基于遗传算法的朴素贝叶斯分类研究

摘　要

分类是数据挖掘领域中重要的研究分支，国内外已经取得了许多令人瞩目的成就。朴素贝叶斯分类器由于计算高效、精确度高，并具有坚实的理论基础而得到广泛的应用。然而，朴素贝叶斯分类器的条件独立性假设限制了对实际数据的应用。遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法，具有简单、通用、稳健等特性，使其在复杂实际问题的求解中显示出巨大的优越性，而且能在概率意义下收敛到问题的全局最优解。

本文基于遗传算法，对朴素贝叶斯分类问题进行研究，主要工作如下：

（1） 概述数据挖掘的研究背景，数据挖掘的主要任务，描述了数据挖掘中分类问题的定义、方法以及分类模型评价的标准等。

（2） 描述了朴素贝叶斯分类模型，朴素贝叶斯分类模型的一般原理，以及存在的问题。

（3） 阐述了遗传算法的基本思想，并描述了遗传算法的一种改进算法即自适应遗传算法。

（4） 将遗传算法引入到朴素贝叶斯分类研究中，提出一种基于遗传算法的朴素贝叶斯分类算法（G_NBC），该算法为避免数据预处理时，训练集的噪音及数据规模过大使属性约简的效果不太理想，并进而影响分类效果的问题，在训练集上通过随机属性选取生成若干属性子集，并以这些子集构建相应的朴素贝叶斯分类器，进而采用遗传算法进行优选。实验表明了该算法的有效性。

关键词： 数据挖掘，分类，朴素贝叶斯，遗传算法

The Research of Naïve Bayesian Classification Based on Genetic Algorithms

Abstract

The classification is an important research branch in the data mining domain. It has obtained many amazing achievements. Owing to its highly efficient and highly precise calculation, as well as its strict theoretical foundation, Naïve Bayesian Classifier has obtained widespread application. However, its condition independence assumption limits its real application. The genetic algorithms are one kind of auto-adapted global optimization probability search algorithms which form through simulating the heredity and evolution process of the biology in the natural environment. Its simple, all-purpose, steady character has made great achievements in the solution to difficult, complex problems, and can convergence to the global minimum.

Based on the genetic algorithms, Naïve Bayesian classification method was studied in this dissertation. The main work were as follows:

First, the research background and the primary mission of data mining were outlined, and the definition, the method as well as the classification model appraisal standard of the classification question in the data mining domain were described.

Second, the Naïve Bayesian classification model, the general principle of the model, as well as some existing questions was described.

Third, the basic ideas of the genetic algorithms were elaborated, and one kind of improvement genetic algorithms namely auto-adapted genetic algorithms was described.

Last, by introducing the genetic algorithms to the Naive Bayesian classification research, a Naive Bayesian Classification algorithm based on genetic algorithms (G_NBC for short) was proposed in this article. In order to avoid the effect of the training sets' noise and the data scale causing the influence of feature reduction not to be too approximately ideal, and then effecting the classification influence, this algorithm generates certain attribute subsets of the training sets through the random attribute selection, and constructs the corresponding Naïve Bayesian classifiers, and then optimizes the Bayesian classifiers by using genetic algorithms. The experiments at the end of this dissertation confirmed the validity of this algorithm.

Key words: Data Mining, Classification, Naïve Bayesian Classifier, Genetic Algorithms

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注和致谢的地方外，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。

学位论文作者签名：[illegible] 签字日期：06 年 5 月 30 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，允许论文被查阅和借阅。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）

学位论文作者签名：[illegible] 导师签名：[illegible]

签字日期：06 年 5 月 30 日 签字日期：06 年 5 月 30 日

学位论文作者毕业后去向：

工作单位：[illegible] 电话：[illegible]

通讯地址：[illegible] 邮编：[illegible]

致　谢

首先，衷心感谢我的导师胡学钢教授。在论文的写作过程中，导师胡学钢教授给予了我悉心的指导。胡老师从开始的选题到研究方法，他都给予我耐心的指导和反复的启发，在我做课题遇到困难的时候，胡老师通过亲切的聊天给我指明了继续的方向和正确的研究方法。胡老师同时又是一个非常细心的人，论文中多处的修改和对细微处的严谨的要求，不仅仅是对我课题的帮助，这些教诲会给我未来的工作和学习产生很大的影响，让我受益终生。

此外，在我研究生的学习期间中，王浩副院长、吴共庆老师以及给研究生课程班上过课的老师都给予了我有益的指导和帮助，使我学到了许多知识，受益匪浅。在此衷心的感谢各位老师！

另外，还要感谢好友海深及我的同事，在我读书期间对我工作上的帮助，使我能够集中精力完成我的学业。

同时，还要感谢徐勇、胡春玲等人工智能和数据挖掘实验室的同学，感谢他们对我的帮助！

最后要感谢我的爱人程转流对我始终如一的支持，让我可以安心完成自己的学业。

再次感谢所有关心和支持我的老师、同事、朋友、同学和家人，谨以此文作为献给他们的礼物！

作者：胡为成

2006 年 5 月

第一章 绪 论

随着信息技术的快速发展和信息搜集能力的日益提高，产生了海量的数据。这些海量的数据或是以静态的形式存储在企业的物理存储器上，或是不被存储而瞬时出现的动态数据。面对如此丰富的海量数据，传统的数据处理方法和能力已远远不能满足实际的需求。面对日趋激烈的市场竞争，人们需要从这些蕴含着丰富决策信息的数据中抽取能帮助领导进行决策的知识。在需求的强烈驱动下，数据挖掘技术应运而生[1]。本章概述了数据挖掘的概念和研究现状，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

