分类号：　　密级：

兰州大学

研 究 生 学 位 论 文

论文题目（中文）	基于贝叶斯方法进行中医证候诊断的研究
论文题目（外文）	**Based on Bayesian Method to Research the Diagnosis of Syndromes in TCM**
研 究 生 姓 名	张彦净
学 科 、 专 业	软件工程·软件工程
研 究 方 向	生物医学数据挖掘
学 位 级 别	硕　士
导师姓名、职称	郑光　副教授
论 文 工 作 起 止 年 月	**2016 年 12 月至 2017 年 04 月**
论 文 提 交 日 期	**2017 年 04 月**
论 文 答 辩 日 期	**2017 年 05 月**
学 位 授 予 日 期	

校址：甘肃省兰州市

学　　院： 信息科学与工程学院　　**学　　号：** 220140928680

学生姓名： 张彦净　　**导师姓名：** 郑光

学科名称： 软件工程·软件工程

论文题目： 基于贝叶斯方法进行中医证候诊断的研究

原 创 性 声 明

本人郑重声明：本人所呈交的学位论文，是在导师的指导下独立进行研究所取得的成果。学位论文中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究成果做出重要贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任由本人承担。

论文作者签名：_____　　日　期：_____

关于学位论文使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用学位论文的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本学位论文。本人离校后发表、使用学位论文或与该论文直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本学位论文研究内容：

□可以公开

□不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

基于贝叶斯方法进行中医证候诊断的研究

摘要

证候和临床症状，是中医诊断的临床基点，具有表现多样，相互之间关系复杂的特点，学习掌握困难。机器学习提供了一种新的可能，而贝叶斯方法是机器学习的一个重要分支，是基于贝叶斯定理的统计学分类方法，具有严谨的数学理论作支撑。通过使用贝叶斯方法对证候和临床症状进行分析识别，便于深入把握其内在联系，为中医理论的学习掌握和临床研究给出有益探索。

本文重点研究贝叶斯方法中的两个重要算法朴素贝叶斯分类算法和贝叶斯网络在中医诊断中的应用。主要研究工作如下：

1）从《中医诊断学》中以人工方法手动总结、整理中医诊断的相关数据，应用于贝叶斯方法。

2）将朴素贝叶斯分类算法应用于中医诊断中，使用 C#语言开发窗体应用程序进行中医诊断中的证候分类研究，实现了分类预测、计算预测准确率等功能，并添加了汉字拼音首字母提示输入功能。

3）突破朴素贝叶斯分类的条件独立性假设，提出利用信息论中条件互信息的概念构造中医诊断的网络结构模型的方法。

论文的初步成果是，使用朴素贝叶斯分类算法的思想开发中医证候诊断的分类器，实现对临床症状的证候进行分类，分类准确率可到 98%以上。采用图论与信息论的知识构造贝叶斯网络结构图，对于更深入的分析临床症状之间以及临床症状与证候之间的依赖关系有指示意义。

关键词：朴素贝叶斯分类，贝叶斯网络，中医诊断，证候，临床症状

Based on Bayesian Method to Research the Diagnosis of Syndromes in TCM

Abstract

Syndrome and clinical symptoms, are the clinical basis of diagnosis of Chinese medicine, with a variety of performance, the relationship complex, learning difficulties. Machine learning provides a new possibility, and Bayesian method is an important machine learning branch, that based on the Bayesian theorem's statistical classification method, with rigorous mathematical theory as a support. Through using the Bayesian method to analysis and identification the syndrome and clinical symptoms, easy to grasp the intrinsic link, for the study of TCM theory and clinical research gives useful tips.

This dissertation focuses on two important algorithms about Bayesian method: naive Bayesian classification algorithm and Bayesian network. The main work of this dissertation as follows:

1) Collect and summarize the TCM data from the "Chinese medicine diagnosis" in artificial way, applied to the Bayesian method.

2) The naive Bayesian classification algorithm was applied to analysis of traditional Chinese medicine, and using the C# language to develop a form application program. Implements the classification and calculate the prediction accuracy, and add the Chinese characters' pinyin first letter prompts for input.

3) Break through the conditional independence hypothesis of naive Bayesian classification, construct the network structure model of TCM diagnosis by using the concept of conditional information in information theory.

The preliminary results of this dissertation are the use of naive Bayesian classification algorithm to develop TCM syndrome diagnosis classifier to achieve the classification of clinical symptoms of syndrome, classification accuracy rate can be more than 98%. The Bayesian network structure is constructed by the theory of graph theory and information theory. It is instructive to analyze the relationship between clinical symptoms and clinical syndrome.

Key words: Naive Bayesian classification, Bayesian Network, Diagnosis of TCM, Syndrome, Clinical Symptoms

目 录

第一章　绪论

1.1 课题研究背景

改革开放 30 多年以来，人民的生活条件和生活环境日新月异，近年来又随着雾霾的加重，生活环境开始恶化，人类表现出多种疾病现象，临床症状表现繁杂，疾病成因复杂，同时伴随多种并发症，对我国现有的临床疾病诊断模式和疾病预防体系都带来了不小冲击。中医学作为华夏土地上的医理典范，承载的是中华儿女在同疾病作斗争过程中积累的经验知识，在长期医疗实践中不断发展并逐步形成一门健全的医学特定的理论体系[1]。

疾病症状（Disease Symptoms）和证候（Syndrome）长期以来都是中医诊断学的研究热点[2]，特别是在与数据挖掘技术、人工智能技术和统计学等多学科交叉结合之后，对于证候的研究又取得了重大进展。近年来，随着统计推理和决策理论的发展产生了以信息收集、信息处理、做出诊断为主要过程的中医诊断方法。目前建立智能化的中医诊断系统已成为研究症状和证候的热点方向之一[3]。首先就是明确病位即临床症状和病性即证候的辨证要素，再利用辨证的方法判断临床症状和证候之间的影响关系。在确定症状的证候属性时，朴素贝叶斯分类方法具有与人脑相类似的思维诊断模式，加之与中医辨证诊断过程的特定的吻合性，逐渐成为研究中医证候的基础方法。特别是在获取的信息不完备、不精准时，需要借助柔性的方法来进行处理分析，贝叶斯网络对于解决复杂系统中由于不确定性的因素而引起的故障具有特殊的优势，尤其是其结合了图论和概率论知识之后，成为能够方便地表达和分析不确定性事物的模型。

1.2 课题研究目的及意义

数据挖掘也被称为数据库知识发现（Knowledge Discovery in Database, KDD），是随着数据库技术和人工智能技术的交叉融合发展兴起的一门专业学科，是近十年来使用广泛的一种重要方法，一般指借助计算机从大量数据中利用算法搜索进而发现隐藏于其中的有价值的知识和信息的复杂过程。生物医学数据挖掘则是数据挖掘的一个重要应用，包括数据的收集、整理、分析和应用等多方面，涉及数据库、统计学、机器学习、数学等多个领域[4]。

目前人体已经表现出来且被整理命名过的证候类型约有 125 种，共涉及临床症状类型约有 1084 种，通过结合人体和中医理论将不同的临床症状表现归属到某些特定的证候中，这对于研究证候和临床症状之间的关系及后续的疾病诊治具有重要的指导性意义。但是不同证候的临床症状之间往往具有交叉重复，在实际的诊治过程中，会根据患者表现出的临床症状集合，来推断其所属的证候类型，进而做出进一步的治疗，推断一组临床症状的证候类型的过程就是使用了分类的思想，在统计学中可用概率表达。贝叶斯方法将统计学中的概率的概念解释为信念度（Degree of Belief），是会随着观测数据的更新而不断被更新的[5]。此外，通过将事件的先验知识作为重要前提信息纳入研究，在使用贝叶斯方法时结合新获取的数据信息，更新先验知识到后验分布中，实现了对数据的充分利用[6]。

当下比较常见的中医诊断主要依赖于医生的临床经验和医术水平，随着在中医疾病治疗的临床实践发展，积累了大量的临床数据，如果能从现已积累的临床数据中提取特征信息和有用数据，结合中医辨证论治的思想，利用数据挖掘方法研究中医诊断系统，这将对中医证候的辨证施治具有极大意义。

利用贝叶斯方法进行中医诊断的证候的辨识，对中医诊断学的发展与更广泛的应用意义非凡[7]。本文以贝叶斯方法为背景，具体研究朴素贝叶斯分类算法和贝叶斯网络在中医诊断中的应用，通过实验证明其准确性和高效性。将前人的治疗经验进行采集记录，建立中医诊断的数据库，为临床医学研究和数据分析提供数据源。不仅能够发挥中医诊断的独特优势，而且对于探索中医诊断的现代化之路也有积极意义。将数据挖掘技术引入中医诊断学中，结合二者优势，对中医临床症状和症候表现进行记录，建立中医诊断信息系统数据库，结合数据挖掘、数据分析的方法，从海量数据中分析并提取出有价值的知识信息，以此来指导医生的临床实践，提高诊治的准确率，为后续的治疗和用药提供理论支持。

1.3 国内外研究现状

传统的中医诊断可以溯源至春秋战国时期的名医扁鹊（公元 407-公元前 310 年）在总结前人的经验之上，提出的“望”、“闻”、“问”、“切”四诊法，成为后代中医治疗疾病的先锋指导[8]。

进入 20 世纪之后，伴随着信息技术和计算机应用的兴起，人工智能和数据挖掘技术的长足发展，使得中医诊断的智能化应用层出不穷。到 20 世纪 70 年代先后出现的结合了专家系统知识[9]的中医诊断系统，到 80 年代末，自称为“信息革命的促进派”的生物医学专家秦笃烈先生，主编《中医计算机模拟及专家系统概论》一书，书中涉及 3 类极典型的专家系统，第一类：只涉及单一病域的中

医诊断专家系统；第二类：逐渐向整体思维靠拢，如袁冰等人设计的“董建华热病诊疗系统”；第三类：是采用智能化的概念而打造的智能化中医诊断辨证论治系统，如朱文锋在 1985 年研究出的“中医辨证论治电脑系统”[10]。

20 世纪 90 年代后期，一种模拟生物神经系统的数学处理方法——人工神经网络[11]，因其对于复杂映射关系具有超强的逼近能力，为研究建立中医证候的非解析模型提供了可靠的依据。随后朱文锋等人又将贝叶斯方法运用于中医辨证系统中，采用中医的证候、证素、证名等并结合大量临床实践数据而建立起的中医辨证数据库[12]，在进行数理统计与计量分析、逻辑推理验证证候等方面取得了与临床中医专家经验吻合度极高的结果[13]。

英国的一位名叫 Thomas Bayes 的牧师在 1736 年提出了贝叶斯定理，其遗作《An essay towards solving a problem in the doctrine of chances》[14]在 1763 年由其朋友 Richard Price 整理发表，贝叶斯理论的学术价值也被世人知晓[15]，之后法国数学家 Laplace 在前人的基础上，对贝叶斯理论作了进一步的证明，并在解决天体力学问题和医学统计研究中的问题中使用该方法[16]。

20 世纪初，意大利和英国的学者对贝叶斯理论的发展做出新的重要贡献[17]。二战之后，瓦尔德（Wald A.）提出的基于统计的决策理论中[18]，贝叶斯理论起着极其重要的作用。信息论的发展也对贝叶斯学派的研究做出重要贡献。英国著名的生物行业杂志《BioMetrika》在其子行业“统计学与概率学”中重新以全文形式刊登了贝叶斯的论文；到了 20 世纪 50 年代，以罗宾斯（Robbins H.）为主的学者，提出把经验贝叶斯方法与经典方法相结合的思想[19]，引起了统计学界的广泛关注，并很快显示出了其优点，成为热门的研究对象。

20 世纪初期，美国知名遗传学家 Sewall Wright 提出了有向无环图（Directed Acyclic Graphical, DAG）的因果表示模型，随后被用于心理学、经济学、社会学等领域[20]。20 世纪中叶，提出一个新的概念——决策树，用来表达决策分析的相关问题，形成了完整的决策分析理论[21]，1988 年 Pearl 通过对前人的工作进行总结，提出了新的网络——贝叶斯网络[22-24]，至 90 年代，有效的推理预测算法和网络学习算法的出现进一步推动了贝叶斯网络的发展[25、26]。

国内对于贝叶斯网络的研究兴起于近几年，以清华大学林士敏为首的通过剖析贝叶斯网络的构造过程，探索从先验信息和样本数据中确定网络结构模型的方法[27]。从 21 世纪初开始，国家自然科学基金资助的项目有关贝叶斯方法的项目逐渐增多。在应用方面，霍利民等提出的使用贝叶斯网络实现配电系统可靠分析的方法，取得了较好的效果[28]；李俭川等将贝叶斯网络应用在复杂设备诊断中存在的不确定性和关联性问题，以 SINS/GPS 组合导航系统用于解决复杂系统的故

障诊断[29]。许文杰等将贝叶斯网络用于主观性和模糊性较大的中医辨证过程，为中医辨证的规范化研究提供平台[30]。

1.4 论文的组织结构

本篇论文共划分六章，结构安排如下：

第一章：绪论。本章介绍了本文研究的目的和意义，并分析中医诊断学和贝叶斯方法的研究现状。

第二章：中医的相关知识介绍。本章主要介绍的是中医诊断的部分理论知识，并给出简单的中医诊断流程模型。

第三章：贝叶斯方法。本章主要介绍贝叶斯方法中的概率论相关知识，再分别介绍朴素贝叶斯分类和贝叶斯网络，并举例介绍。

第四章：基于朴素贝叶斯分类的中医证候诊断研究。本章详细介绍朴素贝叶斯分类算法在中医诊断中的具体应用，包括数据处理、开发应用程序、分类预测、计算准确率等内容。

第五章：基于贝叶斯网络的中医诊断研究。本章重点介绍基于互信息构建证候的贝叶斯网络结构图。

第六章：总结与展望。本章对全篇的所有工作进行总结概述，同时明确进一步的工作方向。

第二章　中医的相关知识介绍

中医是相对于西医而言的医学理论，中医理论主要是指建立在阴阳五行说、藏象学说、经络学说等基础上，通过“望”、“闻”、“问”、“切”进行辨证施治，采用中药各种剂型或结合针灸，拔罐，推拿，按摩等手段对患者疾病进行诊断治疗的一套完整体系[31]。

