算法在第三章 EANBC 算法中已详细介绍，在此不在赘述。下面主要对$P(a_i \mid c_j, K(a_i))$的计算方法进行详细介绍。

4.3.3 $P(a_i \mid c_j, K(a_i))$的计算方法

下面以强属性个数为 1 时为例，对$P(a_i \mid c_j, K(a_i))$的计算方法进行介绍：

假定有 m 个类 $C_1, C_2, \ldots C_m$，强属性为A_k，其属性取值为$\{a_{k1}, a_{k2}, \ldots\ldots a_{kj}\}$，SANBC 算法的先验条件概率分别为：

$P(C = c_1, A_k = a_{k1})$　$P(C = c_1, A_k = a_{k2})$ …… $P(C = c_1, A_k = a_{kj})$

$P(C = c_2, A_k = a_{k1})$　$P(C = c_2, A_k = a_{k2})$ …… $P(C = c_2, A_k = a_{kj})$

……

$P(C = c_m, A_k = a_{k1})$　$P(C = c_m, A_k = a_{k2})$ …… $P(C = c_m, A_k = a_{kj})$

$P(a_i \mid c_j, A_k)$的计算方法为：

$P(a_i \mid c_1, a_{k1}) = P(x = a_i, C = c_1, A_k = a_{k1}) / P(C = c_1, A_k = a_{k1})$

$P(a_i \mid c_1, a_{k2}) = P(x = a_i, C = c_1, A_k = a_{k2}) / P(C = c_1, A_k = a_{k2})$

……

$P(a_i \mid c_1, a_{kj}) = P(x = a_i, C = c_1, A_k = a_{kj}) / P(C = c_1, A_k = a_{kj})$

$P(a_i \mid c_2, a_{k1}) = P(x = a_i, C = c_2, A_k = a_{k1}) / P(C = c_2, A_k = a_{k1})$

$P(a_i \mid c_2, a_{k2}) = P(x = a_i, C = c_2, A_k = a_{k2}) / P(C = c_2, A_k = a_{k2})$

……

$$P(a_i \mid c_2, a_{kj}) = P(x = a_i, C = c_2, A_k = a_{kj}) / P(C = c_2, A_k = a_{kj})$$

……

$$P(a_i \mid c_m, a_{k1}) = P(x = a_i, C = c_m, A_k = a_{k1}) / P(C = c_m, A_k = a_{k1})$$

$$P(a_i \mid c_m, a_{k2}) = P(x = a_i, C = c_m, A_k = a_{k2}) / P(C = c_m, A_k = a_{k2})$$

……

$$P(a_i \mid c_m, a_{kj}) = P(x = a_i, C = c_m, A_k = a_{kj}) / P(C = c_m, A_k = a_{kj})$$

通过比较，SANBC 算法在引入强属性后，其条件概率受到一定的限定，从而限制了条件概率的计算范围，尽管算法的复杂度比朴素贝叶斯分类模型有所增加，但分类的正确率有很大提高。

4.4 SANBC 实验结果及分析

4.4.1 实验数据

实验数据采用第三章已介绍的三个实验数据集 Vote、Tic-Tac-Toe 和 Postoperative-Patient

4.4.2 实验平台

硬件环境：celeron(R)1. 70GHZ, 内存 256M

软件环境：操作系统 Windows XP Professional

实验工具：Visual Foxpro6.0,Weka

实验过程：首先使用 SANBC 分类模型对三个实验数据集进行分类，然后使用 Weka 系统中朴素贝叶斯分类工具对三个实验数据集进行分类，分别得到两种算法分类的正确率，最后将得到的两种实验结果进行比较，完成本实验。实验结果如下表 4-1 所示：

表 4-1　两种算法分类正确率比较

数据集	实例数	属性个数	分类正确率		
			NBC	SANBC	
				m=1	m=2
Vote	435	16	75. 6%	92. 8358%	95. 5223%
Tic-Tac-Toe	958	9	69. 8015%	77. 9519%	81. 1912%
Postoperative-Patient	90	8	74. 7126%	75. 862%	83. 908%