1.1.1 数据挖掘定义及研究背景

进入九十年代，伴随着因特网(Internet)的出现和发展，将整个世界联成一个小小的地球村，人们可以跨越时空地在网上交换数据信息和协同工作。这样，展现在人们面前的已不是局限于本部门，本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋，数据洪水正向人们滚滚涌来。激增的数据背后隐藏着许多重要的信息，人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。于是，一个新的挑战被提了出来:在这被称之为信息爆炸的时代，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率呢?要想使数据真正成为一个公司的资源，只有充分利用它为公司自身的业务决策和战略发展服务才行，否则大量的数据可能成为包袱，甚至成为垃圾。因此，面对“人们被数据淹没，人们却饥饿于知识”[1]的挑战，从数据库中发现知识(Knowledge Discovery in Databases,KDD)及其核心技术——数据挖掘(Data Mining,DM)便应运而生，并得以蓬勃发展，越来越显示出其强大的生命力。

数据挖掘(DM)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[2]。其流程图如图 1-1 所示。这个定义包括以下四个层次的含义：

（1）数据源必须是真实的、大量的、含噪声的；

（2）发现的是用户感兴趣的知识；

（3）发现的知识要可接受、可理解、可运用，最好能用自然语言表达发现结果；

（4）并不是要求发现放之四海皆准的知识，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明，所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的。

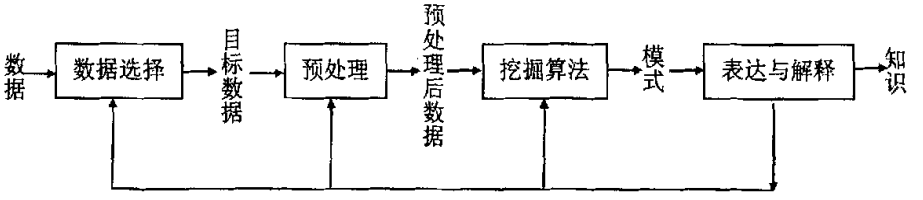

图 1.1 数据挖掘过程

1.1.2 数据挖掘的研究现状

从数据库中发现知识KDD一词首次出现在1989年举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的KDD国际研讨会已经召开了8次，规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。到了1995年，在美国计算机年会(ACM)上，提出了数据挖掘的概念，即通过从数据库中抽取隐含的、未知的、具有潜在使用价值信息的过程。数据挖掘是KDD过程中最为关键的步骤。1997年亚太地区在新加坡组织了第一次规模较大的PAKDD学术研讨会，以后每年召开一次。IEEE的Knowledge and Data Engineering会刊率先在1993年出版了KDD技术专刊，所发表的5篇论文代表了当时KDD研究的最新成果和动态。随后，各类KDD会议、研讨会纷纷涌现，许多领域的国际会议也将KDD列为专题讨论。1999年，IEEE和ACM再次推出KDD专刊，介绍数据挖掘在各个领域的应用成果。此外，并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论。最近，Gartner Group的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。根据最近Gartner的HPC研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。”

目前，国外数据挖掘的发展趋势及其研究方面主要有:对知识发现方法的研究进一步发展，如近年来注重对遗传算法的研究；传统的统计学回归法在 DM 中的应用；DM 与数据库的紧密结合等等。在应用方一面包括:DM 商业软件工具不断产生和完善，注重建立解决问题的整体系统，而不是孤立的过程。用户主要集中在大型银行、保险公司、电信公司和销售业。国外很多计算机公司非常重视数据挖掘的开发应用，IBM 和微软都成立了相应的研究中心进行这方面的工作。许多著名的计算机公司开始开发尝试着 DM 软件的开发，比较典型的如 SAS 公司的 Enterprise Miner, IBM 公司的 Intelligent Miner , SGI 公司的 Set Miner, SPSS 公司的 Clementine，还有 Knowledge Discovery Workbench, DB Miner, Quest 等。Web 数据挖掘产品有 Net percerptions, Accrue Insight 和 Accrue Hit List, Commerce Trend、等。

与国外相比，国内对 KDD 的研究稍晚，目前进行的大多数研究项目是由政府资助进行的，如国家自然科学基金、863 计划、“九五”计划等。1993 年国家自然科学基金开始对数据挖掘研究进行支持。1999 年 4 月在北京召开的第三届亚太地区 KDD 国际会议(PAKDD99)响应热烈，收到论文 158 篇。国内从事数据挖掘研究的人员主要集中在大学，也有部分在研究所或公司。所涉及的研究领域很多，一般集中于学习算法的研究、数据挖掘的实际应用以及有关数据挖掘理论方面的研究。如北京系统工程研究所对模糊方法在数据挖掘中的应用研究、北京大学对数据立方体的研究、华中理工大学、复旦大学、浙江大学等对关联规则的研究等。但到目前为止，国内还没有比较成熟的数据挖掘产品。

当前，DM 研究正方兴未艾，预计今后还会形成更大的高潮，研究焦点可能会集中到以下几个方面:

- 数据挖掘语言的标准化；
- 不完全和不精确的数据的处理:标记数据，计算机处理，存储数据库，创建数据仓库，数据清洁，解决不确定性，格式化数据；
- 数据挖掘过程的可视化，使得知识发现的过程能够被用户理解，也便于在数据挖掘过程中的人机交互；
- 研究在网络环境下的数据挖掘技术，特别是在 Internet 上建立 DM 服务器，与数据库服务器配合，实现数据挖掘；
- 加强对各种非结构化数据的挖掘，如文本数据、图形图像数据、多媒体数据；
- 数据挖掘算法的交叉性研究。

数据挖掘技术的应用首先是满足信息时代用户的急需，因此，研制开发大量基于数据挖掘的决策支持软件工具产品将是首要的任务。

1.1.3 数据挖掘与知识发现

谈到数据挖掘必须提到数据库中的知识发现 KDD (Knowledge Discovery in Databases)。 关于 KDD 与 Data Mining 的关系，有以下几种不同的看法。

(1) KDD 是数据挖掘的一个特例

既然数据挖掘系统可以在关系数据库、事务数据库、数据仓库、空间数据库(Spatial Database)、文本数据(Text Data)以及诸如 WEB 等多种数据组织形式中挖掘知识，那么数据库中的知识发现只是数据挖掘的一个方面，这是早期比较流行的观点，在许多文献可以看到这种说法[3,4,5]，因此从这个意义说，数据挖掘就是从数据库、数据仓库以及其它数据存储方式中挖掘有用知识的过程，这种描述强调了数据挖掘在源数据形式上的多样性。

(2)数据挖掘是 KDD 过程的一个步骤

例如，在 1996 年的知识发现国际会议上，许多学者建议对这两个名词加以区分[5]，核心思想是：KDD 是从数据库中发现知识的全部过程，而 Data dining 则是此全部过程的一个特定的关键步骤。这种观点有它的合理性。虽然我们可以从数据仓库、WEB 等源数据中挖掘知识，但是这些数据源都是和数据库技术相关的，数据仓库是由源数据库集成而来的，即使是像 WEB 这样的数据源，恐怕也离不开数据库技术来组织和存储抽取的信息。因此，KDD 是一个更广义的范畴，它包括数据清洗、数据集成、数据选择、数据转换、数据挖掘、模式生成及评估等一系列步骤，这样我们可以把 KDD 看作是一些基本功能构件的系统化协同工作系统，而数据挖掘则是这个系统中的一个关键的部分。源数据经过清洗和转换等成为适合于挖掘的数据集，数据挖掘在这种具有固定形式的数据集上完成知识的提炼，最后以合适的知识模式用于进一步分析决策工作。从这种狭义的观点上我们可以定义：数据挖掘是从特定形式的数据集中提炼知识的过程。数据挖掘作为 KDD 的一个重要步骤看待，可以使我们更容易聚焦研究，重点有效解决问题。目前人们在数据挖掘算法的研究上基本属于这样的范畴。

(3) KDD 与 Data Mining 含义相同

有些人认为，KDD 与 Data Mining 只是叫法不一样，它们的含义基本相同。事实上，在现今的文献中，许多场合，如技术综述等，这两个术语仍然不加区分地使用着。也有人说，KDD 在人工智能界很流行，Data Mining 在数据库界使用的更多。所以从广义的观点，数据挖掘是从大型数据集(可能是不完全的、有噪声的、不确定性的、各种存储形式的)中挖掘隐含在其中的、人们事先不知道的、对决策有用的知识的过程。从上面的描述中可以看出，数据挖掘概念可以在不同的技术层面上来理解，但是其核心仍然是从数据中挖掘知识，所以有人说叫知识挖掘更合适[1]。

1.1.4 数据挖掘的任务

数据挖掘的任务主要有关联分析、聚类分析、分类、预测、时序模式和偏差分析等[6]。

一、关联分析(Association Analysis)

两个或两个以卜数据项的取值之间存在某种规律性，就称为关联，可以建立起这些数据项的关联规则[7]。数据关联是数据库中存在的一类重要的、可被发现的知识，它反映一个事件和其他事件之间依赖或关联。如果两项或多项属性之间存在关联，那么其中一项的属性值就可以依据其他属性值进行预测。

例如，买面包的顾客中 90%还买牛奶，这就是一条关联规则。在商场中将这两样物品摆放在一起销售，将会提高销售量。

在大型数据库中，这样的关联规则可以产生很多，这就需要进行筛选。一般用“支持度”和“可信度”两个阀值来淘汰那些无用的关联规则。

二、聚类分析(Clustering)

聚类是把数据按照它们的相似性归纳成若干类别，同一类别中的数据距离较小、彼此相似，不同类别中的数据距离偏大、彼此相异[8].聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

聚类方法包括统计分析方法、机器学习方法和神经网络方法等。

在统计分析方法中，聚类分析是基于距离的聚类。这种聚类分析方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分。

在机器学习方法中，聚类是无导师的学习。此时距离是跟据概念的描述来确定的，又称为概念聚类，当聚类对象动态增加时，概念聚类则称为概念形成。

在神经网络中，自组织神经网络方法用于聚类。如 ART 模型、Kohonen 模型等，这是一种无监督学习方法。当给定距离阀值后，各样本按阀值进行聚类。

三、分类(Classification)

分类是数据挖掘中应用得最多得任务。分类就是找出一个类别的概念描述，并用这种描述来构造模型(一般用规则或决策树模式表示)。类别的概念描述代表着这类数据的整体信息，也就是该类的内涵描述[9]。

类的内涵描述分为:特征描述和辨别性描述。特征描述是对类中对象的共同特征的描述。辨别性描述是对两个或多个类之间的区别的描述。

分类的过程是:分析输入数据，通过在训练集中的数据所表现出来的特性，经过有关算法，为每一个类找到一种准确的描述或者模型，并使用这种类的描述对未来的测试数据进行分类。