中医学认为，人体所表现出的症状与五脏六腑等各个部位，往往存在着相互作用相互影响的联系，人体的各个部分不是独立的，而是一个完整有机体。因此，疾病变化的病理本质虽然藏之于“内”，但必然会有一些的症状或体征反映于“外”，局部的病变表现往往折射的是整体的病理状况，整体的病变反应又是通过身体的多方面表现出来的。通过考察人体反映于“外”的不同类别的病症现象，在医学知识的指导下通过概括分析、对比思考，可以认识疾病的本质。

2.1 临床症状与证候

进行中医诊断的基础依据是临床症状，也叫临床表现，是医学中的一个概念，表示疾病发病时人体所表现出来的异常现象。所谓“临床”多指直面病患者来参与疾病的诊断治疗的方法。通过“临床”的过程获取疾病的异常表现[32]，如头痛、胸闷、腹胀等可以描述或感知的症状，如面色晄白、大便腥臭、脉浮数等可以客观检测出来的一些征象。

证候亦简称“证”，在中医学的历史上，对于“证”的解释和使用不太统一，有以“证”为“症状”者，亦有称“病”为“证”者。现代中医约定：“证”是对发作过程中的疾病，所处某一阶段时的病位、病因、病性及病势等所作的病理性概括，反映的是疾病的本质[33]。

需要注意，临床症状中的“症”和证候中的“证”，应该严格区分，前者是人体表现出来一个一个的症状，而后者则是经过辨证得到的有助于疾病治疗的结果[34]。

2.2 中医诊断

中医诊断学，是从古至今历代医者名家进行的临床诊病经验的积累，其理论和方法最早起源于公园前五十几和名医扁鹊所提出的“切脉”、“望色”、“听声”、

“写形”等方法。中医诊断学包含的内容有：诊法、诊病、辨证、病历等内容。

诊法，是指中医中诊察、收集病情资料的基本方法，有“望”、“闻”、“问”、“切”四诊。通过“四诊”可以收集到的病情资料主要有症状、体征和病史。诊病也叫辨病，就是在中医理论的指导下，通过分析四诊资料，对人体的不良表现作出诊断、得出证名的过程[35]。“辨证”是结合中医理论和中国古代哲学思想，采用辨证思维的方法综合分析疾病，并对其本质作出判断的过程。病历，也叫病案，古时候称作诊籍，是临床上对于病患者的病情和诊治过程的详实记录，随着科学技术的发展目前也出现了电子病历。中医诊断坚持的是司外揣内、见微知著、以常衡变的基本原理和整体审查、诊法合参、病症结合的基本原则[36]。

2.3 中医辨证论治

辨证论治是在中医上认识疾病并进行疾病治疗所坚持的基本指导原则，是中医学上研究疾病的一套策略，也称辨证施治。包括辨证和论治两个过程。首先辨证就是认证识证的过程，通过该过程可以综合分析疾病的证候特征，确定疾病的病因、病位等信息。其次，论治即施治，根据辨证的结果，指导医者选择正确高效的治疗方法，对疾病进行治疗。

临床应用中常见的辨证施治的方法大概有以下六种：（1）八纲辨证，它是辨证的纲领，即纲领证；（2）病性辨证，旨在是分辨每种证候的性质，即基础证；（3）脏腑辨证，是以辨别病位为目的的辨证方法，故属具体证；（4）六经辨证，是《伤寒论》中关于辨证论治的纲领，是东汉名医张仲景在《素问 热论》的基础上，依据伤寒病的证候特点和传变规律而总结得出的辨证方法；（5）卫气营血辨证，是清代温病学家叶天士在《外感温热篇》中所提，是一种专门针对外感温热病的辨证方法；（6）三焦辨证，是清代另一位温病学家吴鞠通在《温病条辨》中，对外感温热病做出辨证归纳时所用的一种方法，根据《黄帝内经》中对于三焦所属不同部位的解释，借以阐释三焦所属脏腑在温热病发展的不同阶段的变化规律。

2.4 中医诊断的模型流程图

现代的中医诊断可以分为两个部分，第一部分是建立中医诊断数据库，这一过程首先根据已知典型病例如疾病的类型，收集不同类型疾病的症状表现，对数据进行整理，构建数据库，这一过程需要大量的临床医学治疗经验，是一个费时费力并且对后续的应用研究具有极大影响的过程；第二部分是进行诊断，参照前

一步所建立的数据库系统，分析给定病患，对病患的症状表现进行分析提取，依据一定的方法做出诊断。中医诊断模型图如图 2-1 所示。

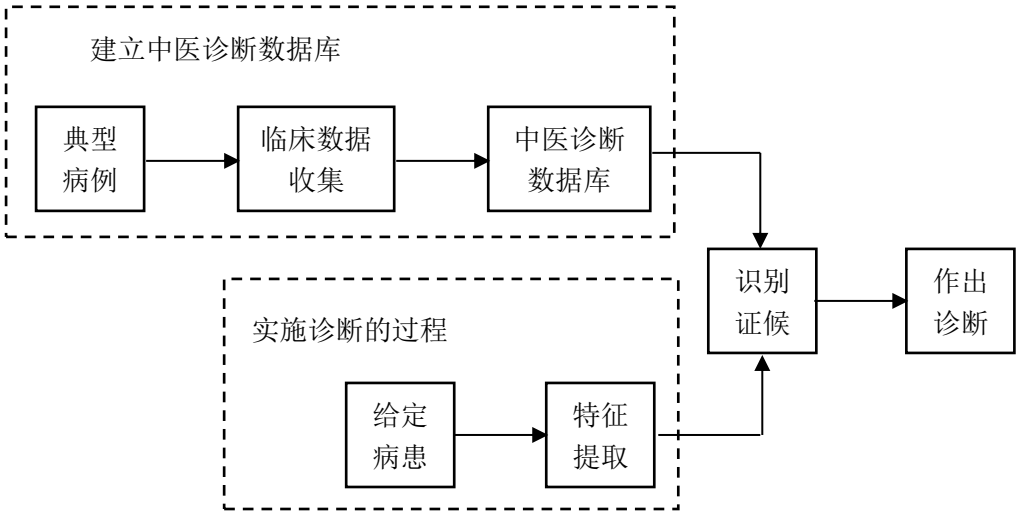


图 2-1　中医诊断模型图

第三章　贝叶斯方法

3.1 贝叶斯方法中相关的概率论知识

3.1.1 先验概率和后验概率

先验概率（Prior probability）表示在对样本观测前依据以往经验来分析各个事件发生的概率知识，通常是作为“执因求果”问题中的“因”出现[37]。可以将先验概率分为两类，一类是可以利用已拥有的历史信息通过计算而得到的先验概率，被称为客观先验概率；另一类是在历史资料无处取得或者资料获取不完整的情况下，借助相关经验人的主观经验判断而得到的先验概率，也被称为主观先验概率[38]。

后验概率（Posteriori probability）是在进行了新的观测之后对原有知识的更新，是可以使用贝叶斯公式对先验概率进行修正之后得到的概率，可看作是“由果觅因”问题中的“因”[39]。

3.1.2 条件概率和联合概率

条件概率（Conditional probability）可以定义为事件 A 在另外一个事件 B 已经确定发生的前提条件下发生的概率，表示为 $P(A \mid B)$[40]。对于只涉及两个事件 A 和 B 的情况，条件概率可用如下形式表示：

$$P(A \mid B) = \frac{P(AB)}{P(B)} \qquad (P(B) > 0)$$

当 $P(B)$ 和 $P(A \mid B)$ 已知时，可以反求 $P(AB)$，得：

$$P(AB) = P(B)P(A \mid B) \qquad (1)$$

同理，由如下条件概率表示形式：

$$P(B \mid A) = \frac{P(AB)}{P(A)} \qquad (P(A) > 0)$$

当 $P(A)$ 和 $P(B \mid A)$ 已知时，亦可以反求 $P(AB)$，得：

$$P(AB) = P(A)P(B \mid A) \qquad (2)$$

公式（1）和（2）都被称为乘法公式[41]，利用它们可以计算两个事件同时发生的概率。

乘法公式也可以扩展到多个事件的积的情况。设有三个事件 A、B、C，则这三个事件的积事件可以表示为：

$$P(ABC) = P(C|AB)P(B|A)P(A)$$

联合概率（Joint probability）[42]表示两个或多个事件同时发生的概率。

3.1.3 全概率公式和贝叶斯定理

如果事件组$\{B_1,B_2,...,B_n\}$满足：（1）$B_i \cap B_j = \emptyset$，$i \neq j$，$(i,j=1,...,n)$，且$P(B_i)>0, i=1,...,n$；（2）$B_1 \cup B_2 \cup ... \cup B_n = \Omega$，则称事件 B1,B2,...是样本空间 Ω 的一个划分。对于任意一个事件 B，有如下全概率公式：

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

利用全概率公式可以将一个复杂事件系统的概率计算问题，分解为有限的若干个简单事件的概率计算问题，通过计算简单事件的概率，然后相加来计算复杂事件的概率。

与全概率公式解决的问题不同的是，贝叶斯定理是在条件独立的基础上寻找导致事件发生的原因的概率。设$B_1,B_2,...,B_n$是样本空间 Ω 的 n 个独立子集，对于样本空间中的任意一个事件 A（$P(A)>0$），则有：

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

上式称为贝叶斯公式，其中把$\sum_{i=1}^{n} P(A|B_i)P(B_i)$称为先验概率，把$P(B_i|A)$称作后验概率。

事件 A 在事件 B 发生的前提条件下发生的概率，与事件 B 在事件 A 发生的前提条件下发生的概率，尽管不是同一个概念，但是两者之间却存在特定联系，贝叶斯定理正是对这种关系做出了解释。简单来说，贝叶斯定理是基于假设的先验概率和在给定的假设下观测概率，提供一种后验概率的计算方法[43]。

贝叶斯定理是由 18 世纪英国牧师 Thomas Bayes 提出的，Thomas Bayes 是早期概率论和决策论的先驱研究者。贝叶斯公式发表于其去世之后的 1763 年，首次在概率论基础理论中引入归纳推理法则，对于后续研究统计决策、概率推理和参数估计等领域起到了重要的推进作用，其影响延续至今，信息时代的经济学理论、数据处理与知识挖掘、信息检索、人工智能等诸多方面都有深入和广泛的应用。

3.2 朴素贝叶斯分类

3.2.1 分类问题概述

日常生活中，对于分类问题，我们都不会陌生。例如，当你身旁走过一个人时，大脑会不自觉的判断 TA 是男是女；和朋友走在路上时，你可能会对身旁的朋友说“刚刚过去的那个人长得好帅，那边一个身体健康的老太太”等等之类的话，这些都算是一种分类的活动。

从数学角度来说，分类问题可以如下方式定义：

已知集合：$C=\{y_1,y_2,...y_n\}$和$I=\{x_1,x_2,...x_m,...\}$，确定映射规则$y=f(x)$，使得任意$x_i\in I$有且仅有一个$y_i\in C$使得$y_i=f(x_i)$成立。其中，C 称为类别集合，C 中的每个元素都是一个类别，I 称为属性集合，I 中的每个元素都是一个待分类项，映射规则 f 被称作分类器。分类的过程就是构造分类器 f，并使用分类器进行分类操作。

例如，医生对病人的病情进行诊断的过程就是一个典型的分类过程，在此过程中，需要观察病人身体的异常表现、不良反应及各种化验检测等数据信息来推测病情，这时的医生就充当了一个分类器的功能，而医生诊断疾病的准确率，与他所受到的医学知识（构造方法）、病人的症状是否明显（待分类数据的特性）以及医生临床症状经验的多少（训练样本数量）都有紧密联系[44、45]。

3.2.2 朴素贝叶斯分类算法简介

基于贝叶斯定理的贝叶斯分类算法是一类简单实用且高效的分类算法，有严谨的数学理论作支撑。在假设待分类项的各个属性相互独立的前提下，构造出来的分类算法就是朴素的，称为朴素贝叶斯分类算法（Naive Bayesian classification，NB），是简单高效的一种贝叶斯分类方法[46]。该算法的思想基础是：对于待分类的属性集合，分别计算属性集合在各个类别条件下发生的概率，把概率值最大的那个类别看作是待分类的属性集合的类别[47]。可简单理解为，下述情况，如果有人问你，水是什么状态的？你很可能会回答液态。因为我们在生活的环境中，多数情况下水是以液态形式存在的，当然水也有固态和气态的形式。但是在没有更多的前提条件的情况下，我们通常会选择可能性最大的那种情况。