4.4.3 实验结果分析

实验的主要目的是对 SANBC 分类模型与 NBC 分类模型在三个数据集上的分类正确率进行比较，每个分类正确率是在测试集上成功预测的实例占总实例

的百分比。实验结果明显表明，SANBC 分类模型在每个实验数据集上均取得了较好的分类性能，其分类性能明显优于 NBC 分类模型。从实验的过程分析，当强属性的个数为 1 时，前两个数据集分类的正确率有较大幅度的提高；当强属性个数为 2 时，第三个数据集分类的正确率有较大幅度的提高；继续增加强属性的个数，分类的正确率也许可能进一步提高。但对于不同的数据集，最佳的强属性的个数可能会有所不同，强属性的个数多少是最佳的，这还有待于进一步探讨。

4.5 本章小结

本章介绍一种基于强属性限定的贝叶斯分类模型 SANBC，给出了构造 SANBC 分类模型的算法，通过在 UCI 机器学习数据集上对 SANBC 分类算法和 NBC 分类算法进行比较，在分类的正确率方面，SANBC 分类算法优于 NBC 分类算法。

第五章　朴素贝叶斯分类模型在指导学生选择专业方向中的应用

本章主要将朴素贝叶分类技术应用到实际的教学指导中，用于指导学生根据自己的专业知识结构合理地选择专业方向。

5.1 问题的提出

目前，为适应教育现代化的需要，培养适应社会主义现代化建设的创新人才，大学的培养目标正从原来的培养“适应性人才”转向培养“创新性人才”。相应地，培养出既能适应又能引导社会发展的“创新性人才”，这便成为了大学课程结构的改革取向。而课程体系是人才培养模式的具体表现和核心内容。近几十年来，全世界各国都在研究符合时代特征的、体现现代教育思想的口等教育人才培养模式，建立与此相适应的课程结构体系。我国现课程设置体现“新、精，现”的特点，课程体系体现“大专业、宽专业基础、多方向”的原则。但面对众多课程，学生对课程间的关联关系不甚清晰，对专业方向的选择非常迷茫，因此，学生在选择专业方向时往往受到三个方面因素的影响：一是他人因素的影响，如教师、父母、同学；二是个人因素，如个人的兴趣、喜好；三是社会因素的影响，如通常选择择业前景好的专业方向。当然个人的兴趣、教师、父母、同学的建议以及好的就业前景对学生的选择并不一定就起误导作用，但如何根据自己的专业知识结构科学合理地选择专业方向往往是困惑大多数学生的一个问题。

因此，将数据挖掘技术应用到教育领域内的学生来源分析、课程相关性、专业方向选择指导、学习评价度量、学生生涯规划等方面必然有十分广阔的前景。本章主要从朴素贝叶斯分类模型在指导学生选择专业方向方面的应用进行探讨。

5.2 建立专业方向选择的分类模型

为了更好地指导学生根据自己的专业知识结构以及专业知识的掌握程度科学合理地选择专业方向，就必须运用数据挖掘中的分类技术，充分利用以往各届学生选择专业方向的先验知识，建立专业方向选择的分类模型，对需要指导的学生根据他们的学习情况进行分类预测，帮助他们合理地选择专业方向。

5.2.1 运用朴素贝叶斯分类器建立分类模型

朴素贝叶斯分类器已在第二章、第三章进行了详细地介绍，下面主要介绍朴素贝叶斯分类器在专业方向选择中的应用。

一、建立学生专业方向选择数据库

数据来源：淮南师范学院计算机科学与技术专业某届本科生的专业基础课、专业课成绩以及专业方向的选择。

专业课成绩以及专业方向的选择。

数据预处理：采用四级等级制对学生的成绩进行评价，即优、良、中、及格。

数据说明：本实验数据共有 72 个实例，19 个条件属性和 1 个类别属性，所有属性均为离散值，其中类别属性为计算机专业方向：硬件方向和软件方向。条件属性信息如下：