四、预测(Predication)

预测是利用历史数据找出变化规律，建立模型，并由此模型对未来数据的种类及特征进行预测[8]。

典型的预测方法是回归分析，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。预测时，只要输入任意的时间值，通过回归方程就可求出该时间的状态。

近年来，发展起来的神经网络方法(如 BP 模型)，实现了非线性样本的学习，能进行非线性函数的判别。

分类也能进行预测，但分类一般用于离散数值；回归预测用于连续数值；神经网络方法预测既可以用于连续数值，也可以用于离散数值。

五、时序模式(Time-Series Pattern)

时序模式是指通过时间序列搜索出的重复发生概率较高的模式[10]。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。

在时序模式中，需要找出在某个最小时间内出现比率一直高于某一最小百分比(最小支持度阀值)的规则。这些规则会随着形势的变化作适当的调整。

时序模式中，一个有重要影响的方法是“相似时序”。用“相似时序”的方法，要按时间顺序查看时间事件数据库，从中找出另一个或多个相似的时序事件。

六、偏差分析(Deviation)

数据库中的数据存在很多异常情况，发现数据库中数据存在的异常情况是非常重要的。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。

1.2 数据挖掘中的分类问题

1.2.1 分类的定义

分类就是根据数据集的特点找出类别的概念描述，这个概念描述代表了这类数据的整体信息，也就是该类的内涵描述[9][12]。

分类的目的是：分析输入数据，通过在训练集中的数据所表现出来的特性，为每一个类找到一种准确的描述或者模型。这种描述常常用谓词表示。并使用这种类的描述对未来的测试数据进行分类。尽管这些未来的测试数据的类标签是未知的，我们仍可以由此预测这些新数据所属的类。

分类可描述为：给定一训练数据的集合 T(简称为训练集或训练数据库)，T 中的元素——记录由若干个属性描述。在所有属性中有且仅有一个属性作为类别属性。属性集合用矢量 $X=(X_1, X_2, \ldots, X_n)$ 表示，其中 $X_i(1 \le i \le n)$ 对应各非类别属性，可以具有不同的值域，即对于任一属性 $X_i=\{X_{1i}, X_{2i}\ldots, X_{mi}\}$，mi 随属性的不同而变化。当一属性的值域为连续值域时，该属性称为连续属性(Numerical Attribute)，否则称为离散属性(Discrete Attribute；用 C 表示类别属性，

C={C1,C2..,Ck}，即数据集有 K 个不同的类别。那么，T 就隐含地确定了一个从矢量 X 到类别属性 C 的映射函数 H: f(X)-> C，分类的目的就是采用某种方法(模型)将该隐含函数 H 表示出来。

1.2.2 几种主要的分类模型

分类技术是很多领域，比如统计、模式识别、人工智能、神经网络等领域的研究课题。本节介绍了一些分类算法和知识模型。

一、决策树

决策树方法，在许多的机器学习书或论文中可以找到，这类方法的详细介绍 ID3[1]算法是最典型的决策树分类算法，之后的改进算法包括 ID4, IDS, C4.5[13], C5.0 等。这些算法都是从机器学习角度研究和发展起来的，对于大训练样本集很难适应。这是决策树应用向数据挖掘方向发展必须面对和解决的关键问题。在这方面的尝试也很多，比较有代表性的研究有 Agrawal 等人提出的 SLIQ[11], SPRINT 算法[14]，它们强调了决策树对大训练集的适应性。1998 年，Iichalski 等对决策树与数据挖掘的结合方法和应用进行了归纳[15]，另一个比较著名的研究是 Gehrke 等人提出了一个称为雨林(Rainforest)的在大型数据集中构建决策树的挖掘构架[16]，并在 1999 年提出这个模型的改进算法 BOAT[17]，另外的一些研究，集中在针对数据挖掘特点所进行的高效决策树、裁减决策树中规则的提取技术与算法等方面。

二、贝叶斯分类

贝叶斯分类(Bayesian Classification)来源于概率统计学，并且在机器学习中被很好地研究。近几年，作为数据挖掘的重要方法倍受瞩目。朴素贝叶斯分类(naïve Bayesian Classification)具有坚实的理论基础，和其它分类方法比，理论上具有较小的出错率。但是，由于受其对应用假设的准确性设定的限制，因此需要在提高和验证它的适应性等方面进一步工作。Jone 提出连续属性值的内核稠密估计的朴素贝叶斯分类方法[16]，提高了基于普遍使用的高斯估计的准确性，Domingos 等对于类条件独立性假设应用假设不成立时朴素贝叶斯分类的适应性进行了分析[18]，贝叶斯信念网络(Bayesian Belief Network)是基于贝叶斯分类技术的学习框架，集中在贝叶斯信念网络本身架构，以及它的推理算法研究上，其中，比较有代表性的工作有：Russell 的布尔变量简单信念网、训练贝叶斯信念网络的梯度下降法[19]、Buntine 等建立的训练信念网络的基本操作[20]以及 Lauritzen 等的具有蕴藏数据学习的信念网络及其推理算法、EM[21]等。