朴素贝叶斯分类在辅助智能决策、数据融合、模式识别、医疗诊断、文本理解、地震预报、项目招投标等方面有重要应用。

3.2.3 朴素贝叶斯分类算法的分类流程

朴素贝叶斯分类的路线图如图 3-1 所示，该图清晰的展示了分类的三个阶段。

第一阶段——准备工作阶段，对即将展开的分类任务做必要的准备工作。具体工作是，针对具体待分类问题的性质和特点确定特征属性和类别属性。获取训练样本，通常采用人工的方法收集并整理数据，准备工作的质量对后续的分类具有较大影响，分类器的分类性能的高低很大程度上取决于特征属性的选择、特征属性的划分取值。

第二阶段——分类器训练阶段，这一阶段的首要任务就是构造分类器，具体包括通过统计每个类别属性在训练样本中的计数来计算其概率，估计不同特征属性划分在每个类别属性下的条件概率，并将结果记录。这一阶段输入的是训练样本、特征属性及其划分，输出分类器。这一阶段可以通过编写计算机程序来实现，可根据前面讨论的贝叶斯理论知识编写代码利用计算机来计算完成。

第三阶段——分类器应用阶段，主要任务是利用分类器对待分类项进行分类操作，一般是输入待分类项，分类器会将待分类项的所属类别输出出来。这一阶段也是通过编写程序来实现的。

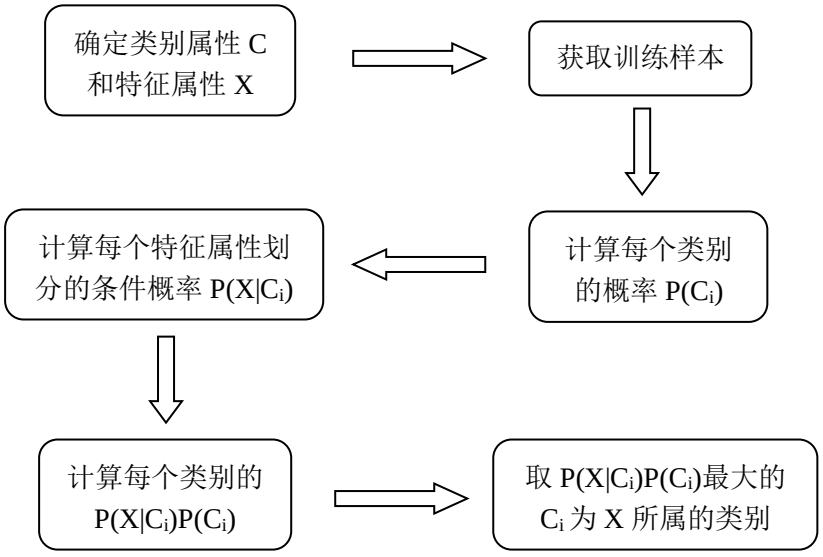


图 3-1　朴素贝叶斯分类路线图

3.2.4 Laplace Smoothing 校准

Laplace Smoothing 又称为加 1 平滑，是比较常用的校准方法，由法国数学家拉普拉斯（Pierre-Simon Laplace, 1749－1827）最先提出使用加 1 的策略估计特定情况下没出现过的事件的概率[48]，目的是为了解决零概率事件问题。零概率问题，就是在计算变量的概率时，若特征属性 X 在某个类别属性 C 下没有出现（即计数为 0）时，会导致该特征属性的概率值计算结果为 0，加之在后续的计算中因为要使用乘法法则计算结果概率，就会使得整个概率值为 0。出现这种计算结

果显然是不合常理的，即不能因为某一个特征项的计数为 0 就断定该整个集合的概率值为 0。

Laplace Smoothing 校准在解决零概率问题时的一般计算方法如下：对于一个随机变量 W，取值范围是$\{1,2,3,\ldots,k\}$，在进行 n 次实验后观测结果值分别为$\{W^{(1)},W^{(2)},W^{(3)},\ldots,W^{(n)}\}$，不使用 Laplace Smoothing 校准时计算不同取值的概率的计算公式为：

$$P_j=\frac{\sum_{j=1}^{n}C\{W^{(j)}=i\}}{n}\qquad(j=1,2,\ldots,k;i=1,2,\ldots,n)$$

使用 Laplace Smoothing 校准之后，计算公式为：

$$P_j=\frac{\sum_{j=1}^{n}C\{W^{(j)}=i\}+1}{n+k}\qquad(j=1,2,\ldots,k;i=1,2,\ldots,n)$$

即在分母上加上取值范围的大小，同时在分子上需加 1。

Laplace Smoothing 校准的应用举例，设某训练数据集 D 上中，类别变量 C 取值为 Yes 时，特征项 T 的三种不同取值 Hot、Cool、Mild 的计数分别为：0、10、990。不使用 Laplace Smoothing 校准时，概率的计算如下：

$$P(T=Hot\mid C=Yes)=\frac{0}{1000}$$
$$P(T=Cool\mid C=Yes)=\frac{10}{1000}$$
$$P(T=Mild\mid C=Yes)=\frac{990}{1000}$$

因为出现概率为 0 的情况，故需要使用 Laplace Smoothing 校准进行修正，于是概率值的计算方法和结果分别为：

$$P(T=Hot\mid C=Yes)=\frac{0+1}{1000+3}=\frac{1}{1003}$$
$$P(T=Cool\mid C=Yes)=\frac{10+1}{1000+3}=\frac{11}{1003}$$
$$P(T=Mild\mid C=Yes)=\frac{990+1}{1000+3}=\frac{991}{1003}$$

可见，使用 Laplace Smoothing 校准后，改善了概率为 0 的情况，同时也能保证概率值的变化不大。

Laplace Smoothing 校准简单明晰，就是对每个特征属性在各个类别下的计数均加 1，这样当训练样本集的量级足够大时，便不会对最终的结果产生影响，同时也解决了概率为 0 的问题[49]。另外，在实际使用中有时会采用加λ（$0<\lambda\leq1$）来代替直接加 1。这时如果对 N 个划分的计数都加上λ，则分母就要加上 N*λ。

3.2.5 朴素贝叶斯分类算法的应用举例

下面以一个两分类的任务，根据天气条件判断是否可以打球来解释朴素贝叶斯分类算法在分类当中的应用。

Step 1. 确定特征属性及划分取值

这里取四种特征属性：Outlook、Temperature、Humidity、Windy 分别表示天气情况、温度情况、湿度情况、风力情况。给出特征属性的取值，属性 Outlook 有 Sunny、Overcast、Rain 三个取值；Temperature 有 Hot、Mild、Cool 三个取值；Humidity 有 High 和 Normal 两个取值；Windy 有 Weak 和 Strong 两个取值。

根据朴素贝叶斯分类算法应用的前提，假设 Outlook、Temperature、Humidity，Windy 这四个特征属性之间完全相互独立，即这四个属性对于是否打球的影响是独立的。

Step 2. 获取训练集

这里，选取了 14 条数据作为训练样本，如表 3-1 所示。

表 3-1　14 条数据的训练样本

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Step 3. 计算训练样本中不同类别的概率

统计训练样本中“打球（Yes）”和“不打球（No）”的数量，得到计数分别为 9 和 5，再除以训练样本总数 14，得到类别概率如下：

$$P(Yes) = \frac{9}{14} \qquad P(No) = \frac{5}{14}$$

Step 4. 计算每个类别下各个特征属性划分的概率

首先根据训练样本，统计各个特征划分的计数，为了清晰查看计算条件概率值，将统计结果先以表的形式展示。表 3-2、表 3-3、表 3-4、表 3-5 分别是依据样本数据统计得到的计数结果，由于出现计数为 0 的情况，会在后面计算概率时，会导致概率值为 0，因此需要使用 Laplace Smoothing 校准，使用 Laplace Smoothing 校准之后的计数如表 3-6 所示。

表 3-2　Outlook 不同取值的计数表

Outlook	PlayTennis=Yes	PlayTennis=No
Sunny	2	3
Overcast	4	0
Rainy	3	2

表 3-3　Temperature 不同取值的计数表

Temperature	PlayTennis=Yes	PlayTennis=No
Hot	2	2
Mild	4	2
Cool	3	1

表 3-4　Humidity 不同取值的计数表

Humidity	PlayTennis=Yes	PlayTennis=No
High	3	4
Normal	6	1

表 3-5　Windy 不同取值的计数表

Windy	PlayTennis=Yes	PlayTennis=No
Weak	6	2
Strong	3	3

表 3-6　校准后的 Outlook 不同取值的计数

Outlook	PlayTennis=Yes	PlayTennis=No
Sunny	3	4
Overcast	5	1
Rainy	4	3

可以依据上述表中的统计信息来计算各个特征属性的条件概率，部分类条件

概率计算如下：

$$P(Outlook = Sunny \mid Yes) = \frac{3}{12} \quad P(Outlook = Sunny \mid No) = \frac{4}{8}$$

$$P(Tremprature = Cool \mid Yes) = \frac{3}{9} \quad P(Temprature = Cool \mid No) = \frac{1}{5}$$

$$P(Humidity = High \mid Yes) = \frac{3}{9} \quad P(Humidity = High \mid No) = \frac{4}{5}$$

$$P(Wind = Strong \mid Yes) = \frac{3}{9} \quad P(Wind = Strong \mid No) = \frac{3}{5}$$

Step 5. 对一组新实例进行分类，判断其是属于打球还是不打球。

待分类实例为：

$x = (Outlook = Sunny; Temprature = Cool; Humidity = High; Wind = Strong)$，分别计算属于类别 Yes 和类别 No 时的概率。

$$P(Yes \mid x) \propto P(x \mid Yes) * P(Yes)$$

$$= P(Outlook = Sunny \mid Yes) * P(Temprature = Cool \mid Yes)$$

$$* P(Humidity = High \mid Yes) * P(Wind = Strong \mid Yes) * P(Yes)$$

$$= \frac{3}{12} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} * \frac{9}{14} = \frac{1}{168}$$

$$P(No \mid x) \propto P(x \mid No) * P(No)$$

$$= P(Outlook = Sunny \mid No) * P(Temprature = Cool \mid No)$$

$$* P(Humidity = High \mid No) * P(Wind = Strong \mid No) * P(No)$$

$$= \frac{4}{8} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} * \frac{5}{14} = \frac{3}{175}$$

Step 6. 对比不同类别下的计算结果，确定待分类实例的类别。

显然 $P(No \mid x) > P(Yes \mid x)$，所以该实例属于类别 No，即不打球。

3.3 贝叶斯网络

朴素贝叶斯分类在应用时需要假定特征属性之间是条件独立的，但这种假定与现实事例不符。通常，影响一个事物的各个属性之间会存在一定的依赖关系，贝叶斯网络考虑了这种依赖关系，因此在处理属性有依赖关系的样本时具有优势，是一种使用类条件概率建模的方法。

3.3.1 贝叶斯网络的定义和性质

贝叶斯网络（Bayesian Network，BN），又称信念网络（Belief Network，BN），是用于描述变量之间相互依赖联系的概率网络图模型，是概率论与图论相结合的产物，常用于分析复杂系统的影响因素之间的关系。

贝叶斯网络包含定性层面和定量层面两方面的内容：

定性层面：贝叶斯网络借助有向无环图（Directed Acyclic Graphical, DAG）来表达随机变量之间的依赖关系，其中节点代表随机变量（Random variables），节点之间的有向边来表示节点间的依赖关系，依赖关系有时也称为具有因果关系，或条件依赖（Conditional dependencies）。在从节点 i 指向节点 j 的有向边，其中 i 称作 j 的父节点，j 称作 i 的子节点。构造有向无环图的过程是贝叶斯网络的结构学习。

定量层面：贝叶斯网络中的每个节点都有一个条件概率表（Conditional Probability Table, CPT），用于定量展示与该节点有直接相关的节点之间的概率分布。如果节点 X 无父节点，则称该节点的概率 P(X) 为先验概率；如果节点 X 有一个或多个父节点 $Y = \{Y_1, Y_2, \ldots, Y_n\}$ $(n \geq 1)$，则 X 在父节点 Y 下的概率可以表示为：

$$P(X \mid Y) = P(X \mid Y_1, Y_2, ..., Y_n)$$

简言之，可将贝叶斯网络定义为 $B = (G, T)$，G 是定性层面的有向无环图结构，T 则是定量层面的条件概率表。

贝叶斯网络的一条重要的性质，就是我们断言每一个节点在其直接前驱节点（父节点）的值给定后，这个节点条件独立于其所有非直接前驱节点[50]。

3.3.2 贝叶斯网络的三种结构形式

贝叶斯网络在结构上是一个有向无环图，图中的节点和边之间的关系通常有三种基本的结构形式。

第一种结构形式：head-to-head，也称为 V 型结构（V-structure）、“冲撞”结构、收敛依靠联系，形式如图 3-2 所示，这一结构形式表示节点 c 有两个父节点 a、b，而 a、b 之间无依赖关系。

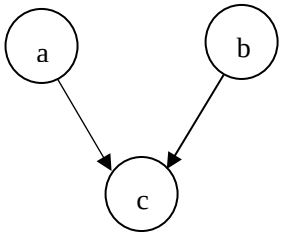


图 3-2　V 型结构示意图

在 c 已知时，a，b 不相互独立。在 c 未知的条件下，a，b 被阻断（Blocked），是独立的，称之为 head-to-head 条件独立。

该结构的联合概率可表示为：

$$P(a,b,c) = P(a)P(b)P(c \mid a,b)\text{。}$$

第二种结构形式：tail-to-tail，也称为“同父”结构（common parent）、发散依靠联系，形式如图 3-3 所示：

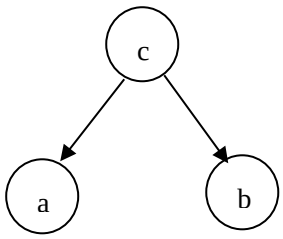


图 3-3　“同父”结构示意图

在 c 未知时，有：

$$P(a,b,c) = P(c)P(a \mid c)P(b \mid c)\text{，}$$

此时无法得出 $P(a,b) = P(a)P(b)$，即 C 未知时，a、b 之间的独立性不确定。在 c 已知时，有：

$$P(a,b \mid c) = \frac{P(a,b,c)}{P(c)}\text{，}$$

将 $P(a,b,c) = P(c)P(a \mid c)P(b \mid c)$ 带入上式中，得：

$$P(a,b \mid c) = \frac{P(a,b,c)}{P(c)} = \frac{P(c)P(a \mid c)P(b \mid c)}{P(c)} = P(a \mid c)P(b \mid c)\text{，}$$