(1)C 程序设计：优、良、中、及格
(2)模拟电路：优、良、中、及格
(3)数据结构：优、良、中、及格
(4)数字电路：优、良、中、及格
(5)汇编语言：优、良、中、及格
(6)操作系统：优、良、中、及格
(7)数据库：优、良、中、及格
(8)计算机组成：优、良、中、及格
(9)微机原理：优、良、中、及格
(10)接口技术：优、良、中、及格
(11)编译原理：优、良、中、及格
(12)计算机图形学：优、良、中、及格
(13)计算机网络：优、良、中、及格
(14)JAVA 程序设计：优、良、中、及格
(15)软件工程：优、良、中、及格
(16)信号处理系统：优、良、中、及格
(17)现代网络技术：优、良、中、及格
(18)多媒体技术应用：优、良、中、及格
(19)人工智能：优、良、中、及格

用于分类的实验数据集见附录：某届计算机科学与技术专业学生成绩信息表。

二、构造用于专业方向选择的朴素贝叶斯分类模型

运用第三章介绍的朴素贝叶斯分类器的方法建立分类模型。将《某届计算机科学与技术专业学生成绩信息表》作为训练集，充分利用先验知识和现有的统计数据，用概率的方法预测未知事件发生的可能性。就本实验数据集而言，计算机专业方向为类别属性：分别为硬件方向和软件方向，其中选择硬件方向的学生为 17 人，选择软件方向的学生为 55 人，因此，选择硬件方向和软件方向的先验概率分别为：P(C1 = 软件) = 55/72 = 76.4%, P(C2 = 硬件) = 17/72=23.6%。

5.2.2 对学生实例进行分类预测

为说明用于专业方向选择的朴素贝叶斯分类模型预测一个未知样本的类标

号的工作过程，我们利用《某届计算机科学与技术专业学生成绩信息表》为训练数据样本，利用现有的统计数据和先验概率预测未知样本的所属类，即预测某个学生 X 选择硬件方向或软件方向的最大可能性。

实例：对我院 06 届计算机科学与技术专业本科生的专业选择方向进行预测。某个学生 X 各科成绩分别为：C 程序设计＝及格，模拟电路＝中，数据结构＝中，数字电路＝良，汇编语言＝中，操作系统＝良，数据库＝中，计算机组成＝中，微机原理＝中，接口技术＝优，编译原理＝中，计算机图形学＝优，计算机网络＝良，JAVA 程序设计＝中，软件工程＝良，信号处理系统＝中，现代网络技术＝中，多媒体技术应用＝良，人工智能＝良。

要预测该学生 X 的专业方向，需要最大化 P(X|Ci)P(Ci), i=1, 2，每个类的先验概率分别为 P(C1＝软件)＝55/72, P(C2＝硬件)＝17/72。