三、神经网络分类

神经网络作为一个相对独立的研究分支已经很一早被提出，有许多著作和文献详细介绍了它的原理。由于神经网络需要较长的训练时间和其可解释性较

差，为它的应用带来了困难。但是，由于神经网络具有高度的抗干扰能力和可以对未训练数据进行分类等优点，又使得它具有极大的诱惑力。因此，在数据挖掘中使用神经网络技术是一件有意义但仍需要艰苦探索的工作。在神经网络和数据挖掘技术的结合方面，一些利用神经网络挖掘知识的算法被提出，例如Lu和Setion。等提出的数据库中提取规则的方法[22]，Widro等系统介绍了神经网络在商业等方面的应用技术[23]。

四、类比学习和案例学习

最典型的类比学习（Analogy Learning）方法是k最临近分类(k-Nearest Neighbor Classification)方法，它属于懒散学习法，相比决策树等急切学习法，具有训练时间短，但分类时间长的特点。k 最临近方法可以用于分类和聚类中，基于案例的学习（case-Based Learning）方法可以应用到数据挖掘的分类中。基于案例学习的分类技术的基本思想是，当对一个新案例进行分类时，通过检查已有的训练案例找出相同的或最接近的案例，然后根据这些案例提出这个新案例的可能解。利用案例学习来进行数据挖掘的分类必须要解决案例的相似度、度量训练案例的选取以及利用相似案例生成新案例的组合解等关键问题，并且它们也正是目前研究的主要问题。

五、其它方法

如粗糙集和模糊集(Fuzzy Set)方法等。

另外需要强调的是，任何一种分类技术与算法都不是万能的，不同的商业问题需要用不同的方法去解决。即使对于同一个商业问题可能有多种分类算法。分类的效果一般和数据的特点有关。有些数据噪声大、有缺值、分布稀疏，有些属性是离散的，而有些是连续值的，所以目前普遍认为不存在某种方法能适合于所有特点的数据。因此，对于一个特定问题和一类特定数据，需要评估具体算法的适应性。

1.2.3 分类模型的评价

我们一般从以下几个方面对分类模型的性能进行评价：

（1）预测准确度:预测准确度是分类器的一个重要度量，分类器是已知数据集的描述模型，也可用于对目标数据进行预测。预测准确度可以评价一个分类器对于预测将来数据的准确度，是用得最多的一种比较尺度，特别是对于预测型分类任务，评价分类器的预测准确度有两种常用的方法：保持(holdout)和 K-次交叉验证(k-fold cross-validation)，还有其他的方法也可用于评价分类器的准确度，如 bootstrapping, leave-one-out 等等。

（2）计算复杂度:计算复杂度依赖于算法的实现细节与机器的硬件环境。在数据挖掘中，目标数据大多是大型数据库，数据规模也越来越大。因此，时空性能将是非常重要的一个因素。

（3）模型描述的简洁度与可解释性:对于描述型的分类任务，模型描述越简洁越受欢迎，因为分类器是通过特定算法挖掘出的模式，这些模式最终是面向特定领域用户的，所以挖掘出的模式要易于理解，便于用户进行决策。

（4）健壮性:健壮性是衡量分类器优劣的一个方面，也是分类器抗干扰能力的度量。现实生活中的数据总会存在噪音数据，而对存在噪音数据的数据集，能否得出好的分类器以及给出正确的预测类别，这一点也很重要。

（5）可伸缩性:目前大部分的分类算法是基于数据规模很小的假定。算法的可伸缩性意味着当给定大数据量能否有效的构造模型。

总之，分类的效果一般与应用领域背景以及数据的特点有关。目前，还没有发现一种对于所有的数据集都最优的方法。也就是说，不存在某种方法能适用任何应用，适合各种特点的数据。

1.3 本文的内容组织

本文由五章组成:

第一章主要简述数据挖掘产生的原因以及研究背景，描述了数据挖掘的概念及其与 KDD 的关系，介绍了数据挖掘的主要任务；详细阐述了数据挖掘中分类问题的定义、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。

第二章对贝叶斯分类技术的原理和方法作了全面的介绍:首先给出了贝叶斯分类的原理，即贝叶斯定理，贝叶斯定理是贝叶斯分类技术的理论基础，所有的贝叶斯分类模型都是基于贝叶斯定理的。接着重点讨论了朴素贝叶斯分类模型，给出了朴素贝叶斯分类模型的一般原理，并在此基础上的一系列改进；最后，简要介绍了贝叶斯信念网络。

第三章详细介绍了遗传算法的基本思想，重点讨论实现遗传算法所涉及的几个主要因素：参数的编码，选择、交叉、变异等遗传操作，适应度函数的设计；进一步介绍基本遗传算法的表示和过程，最后简要介绍遗传算法的一种改进：自适应遗传算法。

第四章给出了本文的核心内容，在对前面理论总结研究的基础上，将遗传算法和贝叶斯分类两种软计算方法相结合，提出了一种基于遗传算法的朴素贝叶斯分类方法（G_NBC）。然后选择 UCI 机器学习数据库提供的典型数据库实例，通过实验对 G_NBC 算法和 NBC 算法进行了比较。

第五章对已有的工作进行总结，并对下一步的工作进行了展望。

第二章　贝叶斯理论与贝叶斯分类模型

贝叶斯分类器是建立在经典的贝叶斯概率理论的基础上的基于统计方法的分类模型[24,25]。本章主要讨论贝叶斯分类的基本原理和几种常见的贝叶斯分类模型。

2.1　贝叶斯分类器的一般原理

贝叶斯学派是现代统计学中与经典频率学派并列的两大学派之一，贝叶斯数据分析就是先验分布在经过数据提供的证据修订之后所形成的后验分布。Tomas Bayes 在 1763 年提出了后来以他名字命名的贝叶斯理论：在具体行动之前，无论决策是如何制定的，在结果的证据收集并确认后，决策是可以改变的。