可知在 c 已知时，a、b 相互独立。

所以，在 c 给定的条件下，a，b 被阻断，是相互独立的，称之为 tail-to- tail 条件独立。

第三种结构形式：head-to-tail，称为“顺序”结构、也叫序列依靠联系，形式结构如图 3-4 所示：

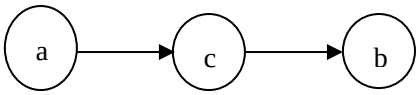


图 3-4　“顺序”结构示意图

c 未知时，有：$P(a,b,c) = P(a)P(c \mid a)P(b \mid c)$，但无法得出 $P(a,b) = P(a)P(b)$，

即 c 未知时，a，b 不相互独立。c 已知时，有：

$$P(a,b\mid c)=\frac{P(a,b,c)}{P(c)},$$

且由 $P(a,c)=P(a)P(c\mid a)=P(c)P(a\mid c)$，可得：

$$P(a,b\mid c)=\frac{P(a,b,c)}{P(c)}=\frac{P(a)P(c\mid a)P(b\mid c)}{P(c)}=\frac{P(c)P(a\mid c)P(b\mid c)}{P(c)}=P(a\mid c)P(b\mid c)。$$

所以在 c 给定的条件下，a，b 被阻断，是独立的，称之为 head-to-tail 条件独立。

第四章　朴素贝叶斯分类在中医证候诊断中的应用

将朴素贝叶斯分类应用于中医诊断中的优势：

1）朴素贝叶斯分类以贝叶斯公式为基础，具有坚定的数学和统计学理论支撑，以概率作为分类的标准，使得分类过程简单清晰，分类结果具有较强的说服力[51]。

2）朴素贝叶斯分类具有成熟严格的分类推理过程，基于条件独立性的假设可以简化变量的复杂关系，利用统计的知识获得变量的分布情况和概率值，从而计算待分类项在不同类别下的值，取值最大的类别认为是待分类项所属类别[52]。

3）朴素贝叶斯分类分类过程时空开销小，算法稳定，对于适用于临床症状较多的中医诊断，对不同的数据特点其分类性能差别不大，健壮性好。

本章就是使用朴素贝叶斯分类思想对一系列未知证候类型的临床症状样本的进行分类预测，并计算预测准确率。

实验前假设：

1）临床症状之间是相互独立的，即认为临床症状“咳嗽”和临床症状“咽干”无直接依赖关系。

2）认为各个临床症状的权值都是相同的，即每一个临床症状对于分类的影响是相同的、没差别的。

4.1 获取实验数据

本论文研究所用到的数据来自朱文锋版的《中医诊断学》一书中，通过人工方法进行收集整理，在领域专家的指导下，排除噪声数据、合并冗余数据。先是手动录入到 EXCEL 表中，然后借助数据库的导入程序将 EXCEL 中的数据导入到数据库表中。至此，用于中医诊断数据以收集完成，共包含 125 种证候和 1084 个临床症状，但现在的数据表结构还不能直接用于朴素贝叶斯分类应用，需作进一步处理。

从 EXCEL 导入到数据库中的数据结构如图 4-1 所示，共 1856 行数据。第一列列名为序号，代表 125 种证候，编号为 1~125；第二列列名为核心，当核心取值为 1 时表示该行的临床症状为主证，当核心取值为 0 时表示该行的临床症状为辅证；第三列列名为临床症状，表示相应证候下的可能出现临床症状表现；第四列列名是证候，是与第一列呈一一对应关系的 125 种证候的名称。

序号	核心	临床症状	证候
1	1	新起恶寒发热	表证
1	0	新起恶风寒	表证
1	0	恶寒发热	表证
1	0	头身疼痛	表证
1	0	喷嚏	表证
1	0	鼻塞	表证
1	0	流涕	表证
1	0	咽喉痒痛	表证
1	0	微有咳嗽	表证
1	0	气喘	表证
1	0	舌淡红	表证
1	0	苔薄	表证
1	0	脉浮	表证
2	1	寒热往来	半表半里证
2	1	胸胁苦满	半表半里证
2	1	心烦喜呕	半表半里证
2	1	默默不欲饮食	半表半里证

图 4-1　从《中医诊断书》中获取的初始数据表结构截图

使用朴素贝叶斯分类时，是把临床症状看作特征属性，证候看作类别属性，根据一系列临床症状集合判断其所属的证候类型，就是分类的过程，也是对疾病作出诊断的过程。通常一个患病者往往会表现出多种症状，医生需要根据这一系列的临床症状集合，分析病情，做出诊治。在数据处理上，需要对数据进行多行变一行的处理，即把属于同一证候的临床症状以逗号分隔的形式放在一个单元格中，结果如图 4-2 所示。

序号	核心	临床症状
1	0	新起恶风寒,恶寒发热,头身疼痛,喷嚏,鼻塞,流涕,咽喉痒痛,微有咳嗽,气喘,舌淡红,苔...
1	1	新起恶寒发热
2	1	寒热往来,胸胁苦满,心烦喜呕,默默不欲饮食,口苦,咽干,目眩,脉弦
3	1	恶寒,畏寒,冷痛,喜暖,口淡不渴,肢冷踡卧,痰清稀,涎清稀,涕清稀,小便清长,大便稀溏,...
4	1	发热,恶热喜冷,口渴欲饮,面赤,烦躁不宁,痰黄稠,涕黄稠,小便短黄,大便干结,舌红,苔...
5	1	面色㿠白,面色暗淡,精神萎靡,身重踡卧,畏冷肢凉,倦怠无力,语声低怯,纳差,口淡不...
6	1	面色赤,恶寒发热,肌肤灼热,烦躁不安,语声高亢,呼吸气粗,喘促痰鸣,口干渴饮,小便...
7	0	鼻塞,流清涕,喷嚏,咽喉痒痛,咳嗽
7	1	恶风寒,微发热,汗出,脉浮缓,苔薄白,突发皮肤瘙痒,丘疹,瘖瘂,突发肌肤麻木,口眼㖞...
8	0	发热,鼻塞,流清涕,脉浮紧,咳嗽,哮喘,咯稀白痰,脘腹疼痛,肠鸣腹泻,呕吐,肢体厥冷,局...
8	1	恶寒重,无汗,头身疼痛,胸腹疼痛,苔白,脉弦紧
9	0	发热恶热,口渴喜饮,气短,肢体困倦,舌红,苔白,苔黄,脉虚数,卒然昏倒,汗出不止,气喘,...
9	1	发热,口渴,神疲,汗出,小便短黄

图 4-2　从《中医诊断书》中获取的初始数据表结构截图

处理该过程的核心代码为：

```
--多行变一行
CREATE FUNCTION dbo.f_str1(@id int,@fuid int)
RETURNS nvarchar(200)
AS
BEGIN
   DECLARE @r nvarchar(200)
   SET @r = ''
   SELECT @r = @r + ',' + [临床症状] FROM [0-证候症状全部初数据] WHERE [序号]=@id
                  and [核心]=@fuid
   RETURN STUFF(@r, 1, 1, '')
END
GO
--调用函数
SELECT distinct([序号]),[核心], [临床症状] = dbo.f_str1([序号],[核心]) FROM  [0-证候症状全部初数据]
```

由于不同证候的临床症状有主证和辅证两种，在甘肃中医药大学张老师的指导下，认为主证出现的概率更高，而辅证在证候的表现中发生概率相对较低，加之不同证候的辅证数目范围较大，有的没有辅证，有的辅证数目达到 28 个。在一条由主证和辅证共同构成的临床症状表现集的数据集中，会导致辅证多的证候的数目远远大于其他证候，在分类概率计算中以出现严重偏差。为解决上述问题，采用排列组合的思想，拆分重组初始数据，取出只有主证、全部主证加一个辅证的组合、全部主证加两个辅证的组合，共得到 9197 行数据。

证候	临床症状
亡阴证	身灼烦渴,唇焦面赤,脉数疾,汗出如油
气虚证	气短,乏力,神疲,脉虚
气陷证	自觉气坠,脏器下垂
气不固证	自汗,大便不固,小便不固,经血不固,精液不固,胎元不固
气脱证	气息微弱,汗出不止
血虚证	面色淡白,睑色淡白,唇色淡白,舌色淡白,脉细
血脱证	面色苍白,心悸,脉微,脉芤
气逆证	咳嗽喘促,呃逆,呕吐
气闭证	突发昏厥,绞痛
血热证	身热口渴,斑疹吐衄,烦躁谵语,舌绛,脉数
血寒证	患处冷痛拘急,畏寒,唇舌青紫,妇女月经后期,经色紫暗夹块

图 4-3　用于朴素贝叶斯分类的样本数据表结构截图

再将得到的 9197 行数据按行打乱，如图 4-3 所示。随机取 7000 行作为训练

集，用于训练分类器，剩余的 2197 行作为测试集，用于计算预测准确率，作为评价该分类器的一个指标。

4.2 开发应用程序

按照中医诊断标准，将临床症状作为特征属性，本文数据集中涉及的 1084 个临床症状，故特征属性的向量表示可以如下：

$$S = \{s_0, s_1, s_2, .., s_{1083}\}$$

对于每一个病例，核对 1084 个临床症状中的每一个临床症状，若出现该临床症状则在那一维上取值为 1，没出现则取值为 0。

证候就是类别属性，通过计算某一病例属于不同证候类别的概率，取概率值最大的那一类证候，认为是该病例所属的类别证候。

朴素贝叶斯分类的基本过程如下：

Step 1. 在训练集中分别统计 125 种证候出现的次数 N_{Xi}（i=0,1,2,…,124），除以训练数据总条数 N 得到 125 种证候的类别概率 $P(Xi) = \frac{N_{Xi}}{N}$；

Step 2. 统计出每种证候下各个特征属性的计数，即 125 种证候下每种临床症状的计数 N_{Si}，计算对应的条件概率 $P(Si \mid Xi) = \frac{N_{Si}}{N_{Xi}}$；

Step 3. 于 Step 2 中可能出现概率为 0 的情况，故须使用 Laplace Smoothing 校准改善概率为 0 的情况；

Step 4. 根据贝叶斯公式计算分属于不同类别属性时的概率值；

Step 5. 取概率值最大的结果即认为是所求的类别证候。

Step 6. 有时，为了分析一组临床症状集合可能属于的多种证候，会根据概率值大小，排序输出概率值从高到低排序靠前的三种证候。

本文实现基于 WinForm 的 C#窗体应用程序，进行素朴贝叶斯分类的研究。

实验硬件条件：

处理器：Inter(R) Pentium(R) CPU G3420 @ 3.20GHZ

安装内存（RAM）：8.00GB

实验软件条件：

操作系统：Windows7 旗舰版 64 位 Service Pack 1

数据库：Microsoft SQL Server 2008 R2

编译平台：Microsoft Visual Studio 2012

4.2.1 预测一组临床症状集合的类别证候

临床上常常需要根据某个具体病例表现出来的症状来诊断其表现为那种症候，进而对疾病进行诊治，因此，依据临床症状获知证候就显得尤为重要，借助朴素贝叶斯分类的思想，分类的过程就是判断临床症状集合所属的类别证候，是一种直观且有效的方法。

本节实现对于输入的一组临床症状集合，通过朴素贝叶斯分类器，判断其所属证候。临床上的诊断往往是针对特定病例，收集其所表现出来的临床症状，按照朴素贝叶斯分类的思想，判断这一系列的临床症状综合表现为那种或哪几种证候，据此进一步对疾病做诊断和治疗，分类预测的界面如图 4-4 所示。

先在左侧输入框中输入某一病例所表现出的全部临床症状的名称，无输入顺序限制，最多可输入的临床症状数目为 16 个，输入完成点击“开始预测”按钮，通过分类器的进行分类，将结果显示在右侧的文本框中，显示的结果是对于该病例可能性最高的前三种证候的名称。结果中的概率值不是实际意义的概率，而是作为比较大小的对比值。该过程同时实现了统计分类所用的时间的功能，在下方显示出本次分类所用的时间。