条件概率分别为：

P(C 程序设计＝及格|硬件)＝9/17

P(C 程序设计＝及格|软件)＝42/55

P(模拟电路＝中|硬件)＝7/17

P(模拟电路＝中|软件)＝21/55

P(数据结构＝中|硬件)＝8/17

P(数据结构＝中|软件)＝11/55

P(数字电路＝良|硬件)＝4/17

P(数字电路＝良|软件)＝13/55

P(汇编语言＝中|硬件)＝3/17

P(汇编语言＝中|软件)＝11/55

P(操作系统＝良|硬件)＝4/17

P(操作系统＝良|软件)＝2/55

P(数据库＝中|硬件)＝7/17

P(数据库＝中|软件)＝12/55

P(计算机组成＝中|硬件)＝6/17

P(计算机组成＝中|软件)＝14/55

P(微机原理＝中|硬件)＝4/17

P(微机原理＝中|软件)＝18/55

P(接口技术＝优|硬件)＝7/17

P(接口技术＝优|软件)＝16/55

P(编译原理＝中|硬件)＝9/17

P(编译原理＝中|软件)＝12/55

P(计算机图形学＝优|硬件)＝6/17

P(计算机图形学＝优|软件)＝4/55

P(计算机网络＝良|硬件)＝6/17

P(计算机网络＝良|软件)＝10/55

P(JAVA 程序设计＝中|硬件)＝7/17

P(JAVA 程序设计＝中|软件)＝32/55

P(软件工程＝良|硬件)＝6/17

P(软件工程＝良|软件)＝7/55

P(信号处理系统＝中|硬件)＝2/17

P(信号处理系统＝中|软件)＝28/55

P(现代网络技术＝良|硬件)＝8/17

P(现代网络技术＝良|软件)＝11/55

P(多媒体技术应用＝良|硬件)＝13/17

P(多媒体技术应用＝良|软件)＝22/55

P(人工智能＝良|硬件)＝2/17

P(人工智能＝良|软件)＝12/55

由以上计算可得到：

P(X|硬件)＝9/17×7/17×8/17×4/17×3/17×4/17×7/17×6/17×4/17×7/17×9/17×6/17×6/17×7/17×6/17×2/17×8/17×13/17×2/17＝1.54E-11

P(X|软件)＝42/55×21/55×11/55×13/55×11/55×2/55×12/55×14/55×18/55×16/55×12/55×4/55×10/55×32/55×7/55×28/55×11/55×22/55×12/55＝9.94E-14

P(X|硬件)×P(C2＝硬件)＝0.0000000000154×17/72=3.64E-12

P(X|软件)×P(C1＝软件)＝0.0000000000000994×55/72＝7.59E-14

由于 P(X|硬件)×P(C2＝硬件)＞P(X|软件)×P(C1＝软件)，因此学生选择硬件方向的可能性最大，这一预测结果与学生实际选择的情况相符。通过这个过程介绍如何利用现有的先验知识对学生专业方向的选择进行预测，它也表明朴素贝叶斯分类方法在实际应用中具有一定的分类准确性。

当然，影响学生专业方向的选择不一定只是学生各门功课的学习成绩，可能学生的个人兴趣、专业特长、实践能力、择业方向以及其它一些未考虑到的因素都会对学生专业方向的选择产生一定的影响。由于作为训练集的这一届学生已毕业，一些相关的信息无法获取，争取在下一步的工作中通过问卷调查的形式获取更多有价值的信息。

5.3 本章小结

本章只是对朴素贝叶斯分类模型在指导学生选择专业方向方面的应用作一个尝试，该方法亦可以通过删减属性(即课程)对学生在不同的学习阶段进行专业方向选择的分类预测,以便指导学生及早为专业方向的选择做好准备，有针对

性地加强相关专业课程的学习。此外，还可以通过属性(课程)相关性分析，对专业课程进行模块划分，使学生对专业课程间的相关性有更加清晰的认识，这将在下一步工作中进行探讨。当然该方法的合理程度对数据有较大的依赖性，由于我们学校只有一届毕业生，历史数据相当缺乏，在今后的工作中逐步收集原始数据，以使数据集更加完善合理，从而更好地指导学生选择专业方向。

第六章　总结与展望

6.1 本文总结

本文的工作是朴素贝叶斯分类模型的研究和应用。贝叶斯分类技术是以贝叶斯定理、最大后验假设、贝叶斯网络理论以及信息学理论为基础的。本文介绍了贝叶斯分类的基本理论，并在此基础上总结了几种常用的贝叶斯分类模型：朴素贝叶斯分类模型、贝叶斯网络模型及增量贝叶斯分类模型。本文的重点是朴素贝叶斯分类模型的改进方法。

朴素贝叶斯分类模型由于其简单和易于实现的特点，受到人们的普遍青睐，其性能可以与神经网络、决策树媲美。目前对于它的研究工作主要集中在探讨它的条件独立性假设和改善分类性能方面。本文在介绍属性相关性度量的基础上，提出了两种贝叶斯分类模型：将属性约简与朴素贝叶斯分类模型相结合，提出一种基于属性相关性度量的朴素贝叶斯分类模型；将强属性的选择方法与贝叶斯变形公式相结合，提出一种基于强属性限定的贝叶斯分类模型。此外，针对我院学生在选择专业方向中存在的问题，将朴素贝叶斯分类模型用于专业方向的选择，以指导学生科学合理地选择专业方向。