2.1.1　贝叶斯理论

经典的频率统计学派对具有不确定性的未知参数 θ 的推断直接利用样本信息，而与具体的应用领域无关。因此，认为总体(研究对象的全体)X 的概率分布——概率密度或概率分布率 f(x; θ)中的未知参数 θ 是一个确定的数，完全由样本集决定。与经典统计学的客观概率不同，贝叶斯概率是观测者对某一事件的发生的相信程度。贝叶斯理论认为 θ 除了与样本信息相关外，还与来自于非样本信息的先验信息相关。先验信息一半来自包含类似 θ 的过去的经验，先验信息具有一定的主观性。因此 f(x; θ)变为条件分布，记为 f(x| θ)。贝叶斯理论正式地将先验信息纳入统计学并利用这种主观信息进行推断。

2.1.2　贝叶斯定理

设试验 E 的样本空间为 Ω，　A、B 为 Ω 的事件，事件 A 发生的概率记为 P(A)，事件 B 发生的概率记为 P(B)，事件 A 事件 B 同时发生的概率记为 P(AB)，在 A 已经发生的条件下，B 发生的概率称为 A 发生的条件下事件 B 发生的条件概率，记为：

$$p(B|A)=\frac{p(AB)}{p(A)} \qquad (2.1)$$

无论事件 A、B 是否是相互独立的事件，下式显然成立：

P(AB)=P(B)P(A|B)=P(A)P(B|A)　　称为概率乘法定理　　(2.2)

假设 $B_1, B_2 \cdots B_n$ 是样本空间 Ω 的一个划分，即满足：

（1）B_i两两互斥，$B_iB_j=\Phi(i\neq j)$；（2）$\sum Bi=\Omega$（i=1,2, …n）则：

$P(A)=P(A\cap\Omega)=P(A\cap\Sigma B_i)=P(\Sigma AB_i)=\Sigma P(AB_i)=\Sigma P(B_i)P(A|B_i)$

---------称为全概率公式　　(2.3)

在式(2-3)中$P(B_i)$是由以前的分析得到的，因此称为先验概率，而$P(A|B_i)$是根据新得到的信息(B_i的信息)重新加以修正的概率，因此成为后验概率。

条件概率P(A|B)说明事件B发生时事件A的概率。相反的问题就是计算逆概率，即当事件A发生时，事件B发生的概率。

根据乘法定理和全概率公式：

$$P(B_i|A)=\frac{P(AB_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{\sum P(B_i)P(A|B_i)}$$

称为贝叶斯公式　　(2.4)

相互独立的随机事件是一系列这样的事件：其中任何一次事件发生的概率，都与此前各事件的结果无关。因此，对于独立随机事件，借助已经发生事件的结果来推测后来事件的概率是不可能的。因此假如事件A、B是相互独立的事件，则有：P(A|B)=P(A)　　(2.5)

所以对于独立事件：P(AB)=P(A|B)P(B)=P(A)P(B)　　(2.6)

2.1.3 极大后验假设与极大似然假设

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为极大后验假设(maximum a posteriori,MAP)，记为h_{MAP}：

$$h_{MAP}=\arg\max P(h|D)=\arg\max P(D|h)/P(D)=\arg\max P(D|h)P(h)\quad(2.7)$$

由于P(D)是不依赖于h的常量，所以在最后一步去掉了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假设为前提的。

在某些情况下，可假定H中每个假设有相同的先验概率(即对H中任意的h_i和h_j，$P(h_i)=P(h_j)$)。这时可把(2.7)式进一步简化，只考虑P(D|h)来寻找极大可能假设。P(D|h)常被称为给定h时数据D的似然度(likelihood)，任何使P(D|h)最大的假设称为极大似然假设(maximum likelihood, ML)记为：h_{ML}

$$h_{ML}\equiv\arg\max P(D|h)\qquad(2.8)$$

在分类过程中，(2.8)式常被用来在启发式搜索时进行模型检测。

2.2 朴素贝叶斯分类模型

2.2.1 朴素贝叶斯分类原理

朴素贝叶斯分类器假定特征向量的各分量间相对于决策变量是相对独立的，并使用概率规则来实现学习或某种推理过程，即将学习或推理的结果表示为随机变量的概率分布，这可以解释为对不同可能性的信任程度。它的理论基础就是贝叶斯定理和贝叶斯假设[26,27]。

朴素贝叶斯分类器将每个训练样本数据分解成一个 n 维特征向量 X 和决策类别变量 C，并假定特征向量的各分量间相对于决策变量是相对独立的。

设特征向量 $X=\{x_1, x_2, \cdots x_n\}$ 表示数据 n 个属性（A_1，A_2，$\cdots A_n$）的具体取值，类别变量 C 有 m 个不同的取值 C_1，C_2，$\cdots C_m$，即有 m 个不同的类别。则：

$$p(X \mid C_k) = p(x_1, x_2, \ldots x_n \mid C_k) = \prod_{i=1}^{n} p(x_i \mid C_k) \qquad 1 \leqslant k \leqslant m \qquad (2.9)$$

由贝叶斯定理知 X 属于 C_k 的后验概率为：

$$p(C_k \mid X) = \frac{p(X \mid C_k)P(C_k)}{p(X)} \qquad 1 \leqslant k \leqslant m \qquad (2.10)$$