图 4-4　分类预测界面

进行分类的部分代码：

```
jointCounts = bc.MakeJointCounts(data_D, attributes, attributeValues);//存放症状计数
```

```
dependentCounts = bc.MakeDependentCounts(jointCounts, attributeValues[1].Length);//存放证候的类别计数
ppp = bc.Classify(jointCounts, dependentCounts, withLaplacian, attributeValues, testData);
public     int[] MakeDependentCounts(int[][][] jointCounts, int numDependents)
{
   int[] result = new int[numDependents];
   for (int k = 0; k < numDependents; k++){
         for (int j = 0; j < jointCounts[0].Length; j++)
               result[k] += jointCounts[0][j][k];
    }
   return result;
}
public     int[][][] MakeJointCounts(string[] data, string[] attributes, string[][] attributeValues)
{ //只有一个特征属性：临床症状
   int[][][] jointCounts = new int[attributes.Length - 1][][];
   //临床症状取值个数为：attributeValues[0].Length——44 个
   jointCounts[0] = new int[attributeValues[0].Length][];
   //每种症状都可能属于的证候数目为：attributeValues[1].Length——10 个
   for (int i = 0; i < attributeValues[0].Length; i++){
      jointCounts[0][i] = new int[attributeValues[1].Length]; }
         for (int i = 0; i < data.Length; i++){
            string[] tokens = data[i].Split('\t');
            int Index 症状 = AttributeValueToIndex(0, tokens[0], attributeValues);//症状
            int Index 症候 = AttributeValueToIndex(1, tokens[1], attributeValues);//证候
            ++jointCounts[0][Index 症状][Index 症候];
            }
   return jointCounts;
}
public double PartialProbability(string ZH, int[][][] jointCounts, int[] dependentCounts, bool withSmoothing, string[][] attributeValues, string[] testData)
{
   int xClass = attributeValues[0].Length;//临床症状的总数
   int[] Index 症状 = new int[xClass];//存放每种临床症状的个数
   for (int i = 0; i < Index 症状.Length; i++){
            Index 症状[i] = -1; //给数组赋值，设初始值为-1
   }
   for (int i = 0; i < attributeValues[0].Length; i++){
        for (int j = 0; j < testData.Length; j++){
              if (testData[j] == attributeValues[0][i])
                 Index 症状[j] = i;//测试集中出现的临床症状，相应的计数
        }
   }
}
```

```
//获取概率排序前三的值，返回前三个概率值
public int[] getTopThreeNum(int[][][] jointCounts, int[] dependentCounts, bool withSmoothing, string[][] attributeValues, string[] testData)
{
 double[] ppp = Classify(jointCounts, dependentCounts, withSmoothing, attributeValues, testData);
 //依次取出最大的三个数的下标
 int first = getMax(ppp);//取出最大的值的下标
 //将最大值替换为 0,接着取出次大的值，依次可以取出前三大的值
 ppp[first] = 0;
 int second = getMax(ppp);
 ppp[second] = 0;
 int third = getMax(ppp);
 int[] TopThree = { first, second, third };
 return TopThree;
}
```

4.2.2 计算预测准确率

本节实现通过计算预测准确率来评价朴素贝叶斯分类器的分类效率，预测准确率是一种直观且快速的评价标准。通过逐条对测试集中数据进行分类操作，对比经分类器所得到的证候与实际证候是否一致，判断准确率。

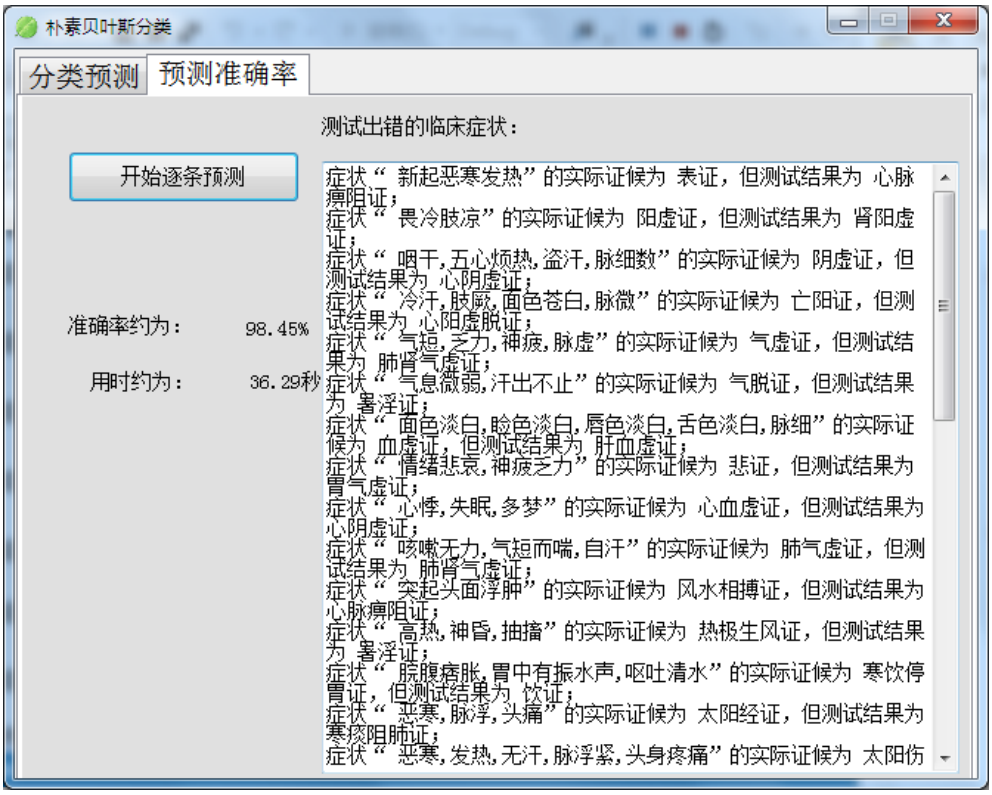


图 4-5　预测准确率的界面

为了评估由 7125 条数据作为训练集构造的分类器的分类效果时，通过测试 2197 条数据组成的测试集来统计分类的准确率。训练集和测试集比例接近 3:1。计算测试集中全部 2197 条测试数据的类别属性，将程序计算出来的证候类别结果跟已知证候类型进行对比，若相同则判定为分类正确，若不同则判定分类出错，统计结果，以此来计算该算法的准确率。同时将分类出错的显示在右侧文本框中。内容形式为：症状“高热,神昏,抽搐”的实际证候为 热极生风证，但测试结果为暑淫证。同时将出错结果以文本形式保存在本地文本文档中，便于后续分析，对改进算法有建设意义。用计算正确的条数除以测试集总条数所得结果可看作是分类器的预测准确率，并统计预测时间。

预测准确率的界面如图 4-5 所示，本次实验使用 2197 条数据进行测试，共有 34 条测试数据分类出错，得出预测准确率为 98.45%，用时为 36.29 秒。

表 4-1　分类出错的测试集

编号	测试集	实际证候	测试的证候
1	寒热往来,胸胁苦满,口苦,脉弦	少阳病证	半表半里证
2	气短,乏力,神疲,脉虚	气虚证	肺肾气虚证
3	咳嗽无力,气短而喘,自汗	肺气虚证	肺肾气虚证
4	发热,微恶风寒,脉浮数	卫分证	风热犯肺证
5	发热,微恶风寒,脉浮数,苔薄黄,咽喉肿痛	卫分证	风热犯肺证
6	发热,微恶风寒,脉浮数,苔薄黄	卫分证	风热犯肺证
7	寒热往来,胸胁苦满	少阳病证	肝胆湿热证
8	心烦不寐,舌尖红,脉细数,口燥咽干	少阴热化证	肝肾阴虚证
9	面色淡白,睑色淡白,唇色淡白,舌色淡白,脉细	血虚证	肝血虚证
10	面色淡白,睑色淡白,唇色淡白,舌色淡白,脉细,妇女月经色淡	血虚证	肝血虚证
11	恶寒,脉浮,头痛	太阳经证	寒痰阻肺证
12	恶寒,脉浮,头痛,恶风寒	太阳经证	寒痰阻肺证
13	恶寒,发热,无汗,脉浮紧,头身疼痛	太阳伤寒证	寒淫证
14	畏冷肢凉,苔白滑,脉沉迟无力	阳虚证	脾肾阳虚证
15	畏冷肢凉	阳虚证	肾阳虚证
16	气息微弱,汗出不止	气脱证	暑淫证
17	高热,神昏,抽搐	热极生风证	暑淫证
18	心悸,失眠,多梦,健忘,面色萎黄	心血虚证	思证
19	情绪悲哀,神疲乏力	悲证	胃气虚证

续表 4-1　分类出错的测试集

编号	测试集	实际证候	测试的证候
20	新起恶寒发热	表证	心脉痹阻证
21	突起头面浮肿	风水相搏证	心脉痹阻证
22	发热恶寒,小便不利	太阳蓄水证	心脉痹阻证
23	冷汗,肢厥,面色苍白,脉微	亡阳证	心阳虚脱证
24	冷汗,肢厥,面色苍白,脉微,脉微欲绝	亡阳证	心阳虚脱证
25	畏冷肢凉,苔白滑	阳虚证	心阳虚证
26	畏冷肢凉,自汗,气短	阳虚证	心阳虚证
27	畏冷肢凉,自汗	阳虚证	心阳虚证
28	咽干,五心烦热,盗汗,脉细数	阴虚证	心阴虚证
29	心悸,失眠,多梦	心血虚证	心阴虚证
30	咽干,五心烦热,盗汗,脉细数,形体消瘦	阴虚证	心阴虚证
31	咽干,五心烦热,盗汗,脉细数,形体消瘦,两颧潮红	阴虚证	心阴虚证
32	心烦不寐,舌尖红,脉细数	少阴热化证	血分证
33	畏冷肢凉,小便清长	阳虚证	阴证
34	脘腹痞胀,胃中有振水声,呕吐清水	寒饮停胃证	饮证

4.2.3 分成六类病症应用朴素贝叶斯分类

疾病的临床表现千变万化，错综复杂，在诊断疾病的过程中，为了执简驭繁，提纲挈领，有时则需通过判断某一病例所属的病位、病性等来更加直接地了解病情的性质和表现，确定证候的辨证方法。根据中医辨证论治的特点，临床常用的辨证方法有以下六种：八纲辨证、病性辨证、脏腑辨证、六经辨证、卫气营血辨证、三焦辨证，六种辨证方法所包含的证候个数及证候举例如表 4-2 所示。

表 4-2　6 类辨证方法及其证候

辨证方法	证候数量（个）	主要证候
八纲辨证	6	半表半里证、表证、寒证、热证、阳证、阴证
病性辨证	31	风淫证、寒淫证、湿淫证、惊证、恐证等
脏腑辨证	69	心脾气虚证、心肾不交证、肾阳虚证、心血虚证等
六经辨证	12	太阴病证、厥阴病证、阳明经证、少阳病证等
卫气营血辨证	4	气分证、卫分证、血分证、营分证
三焦辨证	3	上焦病证、下焦病证、中焦病证

程序实现的是对于给定的一组病例的临床症状集合判断分别在六种辨证方法下属于哪种证候的功能。需要对六个辨证方法训练六个分类器。输入一组临床症状之后，点击“开始预测”按钮，程序通过六个分类器分别进行分类操作，结果输出被测试临床症状集合在六种辨证方法下所属于的证候类型。实现该功能的界面如图 4-6 所示。

对于临床症状组合“面色萎黄，健忘，倦怠少食，失眠”的分类结果，可作如下解释：属于八纲辨证中的寒证；病性辨证中的思证，即证候的性质属于思；脏腑辨证中的肾阴虚证，即病位在肾；辨六经证中的阳明腑证；辨卫气营血证中的血分证；三焦辨证中的中焦病症。

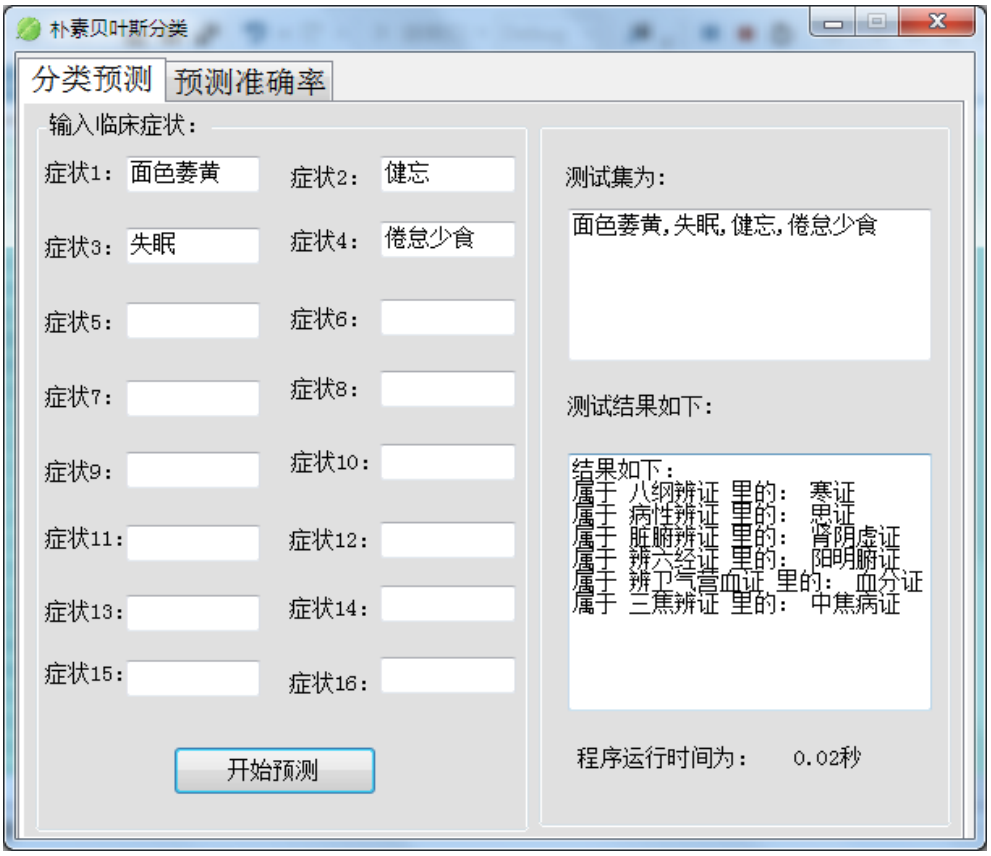


图 4-6　分成六类病症应用朴素贝叶斯分类的界面

4.2.4 拼音首字母提示输入

为了便于在实际操作分类器时，不必须输入临床症状的全部汉字，就能得到临床症状的名称，简化输入过程，在输入临床症状时只输入临床症状名称的拼音首字母，从而显示临床症状的全称，设计了拼音首字母提示输入功能。先根据所有临床症状的名称，按照汉语拼音的规则，构造与之完全一一对应的拼音首字母缩写，保存在文本文档中，结构如图 4-7 所示。