综上所述，本文的创新性工作表现在如下四个方面：

(1)研究基于x^2估计的属性相关性度量，在此基础上提出条件属性的约简方法。

(2)在对朴素贝叶斯分类模型的研究的基础上，将条件属性的约简方法与与朴素贝叶斯分类方法相结合，提出一种基于属性相关性度量的朴素贝叶斯分类模型 EANBC。实验表明，EANBC 算法分类的正确率优于 NBC 算法。

(3)通过分析贝叶斯定理的变形公式和属性相关性度量，介绍强属性的选择方法，并提出一种基于强属性限定的贝叶斯分类模型 SANBC，实验表明，SANBC 算法分类的正确率优于 NBC 算法。

(4)朴素贝叶斯分类模型在专业方向的选择方面的应用。指导学生根据自己的专业知识结构合理地选择专业方向，为学生的学习起一个科学导向作用。

6.2 工作展望

(1)本文虽然提出一种基于属性相关性度量的朴素贝叶斯分类模型 EANBC，但条件属性约简子集是否是最优约简子集仍有待于进一步研究。

(2)本文的算法只与朴素贝叶斯分类算法进行了比较，但与其它一些分类技术相比有没有其优越性，性能如何，这也是下一步将要考虑的问题。

(3)本文的算法跟其它相关分类算法，能否协同工作，有无结合点，也是将要考虑的问题。

(4)联合分类器是分类领域的发展趋势，能否将朴素贝叶斯分类模型与其它模型结合，从而提高分类性能同样是下一步研究的目标。

参考文献

[1] 赵斌. 关联规则分布式挖掘算法研究和实现，南京师范大学硕士学位论文，2003.

[2] U. Fayyad M. Piatetsky-Shapiro G. Smyth, From Data Mining to Knowledge Discovery; An Overview. In: Advances in Knowledge Discovery and Data Mining. Menlo Park, California: AAAI Press/The MIT Press, 1996, p1-35.

[3] 王德兴. 基于量化概念格的关联规则挖掘模型研究，合肥工业大学硕士学位论文，2003.

[4] 姜卯生. 数据挖掘中基于贝叶斯技术的分类问题的研究，合肥工业大学硕士学位论文，2004.

[5] W. J. Frawley, G. Piatetsky, C. Shapiro, J. Matheus, Knowledge Discovery in Databases: An Overview. In Piatetsky-Shapiro,W. J. Frawley eds. Knowledge Discovery in Databases. Menlo Park, California: AAAI Press/The MIT Press,1991，p1~27.

[6] A Perspective on Databases and Data Mining， In Proceedings of the First International Knowledge on Discovery and Data Mining，(KDD95), p150-155.

[7] 谈恒贵等. 数据挖掘算法综述，微型机与应用，2005 年第 2 期.

[8] HanJiawei,KamberM. 数据挖掘概念与技术[M].范明，孟小峰等译.北京:机械工业出版社,2001.

[9] 张海笑. 数据挖掘中分类方法的研究，山西电子技术 2005 年第 2 期.

[10] 于莉. 基于高校学生信息库的数据挖掘,2004

[11] 基于粗糙集合和朴素贝叶斯模型的分类问题研究，合肥工业大学硕士学位论文，2005.

[12] Jiawei Han and Michelins Kamber. DATA MINING Concepts and Techniques, Higher Education Press，Morgan Kaufmann Publishers,2001.

[13] Tom M. Mitchell 著,曾华军，张银奎等译. 机器学习，机械工业出版社，2003 年.

[14] 张云涛,龚玲著. 数据挖掘原理与技术，电子工业出版社 2004 年 4 月第 1 版.

[15] Mia K. Stern, Joseph E. Beverly Park Woof . Native Bayes Classifiers for User Modeling.

[16] Pedro Domingos, Michael Pazzzani. On the Optimality of the Simple Bayesian Classifier under zero-one Loss. Machine Learning. 29,103-130,1997.

[17] C.Elkan. Boosting and Naive Bayesian Learning. In Technical Report CS97, Dept .of Computer Science and Engineering, Univ. Calif. at San Diego,Sept.1997.

[18] 张璠. 多种策略改进朴素贝叶斯分类器，微机发展，2005 年第 15 卷，第 4 期.

[19] 石洪波. 一种限定性的双层贝叶斯分类模型 软件学报 2004.Vol.43 No.2.