朴素贝叶斯分类器将未知类别的决策变量 X 归属于类别 C_k 当且仅当：

$p(C_k \mid X) > p(C_j \mid X)$ 对于 $1 \leqslant j \leqslant m$，$j \neq k$ 即 $p(C_k \mid X)$ 最大。

由于 $P(X)$ 对于所有类别均是相同的，因此：

$$p(C_k \mid X) \propto p(X \mid C_k)P(C_k) = P(C_K)\prod_{i=1}^{n} p(x_i \mid C_K) \quad 1 \leqslant k \leqslant m \qquad (2.11)$$

由于类别的事前概率是未知的，因此，可以假设各类别出现的概率相同，即 $P(C_1) = P(C_2) = \ldots = P(C_m)$。这样求公式（2.11）的最大转换为求 $p(X \mid C_k)$ 最大，否则就要求 $p(X \mid C_k)P(C_k)$ 得最大。可以通过训练样本数据集合估计 $P(C_k)$ 和 $p(x_i \mid C_k)$（$1 \leqslant i \leqslant n$，$1 \leqslant k \leqslant m$）：

$$P(C_k) = s_k / s \qquad (2.12)$$

$$p(xi \mid C_k) = s_{ki} / s_k \text{。} \qquad (2.13)$$

其中，s_k 为训练样本数据集合中类别为 C_k 的样本个数，s 为整个训练样本数据集合的容量。s_{ki} 为训练样本数据集合中类别为 C_k 且属性 A_i 的取值为 x_i 的样本个数。

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。

2.2.2 朴素贝叶斯分类模型的不足

朴素贝叶斯分类模型中的类条件独立性假设是它的先天不足所在，独立性假设在许多实际问题中并不成立，如果在这些问题中忽视这一点，会引起分类的误差。为了克服这一不足，已有许多相关文献对朴素贝叶斯分类算法作了一些改进，主要是放宽条件独立性的限制。下面将就这些改进方法作一简单的介绍，并提出一种我们自己的改进方法。

2.2.3 朴素贝叶斯分类模型的改进

一、从属性变量间的关系来改进朴素贝叶斯分类器

朴素贝叶斯分类器关于变量独立性的假设虽然大大减少了参数量，但在现实生活中，这种独立性假设经常是不满足的。经过分析得知，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，并且空间复杂度也很高。因此，可以从属性变量间的关系来改进朴素贝叶斯分类器，研究具有较宽松条件限制的贝叶斯网络分类器。

（1）属性分组

适用于属性可以分为独立的子集合的情况。

Kononerko 提出一种采用穷尽搜索的属性分组技术[28]，假定同一组内的属性之间可能是相互依赖的，但组与组之间是满足独立性假设的属性集合。也就是说，独立性假设弱化为这些属性组之间的独立性。但是，这种算法的复杂性要远远高于朴素贝叶斯分类器，而且在现实世界中，属性可以完全被分成独立的子集合只是少数情况。

（2）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[29]。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。

（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[30]，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, LBR)学习技术[31]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 HLBR[32]，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）通过属性约简改善属性间的依赖性

运用粗糙集合理论可以对条件属性集进行约简处理而不改变分类质量，然后结合信息熵理论可以计算出约简后的属性依赖度，从而可以选择一个近似独立的约简后的属性集。这样，既可以满足朴素贝叶斯分类的类条件独立的基本要求，又可以通过约简降低特征维数，缩减求解问题的规模[33]。

二、利用遗传算法对多个朴素贝叶斯分类器进行优选

我们提出一种基于遗传算法的朴素贝叶斯方法，该方法为避免数据预处理时，训练集的噪音及数据规模使属性约简的效果不太理想，并进而影响分类效果，在训练集上通过随机属性选取生成若干属性子集，并以这些子集构建相应的朴素贝叶斯分类器，然后结合遗传算法，将这些朴素贝叶斯分类器作为初始种群，设立适应度函数，采用遗传算子进行优选，这样最后一代的最优秀个体即是要求的分类器。具体方法的介绍将在第四章给出。

2.2.4 朴素贝叶斯分类器的提升

提升方法[26]（Boosting）总的思想是学习一系列分类器，在这个序列中每一个分类器对它前一个分类器导致的错误分类例子给予更大的重视。尤其是，在学习完分类器 H_k 之后，增加了 H_k 导致分类错误的训练例子的权值，并且通过重新对训练例子计算权值，再学习下一个分类器 H_{k+1}。这个过程重复 T 次。最终的分类器从这一系列的分类器中综合得出。

在这个过程中，每个训练例子被赋予一个相应的权值，如果一个训练例子被分类器错误分类，那么就相应增加该例子的权重，使得在下一次的学习中，分类器对该例代表的情况更加重视。

对多类分类问题的提升方法如下：

Input：N 个训练实例：〈（x_1，y_1），…，（x_N，y_N）〉

N 个训练实例上分布 D：w，w为训练实例的权向量。

T 为训练重复的趟数。

1）Initialize;

2）初始化训练实例的权向量。$w_i=1/N$，i=1,…N

3）for t=1 to T

4）给定权值 w_i^t 得到一个假设 $H^{(t)}$：X—>Y

5）估计假设 $H^{(t)}$ 的总体误差，$e^{(t)}=\sum_{i=1}^{N} w_i^{(t)} I(y_i \neq h_i^{(t)}(x_i))$