为方便使用，对于单字母的声母和双字母的声母，其拼音首字母都只取拼音的第一个字母，即对“臭 chou”、“常 chang”、“痴 chi”等两个字母“ch”为声母的字和“错 cuo”、“惨 can”等一个字母“c”为声母的字，规定拼音首字母均为“c”。

增加了汉字拼音首字母提示输入之后，例如，当在临床症状文本框中输入一个字母 m 的时候如图 4-8 中（a）所示，显示临床症状第一个汉字的拼音首字母为 m 的所有临床症状的下拉列表，可以通过滑动右侧滚动条找到想输入临床症状名称，按回车键或者用鼠标点击该临床症状就可以将临床症状显示在文本框中。也可继续输入字母 f，则显示临床症状名称前两个汉字拼音首字母分别为 m、f 的临床症状，如图 4-8 中（b）所示。

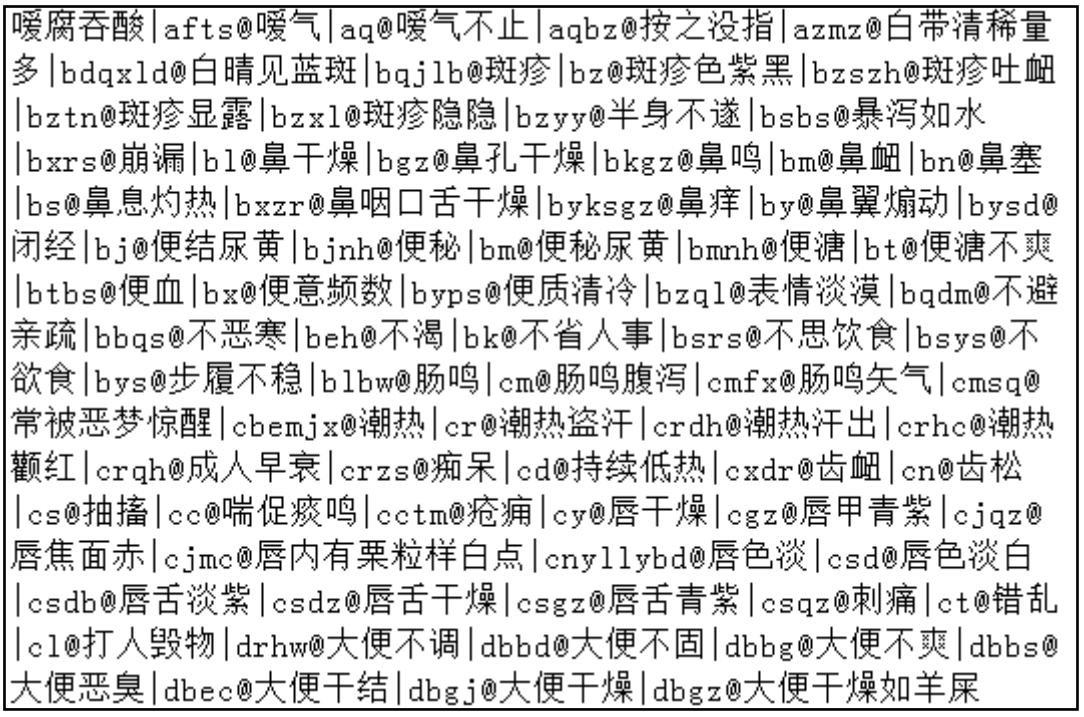


图 4-7　临床症状名称及其拼音首字母对应结构的截图

（a）　　（b）

图 4-8　拼音首字母提示输入

4.3 本章小结

本章主要阐述的是朴素贝叶斯分类在中医诊断中的应用，具体是中医学中的临床症状在证候类型中的分类研究，使用高级程序设计语言 C#进行界面实现，编写一个朴素贝叶斯分类器，实现了分类功能、计算准确率功能，并添加了拼音首字母提示输入功能

通过朴素贝叶斯分类的应用，在中医专家对分类诊断结果的进行评估之后，认为结果基本符合中医理论现状，具有合理性和指导性。尤其是区分六类病症之后的应用，能对一组临床症状表现出细致的分类，这对于指导中医诊断有实践意义。

第五章　基于贝叶斯网络的中医诊断研究

5.1 中医诊断与贝叶斯网络

中医诊断的过程可以描述如下：首先是根据患者身体表现出的异常症状获取临床症状集合，这样构成了疾病的状态集合；其次研究这些具有特定的关联关系的临床症状和证候，构建中医诊断的知识表示系统；最后运用知识表示系统指导疾病的治疗。贝叶斯网络就是以有向图的形式模拟人类推理思维过程的模型，可用于指导中医诊断。贝叶斯网络适合于表达诊断类型的问题，在故障诊断和疾病诊断等方面都有较好的效率和适用性，可以良好的表达用于诊断的变量之间的层次关系和关联关系。

中医诊断中的贝叶斯网络的研究主要有以下三部分：

1）网络节点和网络结构

网络结构图中的节点表示中医诊断中涉及的临床症状和证候信息，节点之间的有向边表示临床症状和证候之间的关系，或临床症状之间的依赖关系。临床症状和证候的类型为布尔类型，取值为 0，表示节点所指的变量不出现；反之，取值为 1，表示节点所指的变量出现[53]。

2）条件概率表

对网络图中的每个节点都附以与该变量相联系的条件概率分布情况，以条件概率表的形式给出，表明了临床症状和证候之间的概率依赖关系，是对中医诊断知识的定量表达。

3）推理模式

结合定量知识和定性知识进行推理研究，是贝叶斯网络学习的最终目的。尽管 Cooper 曾证明了贝叶斯网络推理的应用是一个 NP-complete 问题[54]，近年来对于特定类型网络的推理算法也取得了一定进展。

本文主要研究贝叶斯网络结构学习，提出基于互信息的方法充分利用数据集来构造网络结构。

5.2 贝叶斯网络结构学习

常用的结构学习方法主要有两类：一种是基于依赖性测试的学习，在给定的数据集中，利用条件独立性测试，评估变量之间的依赖关系。优点是该方法将条件独立性测试和网络结构的搜索分离开，直观的寻找变量之间的关系。缺点是条

件独立性测试的结果误差对网络结构的构建影响较大，且条件独立性测试的次数会随着变量的数目成指数级增长。另一种是基于搜索评分的学习，原理是在所有变量的结构空间内，按照一定的搜索策略及评分准则构建贝叶斯网络结构，从所有可能的网络结构中搜索最佳的结构，但这一过程被证明是 NP-complete 问题。

一般的贝叶斯网络是由相关领域的专家根据事物之间的关系来构建，但这样构建的网络模型，其客观性无法保证，因此这里希望通过借助客观的数据信息，通过观察和分析数据得出贝叶斯网络。

完整数据集下构建贝叶斯网络结构的三种方法：1）依靠专家建模；2）从数据中学习；3）从知识库中创建。在实际应用中常常综合这些方法，以领域专家为主导，以数据库和知识库为辅助手段，来保证建模的效率和准确性。但是，在不具备专家知识或知识库的前提下，从数据中学习贝叶斯网络结构的研究则显得根据实践价值。

建立贝叶斯网络的过程，可以看作是对实际问题进行图形网络化的过程。首先将实际问题的事件抽象为节点，节点必须有明确意义，并且至少有是或非两个取值状态，或多个可以确定的取值状态，取值状态在概率意义上是完备且互斥的。即所有状态在某一时刻只发生一个，所有状态的概率之和为 1。其次确定节点之间的连线，只在有明确因果或相关关系的节点之间添加连线，同时应防止出现环。可以使用经验判断，专家指导等方法[55、56]。

5.3 基于条件互信息学习贝叶斯网络结构

本文结合信息论的知识，引入一种基于互信息知识学习贝叶斯网络结构的方法。

互信息（Mutual Information）是信息论里一个十分有用的概念，可以解释为一个离散随机变量 y_j 的取值对于另一个变量 x_i 取值的确定性能力，互信息值可用来衡量两个变量之间的依赖关系[57、58]。互信息可按下述公式计算：

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

而条件互信息（Conditional Mutual Information）简单可以理解为有条件的互信息[59]。

定义如下：在 XYZ 联合集中，在给定 z_k 的条件下，x_i 与 y_j 之间的互信息量就称为条件互信息量。

条件互信息量的计算方法如下式：

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

常用以 2 为底的对数进行计算。

互信息和条件互信息有如下性质：

1、对称性，即有 I(X,Y)=I(Y,X)和 I(X,Y|Z)=I(Y,X|Z)；

2、非负性，即 I(X,Y)≥0 和 I(X,Y|Z)≥0，当且仅当 X 和 Y 条件独立时有 I(X,Y)=0；当且仅当在给定 Z 时，若 X 和 Y 条件独立有 I(X,Y|Z)=0 成立。

基于条件互信息学习贝叶斯网络结构的基本过程是：

Step 1：首先根据研究问题确定作为网络节点的 N 个变量$\{X_1,X_2,\ldots,X_n\}$；在中医诊断中节点为临床症状和证候，都是离散变量，取值为 1 或 0，表示出现或不出现。

Step 2：建立无向图；

首先可以在证候与每个临床症状之间确定一条边；其次确定临床症状之间边的方法可描述为如下：结合已有数据集 D，统计临床症状的两两组合在数据集中出现的次数 $C(X_i, X_j)$，取一个很小的数值 e_1，当 $C(X_i, X_j)>e1$ 时，连接 X_i 和 X_j，即在 X_i 和 X_j 之间确定一条边。以此方法建立一个无向图。

Step 3：对无向图进行“剪枝”操作，结合条件互信息的概念，判断两个变量之间是否条件独立。在给定节点变量的概率分布时，通过计算 I(X,Y|Z)是否为 0 来判断在给定 Z 的条件下，X、Y 是否相互独立，若 I(X,Y|Z)=0，则可判定 X、Y 相互独立，删除 X、Y 之间的边，否则 X、Y 互相依赖。

但是在实际计算过程中，节点变量的概率分布不易获得，因此常常借助于样本数据集 D，利用统计的思想进行计算概率分布来近似估计变量的概率值。使用统计计算的思想计算节点的概率值的方法如下：假设节点 A 有 m 个取值状态 $A_1, A_2, A_3, \ldots, A_m$，则计算节点 A 不同取值状态的概率的方法如下：

$$P(A_i)=\frac{A_i\text{在训练集中出现的次数}}{\text{训练集总数}}。$$

如果 As 表示节点 A 的一个状态，Bs 表示节点 B 的一个状态，则在 As 发生时 Bs 发生的概率，可以用条件概率的形式表示为：

$$P(Bs|As)=\frac{\text{As和Bs共同发生的总次数}}{\text{As发生的总次数}}。$$

这时判断变量之间的条件独立性的过程是：对于临床症状 X_i, X_j 和证候类别 C，给定一个阈值 e_2，计算 $I(X_i, X_j|C)$，若 $I(X_i, X_j|C)<e_2$ 则判定临床症状 X_i, X_j 条件独立，可以删除边，否则，在给定 C 时，判定 Xi, X_j 有依赖关系。

而对于临床症状与证候之间的边的操作，仍是通过计算互信息 $I(X_i, C)$来判断，对于给定一个阈值 e_3，若 $I(X_i, C)<e_3$，则移除 Xi 和 C 之间的边，达到“剪枝”的目的。

Step 4：确定边的方向，得到有向图；

对于存在边的两个临床症状 X_i, X_j，判定其边的方法要用到临床症状与证候之间的互信息的差值。对于给定一个阈值 e_4:

1、当$|I(X_i,C)-I(X_j,C)|\leqslant e_4$时，表示边的方向是$X_i \rightarrow X_j$或$X_i \leftarrow X_j$；这一情况的出现不可避免，因此需要结合其他知识进行分析研究，以确定最可能的边的方向。

2、当$|I(X_i,C)-I(X_j,C)|\geqslant e_4$时，若$I(X_i,C) < I(X_j,C)$，规定边的方向为：$X_i \rightarrow X_j$；反之则为$X_i \leftarrow X_j$。

在计算过程中可能会出现$X_i \rightarrow X_j$或$X_i \leftarrow X_j$两个方向的边存在的情况，这表示两个变量之间训在依赖关系，此时边的方向需要借助其他知识进行确定，在这种情况下，贝叶斯网络会存在多种可能的结构形式。

Step 5：结合相关领域知识修正有向图，得到最具可能性的贝叶斯网络结构图。

5.4 实验及结果展示

实验选取的是中医诊断系统的临床症状和证候信息数据表，共 9197 行数据作为训练样本集，命名为 D。基于互信息方法构造脾肺气虚证及其临床症状之间关系的贝叶斯网络结构图。其中一个类别属性脾肺气虚证 X，9 个临床症状“咳嗽”、“气喘”、“咯痰”、“食少”、“腹胀”、“便溏”、“气短”、“神疲乏力”、“脉虚”为特征属性，构造贝叶斯网络，以分析其相互之间的影响关系。