[20] 林士敏，田凤占，陆玉昌.《用于数据挖掘的贝叶斯分类器研究》,计算机科学 2000 27(10) 73-76.

[21] Mehran Sahami. Learning Limited Dependence Bayesian Classifers. Computer Science Department Stanford University,1997 [5] Lior Rokach,Oded Mainon. Theory and Applications of Attribute Decomposition. Department of Industrial Engineering.

[22] 张剑飞. 贝叶斯网络学习方法和算法研究，东北师范大学硕士学位论文,2005.

[23] 罗宁,穆志纯. 基于贝叶斯网的分类器及其在 CRM 中的应用,计算机应用,第 24 卷第 3 期 2004 年 3 月.

[24] Maurice Pagnucco. Reasoning Under Uncertainty and Bayesian Networks.COMP9414, 2002.

[25] David Maxwell，David Herkerman, Christopher Meek. Technical Report: A Bayesian Approach to Learning Bayesian Networks with Local Structure .March 1997.

[26] 宫秀军，刘少辉，史忠植. 一种增量贝叶斯分类模型,计算机学报 2002,25(6):645-650.

[27] 宫秀军，孙建平，史忠植. 主动贝叶斯网络分类器,计算机研究与发展 2002 39(5):574-579.

[28] 王大玲. 一种基于关联性度量的决策树分类方法,东北大学学报 2001.Vol.22 No.5.

[29] 韩家新. 一种以相关性确定条件属性的决策树,微机发展 2003.Vol.13 No.5.

[30] 张静. 基于属性相关性的属性约简新方法,计算机工程与应用 2005.28.

[31] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Seattle: MorganKaufmann Publishers, 2000. 265～314.

[32] Lu RQ. Artificial Intelligence. Beijing: Science Press, 1989.1134～1147 (in Chinese)

[33] Zheng Z,Webb GI.Lazy learning of Bayesian rules.Machine Learning,2000.41(1):53～84.

附录：某届计算机科学与技术专业学生成绩信息表

序号	C语言	模拟电路	数据结构	数字电路	汇编语言	操作系统	数据库	计算机组成	微机原理	接口技术
1	中	良	中	良	优	中	及格	及格	中	良
2	及格	及格	及格	及格	及格	及格	及格	中	及格	良
3	中	中	良	中	良	中	中	中	良	良
4	及格	及格	中	中	良	中	良	中	中	优
5	及格	中	中	中	良	及格	中	良	及格	良
6	及格	中	及格	良	及格	及格	及格	及格	中	良
7	中	及格	中	及格	中	中	中	及格	中	良
8	及格	及格	及格	及格	及格	中	及格	及格	及格	良
9	及格	中	及格	良	中	中	及格	及格	及格	中
10	优	良	良	优	优	良	良	良	良	优
11	及格	中	及格	中	优	良	及格	及格	良	优
12	及格	及格	及格	中	及格	及格	中	中	及格	优
13	及格	及格	及格	中	良	及格	及格	及格	及格	良
14	中	中	中	良	良	良	中	中	优	优
15	及格	良	中	中	优	及格	良	中	良	优
16	优	中	良	良	中	优	中	良	中	良
17	中	中	良	中	良	中	中	良	良	优
18	及格	及格	及格	中	及格	及格	及格	及格	及格	良
19	及格	及格	及格	良	及格	及格	及格	及格	及格	优
20	及格	中	中	良	及格	中	及格	良	及格	优
21	中	中	中	优	良	良	中	良	良	优
22	中	良	及格	中	及格	中	及格	良	良	中
23	及格	及格	及格	中	及格	及格	中	及格	中	良
24	良	及格	中	中	良	中	及格	及格	及格	中
25	及格	及格	及格	及格	及格	及格	中	及格	及格	优
26	中	中	良	良	良	优	良	良	优	优
27	中	良	良	良	优	良	中	良	良	优
28	及格	及格	及格	及格	及格	及格	及格	中	中	良
29	及格	及格	及格	及格	中	及格	及格	及格	及格	良
30	中	及格	中	及格	及格	良	中	中	及格	良
31	及格	及格	及格	及格	及格	及格	及格	及格	及格	良
32	中	中	中	良	中	中	及格	中	中	良
33	及格	及格	及格	中	及格	及格	及格	中	及格	中
34	及格	及格	及格	及格	中	及格	及格	及格	及格	良
35	及格	中	中	良	良	及格	中	中	中	优
36	及格	及格	及格	及格	及格	及格	及格	及格	及格	良