为后续显示清晰简洁，需要对节点信息进行编码整理，用英文字母简化表示临床症状，对应关系如表 5-1。

表 5-1　对网络节点进行编码整理结果表

临床症状名称	ID	代码	临床症状名称	ID	代码
咳嗽	0295	A	脉虚	0438	F
气短	0544	B	食少	0703	G
神疲乏力	0670	C	咯痰	0365	H
便溏	0029	D	腹胀	0188	J
气喘	0541	E	脾肺气虚证	98	X

表 5-2　临床症状组在样本数据集中的计数

编码	临床症状组	Count	编码	临床症状组	Count
1	A,G	56	19	A,F	79
2	A,J	56	20	A,B	128
3	A,D	56	21	D,B	164
4	E,G	56	22	D,C	175
5	E,J	56	23	J,F	176
6	E,D	56	24	D,F	178
7	E,H	57	25	J,B	182
8	E,B	57	26	A,E	187
9	E,C	57	27	J,C	187
10	E,F	57	28	G,B	256
11	A,H	66	29	G,C	267
12	H,G	66	30	G,F	270
13	H,D	66	31	G,J	347
14	H,C	66	32	C,F	349
15	H,F	66	33	G,D	351
16	H,J	67	34	B,C	361
17	A,C	68	35	B,F	592
18	H,B	75	36	J,D	686

当取 e_1=100 时，得到的无向图如图 5-1 所示：

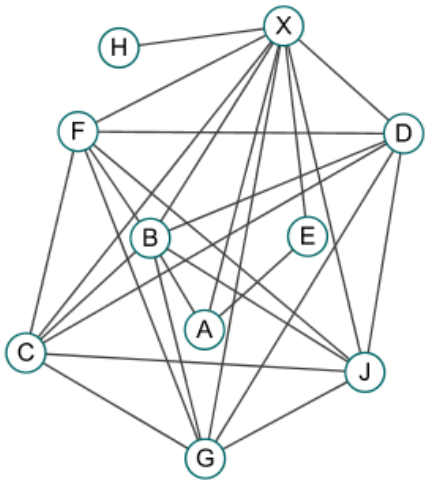
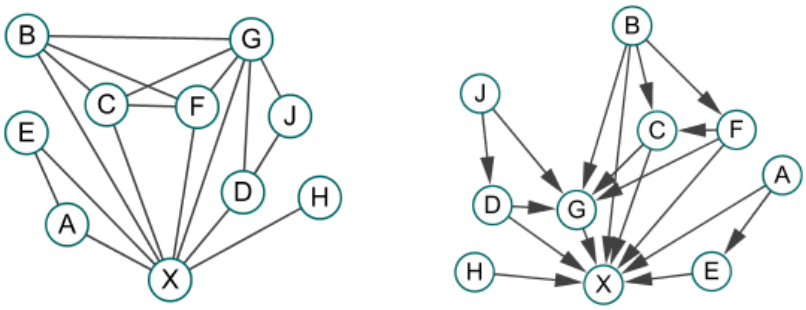


图 5-1　初始无向图

下面对无向图进行“剪枝”，使用互信息进行计算的过程，由于所取得阈值不同时会得到不同的贝叶斯网络结构图，因此这里分别取了三组不同的阈值，进行结果的对比展示。

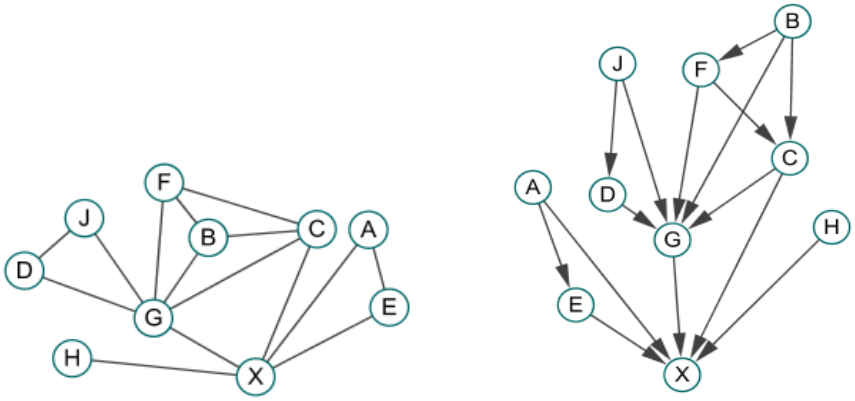
当阈值分别为 e_2=2.5，e_3=3.5 时得到剪枝无向图如图 5-2（a），阈值 e_4=0.1 有向图 5-2（b）。



（a）无向图　　（b）有向图

图 5-2　剪枝后的图一

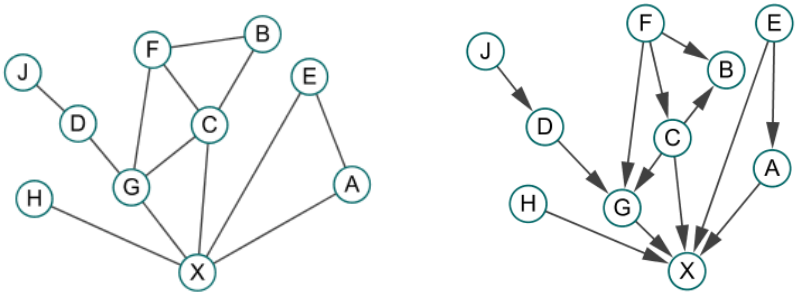
当阈值分别 A 为 e_2=2.5，e_3=4 时得到剪枝无向图如图 5-3（a），阈值 e_4=0.1 有向图 5-3（b）。



（a）无向图　　（b）有向图

图 5-3　剪枝后的图二

当阈值分别为 e_2=2.8，e_3=4 时得到剪枝无向图如图 5-4（a），阈值 e_4=0.1 有向图 5-4（b）。



（a）无向图　　（b）有向图

图 5-4　剪枝后的图三

对于使用互信息概念得到的贝叶斯网络结构图 5-4（b）可作如下解释，该结构表明与脾肺气血证 X 有直接影响关系的临床症状是 A、E、C、G、H 所代指的临床症状，另外 J、B、F、D 临床症状与证候 X 无直接依赖关系，同时临床症状之间的有向边表示它们之间的依赖关系，即当出现一个临床症状时，另一个临床症状也可能会出现。

对于其中有向边寻找中医依据，目前对于 13 条有向边，结合贝叶斯网络的概念和中医知识对其中的 11 条边找到相关的中医知识做支撑，即有 84.6%的边找到了可以支撑的中医知识。其中依据朱文锋《中医诊断学》书中对脾肺气虚证 X 及其临床症状的记录，该证以咳嗽 A、气喘 E、咳痰 H、食少 G、气虚症状 C 共见为辨证的主要依据。故可能存在这样的边：A→X、C→X、E→X、G→X、H→X。《黄帝内经・素问・玉机真藏论篇》[60]有对于脉象有如下描述“其不及，则令人喘，呼吸少气而咳，上气见血，下闻病音”可知脉虚出现会导致气短出现即 F→B；便溏，中医指拉稀薄的大便，和西医的腹泻是同一回事，《黄帝内经・素问・玉机真藏论篇》中有“入五藏，则瞋满闭塞，下为飧泄，久为肠澼。”的描述，支持了腹胀到便溏的边 J→D；姚乃礼主编《中医症状鉴别诊断学》[61]一书中对食欲不振有如下介绍：“导致脾胃湿热食欲不振的原因有：呕恶厌食，脘腹痞闷，周身疲乏倦怠，大便溏而不爽，溲黄而短，舌红，苔黄白而腻，脉濡数或滑；导致脾胃气虚食欲不振有：不思饮食，食后腹胀，或进食少许即泛泛欲吐，气短懒言，倦怠少力，舌淡苔白，脉缓弱；导致脾胃虚寒食欲不振有：饮食无味，不知饥饿，进食稍多则脘腹闷胀欲呕，脘腹隐痛或阵痛，喜暖畏寒，按之则舒，疲倦气短，四肢不温，大便溏薄，舌淡苔白，脉沉迟。”，便溏和食少之间的有边 D→G，神疲乏力和食少之间有边 C→B，脉虚和食少之间有边 F→G；其中对于气喘有如下介绍“风寒闭肺气喘：喘急胸闷，伴有咳嗽，咯痰清稀色白，初起多兼恶寒发热，无汗；风热犯肺气喘：喘急烦闷，伴有咳嗽，或见发热，汗出恶风，口渴，胸痛。”可知气喘会伴随咳嗽即 E→A。

5.5 本章小结

本章通过研究脾肺气虚证的贝叶斯网络结构学习，使用互信息和条件互信息进行计算，确定了其临床症状之间的依赖关系，以及临床症状与证候之间的影响关系，对于研究脾肺气虚证的发病表现具有提示意义。当所选择的阈值不同时，得到的网络结构也不相同，但都保证了数据集中属性的完整性，这种依赖关系对于研究中医诊断有微弱的提示意义。

通过引入互信息的概念，对贝叶斯网络节点进行依赖度的计算，发掘了节点

之间中可能存在的隐含依赖关系，这对于分析证候的临床症状之间的关系起到了提示作用。另外，对于采用互信息计算过程中，用到的阈值，如果选取不同的阈值时，会产生多个不同结构的贝叶斯网络，因此需要在中医专家的指导下，对可能的网络结构进行评估，以确定最可能的网络结构。本论文所使用的贝叶斯网络对于研究中医证候分类没有任何效果。

第六章　总结与展望

6.1 总结

本文以中医诊断临床症状和证候数据为研究基础，将贝叶斯方法中的朴素贝叶斯分类方法和贝叶斯网络应用于中医诊断中，设计人机交互的应用程序，实现了对临床症状组合的分类判别功能和证候的临床症状之间依赖关系的确定，具有一定的实用性和科学性。

针对本文研究所需，对于贝叶斯方法的研究和应用主要有以下内容：根据目前有关贝叶斯方法的研究探索，整理总结贝叶斯方法的理论知识和算法原理；结合中医诊断学的知识应用贝叶斯方法中的两个重要算法朴素贝叶斯分类和贝叶斯网络，使用朴素贝叶斯分类算法对中医诊断学中 1085 个临床症状的组合进行证候的分类研究，使用贝叶斯网络破除朴素贝叶斯的条件独立性假设，重点研究临床症状之间的依赖关系，构建贝叶斯网络结构图。

本文在研究过程中遇到的问题及解决思路有：

1）在进行朴素贝叶斯分类器的开发时，先以一个两分类的实例进行代码的编写，然后将相同的思路用于多分类目标的实现，充分结合统计知识使用计算机语言进行操作和学习。在输出结果时，开始选择输出一个分类结果，但是由于中医诊断中包含 125 种证候，每个证候之间在医学意义上有相似之处，所以将可能的输出结果扩大为 3 个，这对于更好的研究疾病表现和病因，指导治疗更具参考价值。

2）在贝叶斯网络的中医诊断研究中，重点研究贝叶斯网络的结构学习，通过参考从《中医诊断学》一书中获得的数据集，提出一种结合信息论的知识利用充分挖掘数据集中有用信息学习贝叶斯网络结构的方法，用于分析其临床症状之间的依赖关系。

6.2 展望

本文在应用贝叶斯方法时依据了特征属性对结果的影响程度完全相同的假设，这在现实中不太合理，因为不同的特征表现对于结果的影响程度很多时候是不同的，因此后续的研究有必要引入权值的概念，而这些权值的设定则需要来自更多临床医学的数据作支撑，所以这是需要进一步深入研究和分析的要点。另外在对贝叶斯网络的研究上重点放在了构建网络结构上面，这对于在中医诊断中分

析临床症状之间的依赖关系有实际的参考价值，进一步的研究可以放在贝叶斯网络的参数学习和推理应用上，探索不同的临床症状对于证候的贡献度。

参考文献

[1] 张硕 & 商洪才. 中医疾病证候与组学的相关性研究现状[J]. 中西医结合学报. 2011, 9: 1286-1291.

[2] 孙喜灵, 姜伟炜 .中医证候的结构化研究[J]. 世界中医药. 2013: 146-148.

[3] 贾运滨 & 魏江磊. 数据挖掘技术在中医证候研究中的应用述评[J]. 中国中医急症. 2010: 1184-1186.

[4] Gonzalez, G. H., Tahsin, T. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery[J]. Brief Bioinform. 2016, 17: 33-42, doi:10.1093/bib/bbv087.

[5] Funahashi, K. Multilayer neural networks and Bayes decision theory[J]. Neural Netw. 1998, 11: 209-213.

[6] Muralidharan, V., Sugumaran,V.Fault Diagnosis of Monoblock Centrifugal Pump Using Stationary Wavelet Features and Bayes Algorithm[J]. Asian Journal of Science and Applied Technology. 2014, 3: 1-4.

[7] Sharma, R. K., Sugumaran, V. A comparative study of naïve Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal[J]. International Journal of Decision Support Systems. 2015, 1: 115-129.

[8] 伯聪. 扁鹊和扁鹊学派研究[M]. (陕西科学技术出版社, 1990).

[9] Liao, S.-H. Expert system methodologies and applications—a decade review from 1995 to 2004[J]. Expert systems with applications. 2005, 28: 93-103.

[10] 秦笃烈 & 鲍亦万. 中医计算机模拟及专家系统概论[J]. 北京人民卫生出版社. 1989.

[11] Gupta, N. Artificial neural network[J]. Network and Complex Systems. 2013, 3: 24-28.

[12] 朱文锋. 创立以证素为核心的辨证新体系[J]. 湖南中医学院学报. 2004, 24: 38-39.

[13] 郑舞 & 刘国萍. 常见数据挖掘方法在中医诊断领域的应用概况[J]. 中国中医药信息杂志. 2013, 20: 103-107.

[14] Bayes, T., Price, R. An essay towards solving a problem in the doctrine of chances[M]. (C. Davis, Printer to the Royal Society of London, 1763).

[15] Stigler, S. M. The True Title of Bayes's Essay[J]. arXiv preprint arXiv:1310.0173. 2013.

[16] Bayes, T. Essay towards solving a problem in the doctrine of chances[M]. (Biometrika Trust, 1958).

[17] De Finetti, B. Probability, induction, and statistics[J]. 1972.

[18] Wald, A. Sequential analysis[M]. (Courier Corporation, 1973).

[19] Robbins, H. & Monro, S. A stochastic approximation method[J]. The annals of mathematical statistics. 1951: 400-407.

[20] Geiger, D. & Heckerman, D. Parameter priors for directed acyclic

graphical models and the characterization of several probability distributions[C]. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999: 216-225.

[21] Safavian, S. R. & Landgrebe, D. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics. 1991, 21: 660-674.

[22] Pearl, J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M].(Morgan Kaufmann, 2014).

[23] Pearl, J. Heuristics: intelligent search strategies for computer problem solving[J]. 1984.

[24] Pearl, J. & Verma, T. S. A theory of inferred causation[J]. Studies in Logic and the Foundations of Mathematics. 1995, 134: 789-811.

[25] Pearl, J. Causality: models, reasoning and inference[J]. Econometric Theory. 2003, 19: 46.

[26] Ramoni, M. & Sebastiani, P. Learning Bayesian networks from incomplete databases[C]. Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1997: 401-408.

[27] 林士敏, 田凤占. 斯网络的建造及其在数据采掘中的应用[J]. 清华大学学报: 自然科学版. 2001, 41: 49-52.

[28] 霍利民, 朱永利. 一种基于贝叶斯网络的电力系统可靠性评估新方法[J]. 电力系统自动化. 2003, 27: 36-40.

[29] 李俭川, 胡茑庆. 贝叶斯网络理论及其在设备故障诊断中的应用[J]. 中国机械工程. 2003, 14: 896-900.

[30] 徐璡, 许朝霞. 基于贝叶斯网络原理的835例冠心病病例中医证候分类研究[J]. 上海中医药杂志. 2014, 48: 10-13.

[31] 郭, 靄. 黄帝内经素问校注语译[M]. (天津科学技术出版社, 1981).

[32] 李宗明. 临床症状鉴别诊断学[J]. 1995.

[33] 赵金铎编. 中医证侯鉴别诊断学[M]. (人民卫生出版社, 1987).

[34] 李影林编. 临床医学检验手册[M]. (吉林科学技术出版社, 1987).

[35] 韩素杰. 基于民国时期诊法著作的中医诊断学术研究[硕士]. 中国中医科学院. 2015.

[36] 朱文锋. 中医诊断学[J]. 中医教育. 2002, 21: 39-40.

[37] Jaynes, E. T. Prior probabilities[J]. IEEE Transactions on systems science and cybernetics. 1968, 4: 227-241.

[38] Largeault, J., Jaynes, E. (JSTOR, 1986).

[39] Lee, P. M. Bayesian statistics: an introduction[M]. (John Wiley & Sons, 2012).

[40] Sheldon, R. A first course in probability[M]. (Pearson Education India, 2002).

[41] Rudin, W. Principles of mathematical analysis[M]. Vol. 3 (McGraw-Hill New York, 1964).

[42] Fine, A. Hidden variables, joint probability, and the Bell inequalities[J]. Physical Review Letters. 1982, 48: 291.
[43] Jeffreys, H. Scientific inference[M]. (Cambridge University Press, 1973).
[44] Wu, W.-t., Jin, W.-z. Research on choice of travel mode model based on naive Bayesian method[C]. Business Management and Electronic Information (BMEI), 2011 International Conference on. IEEE, 2011: 439-444.
[45] 郑熠煜. 贝叶斯分类方法及其在冠心病诊疗中的应用研究.[大连海事大学]. 2013.
[46] Bermejo, P., Gámez, J. A. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier[J]. Knowledge-Based Systems. 2014, 55: 140-147.
[47] Ng, A. Y. & Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes[J]. Advances in neural information processing systems. 2002, 2: 841-848.
[48] Field, D. A. Laplacian smoothing and Delaunay triangulations[J]. International Journal for Numerical Methods in Biomedical Engineering. 1988, 4: 709-712.
[49] Cai, D., He, X. Learning a spatially smooth subspace for face recognition[C]. Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-7.
[50] Hänninen, M. Bayesian networks for maritime traffic accident prevention: benefits and challenges[J]. Accident Analysis & Prevention. 2014, 73: 305-312.
[51] Buntine, W. Learning classification rules using Bayes[C]. Proceedings of the sixth international workshop on Machine learning. 2016: 94-98.
[52] Patil, T. R. & Sherekar, S. Performance analysis of Naive Bayes and J48 classification algorithm for data classification[J]. International Journal of Computer Science and Applications. 2013, 6: 256-261.
[53] Hesar, A. S., Tabatabaee, H. Structure learning of bayesian networks using heuristic methods[C]. Proc. of International Conference on Information and Knowledge Management (ICIKM 2012). 2012.
[54] Ullman, J. D. NP-complete scheduling problems[J]. Journal of Computer and System sciences. 1975, 10: 384-393.
[55] Dumais, S., Platt, J. Inductive learning algorithms and representations for text categorization[C]. Proceedings of the seventh international conference on Information and knowledge management. ACM, 1998: 148-155.
[56] Schlosberg, C. E., Schwantes-An, T.-H. Application of Bayesian network structure learning to identify causal variant SNPs from resequencing data[C]. BMC proceedings. BioMed Central, 2011: S109.
[57] Kraskov, A., Stögbauer, H. Hierarchical clustering based on mutual

information[J]. arXiv preprint q-bio/0311039. 2003.
[58] Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient[J]. Proceedings of the National Academy of Sciences. 2014, 111: 3354-3359.
[59] Wyner, A. D. A definition of conditional mutual information for arbitrary ensembles[J]. Information and Control. 1978, 38: 51-59.
[60] 战国, 佚名. 全本黄帝内经[J]. 线装经典编委会. 昆明: 云南教育出版社, 2010.
[61] 姚乃礼. 中医症状鉴别诊断学[J]. 北京: 人民卫生出版社, 2000.

在学期间的研究成果

一、发表论文

1. Guang Zheng, **Yanjing Zhang**, He Zhang, Miaofeng Li, Xuwen He, Yutao Qi, Junping Zhan, Hongtao Guo; "Biological Functional Analysis of Chinese Herbal Medicines Against wind-cold-dampness Syndrome", 2016 IEEE International Conference.

二、参与课题

1. HONG KONG BAPTIST UNIVERSITY, Toward integrative medicine for innovative combinational therapeutics: an unique platform unifying approaches of data mining, drug design and bioinformatics to traditional Chinese medicine-based new drug discovery and clinical knowledge discovery. Project No: RC-IRMS/12-13/02.

2. HONG KONG BAPTIST UNIVERSITY, Toward an emerging paradigm of network medicine A novel approach to combinational herbs/compounds-based drug discovery in rheumatoid arthritis. Project No: SDF13-1209-P01.

3. 中国中医科学院中医临床基础研究所.高血压病中医健康服务规范示范研究

4. 甘肃源信电子科技有限公司.常见中药特征化学成分靶蛋白生物信息学分析

致　　谢

文章写至此处，意味着我在兰州大学的学业生涯接近谢幕时刻。细看满篇的文字，欲言还休，近三年的光阴，已悄然逝去，回首过去，不禁想起第一次走进兰州大学的情景，先是被门口苍劲的四个大字吸引，然后又沉浸于毓秀湖中小喷泉的美。三年的时光，陪伴我的不仅有美丽的校园，更有自强不息，独树一帜的百年兰大优良学风，以其春风化雨、润物无声的姿态感染着我，改变着我。

本篇硕士学位论文是由我的导师郑光副教授担纲指导重责下进行并完成的，再次以万分诚挚之心感怀导师的殷勤付出和悉心指导，从开始的选择论文课题、中期的实验设计实施，直至最后论文的定稿完成，郑老师始终给予我的都是耐心的指导和坚定的支持。郑老师渊博的知识体系、严谨的工作作风和昂扬向上的心态品格是我在今后生活学习的榜样。

在此，我也要特别感谢我的父母家人给予我的关怀和支持，没有他们我不可能完成学业。此外感谢飞云楼 501 实验室的郝婷同学、陈菊萍同学、张赫同学，他们是和我并肩战斗过的小伙伴，有过很多快乐的时光，还要感谢已毕业的师兄师姐们和已成为 501 中流砥柱的师弟们，你们身上那种直面难题乐观豁达的信念必将影响我的一生。同时感谢我宿舍 347 的小伙伴们，感谢你们为我营造了一个温馨舒适的休息环境，成为我坚持奋斗的力量来源。

行文的最后，郑重的向各位参与审阅、评议、答辩的老师们表达感谢之情！生命的每一个阶段都有会遇到不同的人，有些人、有些事，在不经意之间会影响我们的一生，这三年，有你们，我很幸运。我会努力把这种幸运化作我前进的动力，努力做得更好，我想这也是我能回馈给你们的最好的礼物吧。谨以此文将最美好的祝愿送给你们，愿永远平安、健康、幸福！

附　　录

针对图 5-4（b）所示的贝叶斯网络中的有向边的寻找中医理论的支撑。

1、依据朱文锋《中医诊断学》书中对脾肺气虚证 X 及其临床症状的记录，该证以咳嗽 A、气喘 E、咳痰 H、食少 G、气虚症状 C 共见为辨证的主要依据。故可能存在这样的边：A→X、C→X、E→X、G→X、H→X。

2、脉虚→气短　F→B

帝曰：秋脉太过于不及，其病皆何如？岐伯曰：太过则令逆气而背痛，愠愠然；其不及，则令人喘，呼吸少气而咳，上气见血，下闻病音。

---《黄帝内经·素问·玉机真藏论篇》

译文：

黄帝说：秋脉太过和不及，都会发生什么病变呢？

岐伯说：太过会使人气逆，背部作痛，郁闷而不舒畅；如果不及，会使人喘促，呼吸气短、咳嗽，在上部会发生气逆出血，在下的胸部则可以听到喘息的声音。

---2010 年中华书局刊印姚春鹏译注《黄帝内经》

4、便溏→食少　D→G　神疲乏力→食少　C→B　脉虚→食少　F→G

导致脾胃湿热食欲不振的原因有：呕恶厌食，脘腹痞闷，周身疲乏倦怠，大便溏而不爽，溲黄而短，舌红，苔黄白而腻，脉濡数或滑。导致脾胃气虚食欲不振有：不思饮食，食后腹胀，或进食少许即泛泛欲吐，气短懒言，倦怠少力，舌淡苔白，脉缓弱。导致脾胃虚寒食欲不振有：饮食无味，不知饥饿，进食稍多则脘腹闷胀欲呕，脘腹隐痛或阵痛，喜暖畏寒，按之则舒，疲倦气短，四肢不温，大便溏薄，舌淡苔白，脉沉迟。

---2000 年人民卫生出版社出版的姚乃礼主编《中医症状鉴别诊断学》

5、腹胀→便溏　J→D

黄帝问曰：太阴阳明为表里，脾胃脉也，生病而异者何也？

岐伯对曰：阴阳异位，更虚更实，更逆更从，或从内，或从外，所从不同，故病异名也。

帝曰：愿闻其异状也。

岐伯曰：阳者，天气也，主外；阴者，地气也，主内。故阳道实，阴道虚。故犯贼风虚邪者，阳受之；食饮不节，起居不时者，阴受之。阳受之，则入六府，阴受之，则入五藏。入六府，则身热不时卧，上为喘呼；入五藏，则填满闭塞，下为飧泄，久为肠澼。

---《黄帝内经 素问 玉机真藏论篇》

译文：

黄帝问道：太阴、阳明两经，互为表里，是脾胃所属的经脉，而所生的疾病不同，是什麽道理？

岐伯回答说：太阴属阴经，阳明属阳经，两经循行的部位不同，四时的虚实顺逆不同，病或从内生，或从外入，发病原因也有差异，所以病名也就不同。

黄帝道：我想知道它们不同的情况。

岐伯说：人身的阳气，犹如天气，主卫互于外；阴气，犹如地气，主营养于内。所以阳气性刚多实，阴气性柔易虚。凡是贼风虚邪伤人，外表阳气先受侵害；饮食起居失调，内在阴气先受损伤。阳分受邪，往往传入六腑；阴气受病，每多累及五脏。邪入六腑，可见发热不得安卧，气上逆而喘促；邪入五脏，会出现脘腹胀满，闭塞不通，在下为大便泄泻不止的症状，时间长了会发展成痢疾。

---2010 年云南教育出版社出版线装经典《全本黄帝内经》

6、气喘→咳嗽 E→A

风寒闭肺气喘：喘急胸闷，伴有咳嗽，咯痰清稀色白，初起多兼恶寒发热，无汗。

风热犯肺气喘：喘急烦闷，伴有咳嗽，或见发热，汗出恶风，口渴，胸痛。

---2000 年人民卫生出版社出版的姚乃礼主编《中医症状鉴别诊断学》