序号	C 语言	模拟电路	数据结构	数字电路	汇编语言	操作系统	数据库	计算机组成	微机原理	接口技术
37	及格	及格	及格	良	良	及格	及格	及格	及格	优
38	及格	及格	及格	及格	及格	及格	及格	及格	及格	良
39	及格	中	及格	及格	良	及格	及格	及格	及格	良
40	及格	中	及格	中	及格	及格	及格	及格	中	良
41	及格	及格	及格	良	良	中	中	及格	中	良
42	中	中	良	及格	及格	及格	及格	中	及格	良
43	及格	中	及格	及格	及格	及格	及格	中	及格	良
44	中	良	及格	及格	中	及格	及格	及格	及格	良
45	及格	中	中	良	中	及格	及格	及格	及格	优
46	及格	中	及格	中	良	及格	及格	及格	中	良
47	及格	及格	中	良	中	及格	中	良	及格	良
48	及格	中	中	中	良	中	及格	及格	中	良
49	及格	优	中	中	良	及格	中	良	中	优
50	中	良	及格	及格	及格	及格	及格	中	及格	良
51	及格	及格	及格	及格	及格	及格	及格	及格	及格	良
52	及格	中	及格	中	中	及格	及格	及格	及格	优
53	及格	及格	及格	及格	护林	及格	及格	及格	及格	良
54	及格	中	及格	中	中	及格	及格	及格	中	良
55	及格	良	中	中	良	中	中	中	中	优
56	及格	良	及格	及格	良	良	及格	及格	良	良
57	及格	中	中	中	良	中	及格	良	中	优
58	及格	中	及格	优	及格	及格	及格	中	及格	良
59	及格	良	及格	及格	良	及格	及格	及格	及格	良
60	及格	及格	及格	中	及格	中	及格	及格	及格	良
61	及格	中	及格	中	中	及格	及格	及格	中	优
62	中	优	及格	良	良	及格	及格	及格	及格	良
63	中	良	及格	中	中	中	中	良	中	良
64	中	及格	中	良	良	优	及格	中	中	优
65	及格	及格	及格	中	良	及格	及格	及格	及格	良
66	及格	良	及格	中	及格	及格	及格	及格	及格	中
67	及格	良	及格	及格	及格	及格	及格	及格	及格	良
68	及格	中	及格	及格	良	及格	及格	中	中	优
69	中	中	及格	中	良	中	中	中	及格	良
70	及格	中	及格	及格	及格	及格	及格	及格	及格	良
71	及格	良	及格	中	中	及格	及格	中	中	良
72	及格	及格	及格	及格	及格	及格	及格	及格	及格	及格

序号	编译原理	图形学	计算机网络	JAVA	软件工程	信号处理	现代网络技术	多媒体应用	人工智能	专业方向
1	及格	良	优	中	良	优	良	中	中	软件
2	中	及格	及格	良	及格	中	良	中	中	硬件
3	及格	良	优	良	良	优	及格	良	中	软件
4	中	良	中	中	良	良	良	良	中	硬件
5	中	良	良	良	及格	良	中	良	中	硬件
6	及格	中	中	中	中	良	良	良	良	软件
7	中	优	良	中	及格	良	中	良	良	硬件
8	及格	及格	良	及格	中	中	及格	中	良	软件
9	及格	良	中	良	中	良	及格	中	及格	软件
10	良	优	优	良	良	优	良	优	中	硬件
11	中	优	优	优	良	良	优	良	中	硬件
12	及格	良	中	良	中	良	良	良	中	软件
13	及格	良	中	中	中	良	良	中	良	软件
14	中	优	优	及格	中	优	良	良	中	硬件
15	中	良	良	良	良	优	良	良	中	软件
16	中	优	良	良	良	良	良	良	良	硬件
17	及格	优	优	良	良	良	良	良	中	硬件
18	及格	及格	中	及格	及格	中	及格	良	良	软件
19	及格	中	中	中	及格	中	中	良	良	软件
20	及格	良	良	中	良	良	良	良	中	硬件
21	中	良	良	良	良	良	良	良	中	软件
22	中	良	良	中	及格	良	良	良	中	硬件
23	及格	中	中	中	中	中	中	中	中	软件
24	及格	良	中	中	及格	良	中	中	中	硬件
25	及格	中	中	中	及格	中	中	良	中	软件
26	中	良	优	良	及格	优	中	优	中	软件
27	良	优	优	及格	中	良	良	良	中	软件
28	及格	中	中	中	及格	优	中	及格	中	软件
29	及格	良	中	及格	及格	中	中	良	中	硬件
30	及格	良	良	中	中	良	中	良	中	硬件
31	及格	良	优	中	及格	中	中	良	中	软件
32	中	优	优	中	中	中	中	良	良	软件
33	及格	及格	中	及格	及格	良	中	中	中	硬件
34	及格	中	中	中	中	中	及格	中	中	软件
35	中	良	优	优	中	良	中	良	中	硬件
36	及格	及格	及格	中	及格	中	中	中	中	软件

序号	编译原理	图形学	计算机网络	JAVA	软件工程	信号处理	现代网络技术	多媒体应用	人工智能	专业方向
37	中	中	中	中	中	中	及格	中	良	软件
38	及格	良	及格	中	中	良	及格	良	中	硬件
39	及格	及格	中	优	及格	中	及格	中	良	
40	及格	及格	及	中	及格	中	中	中	中	软件
41	及格	良	中	中	中	中	及格	中	良	软件
42	及格	良	中	中	及格	中	中	良	中	软件
43	及格	中	良	中	中	良	良	中	中	软件
44	及格	中	及格	中	及格	中	及格	中	中	软件
45	中	中	良	良	中	中	中	中	中	软件
46	及格	良	中	良	中	良	及格	中	中	软件
47	中	良	优	良	中	良	及格	中	中	软件
48	及格	良	良	中	中	良	中	良	中	软件
49	中	优	优	良	中	良	中	中	中	软件
50	及格	良	中	中	中	良	中	中	良	软件
51	及格	中	及格	中	中	中	及格	良	中	软件
52	及格	良	中	及格	中	良	及格	中	中	软件
53	及格	中	中	中	及格	良	中	良	中	软件
54	及格	及格	中	中	良	中	及格	及格	中	软件
55	及格	良	优	中	中	优	中	中	良	软件
56	及格	中	优	良	中	优	良	良	中	软件
57	中	优	优	中	中	优	中	中	良	软件
58	及格	良	及格	及格	中	中	及格	及格	及格	软件
59	及格	及格	及格	中	中	中	及格	中	中	软件
60	及格	优	中	中	中	中	中	良	中	软件
61	及格	良	中	中	中	良	良	良	中	软件
62	中	中	良	及格	中	中	中	中	中	软件
63	中	良	优	良	中	良	中	良	中	软件
64	中	良	良	良	良	良	及格	良	中	软件
65	及格	良	中	中	及格	中	中	良	中	软件
66	及格	及格	及格	及格	及格	中	及格	及格	中	软件
67	及格	及格	及格	中	及格	中	中	中	中	软件
68	及格	良	良	中	良	良	中	中	中	软件
69	及格	良	中	中	及格	良	中	中	中	软件
70	及格	及格	及格	及格	及格	中	中	中	及格	软件
71	及格	中	良	中	中	中	及格	良	中	软件
72	及格	及格	及格	及格	中	中	良	中	中	软件

在读期间发表的学术论文

1.王峻. 一种基于强属性限定的贝叶斯分类模型，微机发展，2007 年第 1 期或第 2 期.

2.王峻. 基于 Apriori 算法的多级数据选择器实现逻辑函数的最优化设计，大学时代学术教育，2006 年第 1 期.