6）计算 $\beta^{(t)}=\frac{e^{(t)}}{1-e^{(t)}}$

7）计算下一轮样本的权值 $w_i^{(t+1)} = w_i^{(t)}(\beta^{(t)})^{1-I(y_i \neq h_i^{(t)}(x_i))}$

8）正规化 $w_i^{(t+1)}$，使其总和为 1

9）End for

10）Output

11）$h(x) = \arg\max \sum_{t=1}^{T}(\log\frac{i}{\beta^{(t)}})I(h^{(t)}(x) = y)$

这里 $I(\Phi)=1$，如果 $\Phi = T$；否则 $I(\Phi)=0$。

上述提升朴素贝叶斯分类器的时间复杂度是 O(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: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。

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 贝叶斯信念网络

朴素贝叶斯分类假定条件独立，即给定样木的类标号，属性的值相互条件独立。这一假定简化了计算。当假定成立时，与其他所有分类算法相比，朴素贝叶斯分类是最精确的。然而，在实践中，变量之间的依赖可能存在。贝叶斯信念网络(Bayesian belief network)说明链和条件概率分布。它允许在变量的子集间定义类条件独立性。它提供一种因果关系的图形，可以在其上进行学习。这种网络也被称作信念网络、贝叶斯网络和概率网络。为简洁计，我们称它为信念网络。

贝叶斯信念网络由 R. Howard 和 J. Matheson 于 1981 年提出，它是一种概率推理方法，它能从不完全、不精确和不确定的知识和信息中做出推理，可以处理不完整和带有噪音的数据集，从而解决了数据间不一致甚至相互独立的问题。其坚实的理论基础、知识结构的自然表述方式、灵活的推理能力、方便的决策机制使其应用越来越广泛。贝叶斯信念网络将不确定事件以网络的形式连结起来，实现对某一与其它事件有关的时间的预测。与传统的不确定信息模型相比，贝叶斯信念网络建立在严格的统计理论的基础上，因此具有坚实的理论基础。

信念网络由两部分定义。第一部分是有向无环图，其每个节点代表一个随机变量，而每条弧 f 代表一个概率依赖。如果一条弧由节点 Y 到 Z，则 Y 是 Z 的双亲或直接前驱，而 Z 是 Y 的后继。给定其双亲，每个变量条件独立于图中的非后继。变量可以是离散的或连续值的。它们可以对应于数据中给定的实际属性，或对应于一个相信形成联系的“隐藏变量”(如影响学生的身体健康的综合因素)。

图 3.1 给出了一个 6 个布尔变量的简单信念网络。弧表示因果关系。例如，

得肺病的学生受家族肺病史的影响，也受其是否吸烟的影响。此外，该弧还表明:给定其双亲 Family History 和 Smoker，变量 Lung Cancer 条件独立于 Emphysema。这意味，一旦 Family History 和 Smoker 的值已知，变量 Emphysema 并不提供关于 Lung Cancer 的附加信息。

定义信念网络的第二部分使每个属性一个条件概率表(CPT)。变量 Z 的 CPT 说明条件分布 P(Z|Parents(Z))，其中 Parents(Z)是 Z 的双亲。表 3.1 给出了 Lung Cancer 的 CPT。

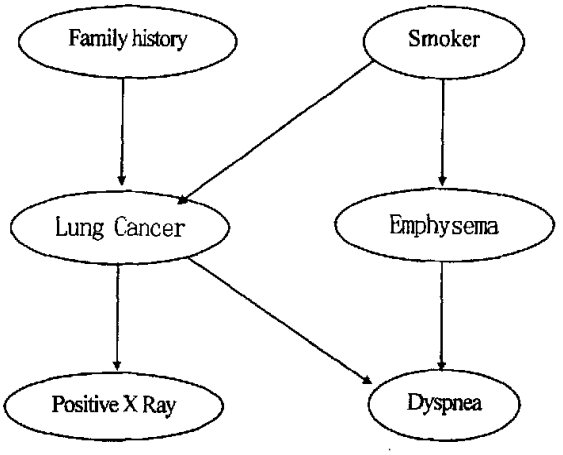

图 2.1　　一个简单的贝叶斯信念网络

表 2.1 条件概率表

Result	FH,S	FH,~S	~FH,S	~FH,~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

对于其双亲值的每个可能组合，表中给出了 Lung Cancer 的每个值的条件概率。例如，有左上角和右下角，我们分别看到

P(LungCancer=“yes”|EamilyHistoty=″yes”,Smoker=“yes”)=0.8

P(Lungcancer=“no”|FamilyHistory=“no”,Smoker=“no”)=0.9

对应属性或变量(Z1,…,Zn)的联合概率由下式计算。其中，P(Zi|parents(Zi))的值对应于 Zi 的 CPT 中的表目。

$$P(Z_1, \cdots, Z_n) = \prod_{i=1}^{n} P(Z_i \mid parents(Z_i)) \qquad (2.14)$$

网络内节点可以选作“输出”节点，代表类标号属性。可以有多个输出节点。学习推理算法可以用于网络。分类过程不是返回单个类标号，而是返回类标号属性的概率分布，即预测每个类的概率。

对于贝叶斯信念网络的学习，也就是找出一个能够最真实地反映现有数据集中各数据变量之间的依赖关系的贝叶斯网络模型。给定离散变量集 $X=\{X_1, X_2, \ldots X_n\}$ 上的数据样本 D，学习的目的是为找到和 D 匹配程度最高的贝叶斯信念网络。贝叶斯信念网络由两部分组成，因此贝叶斯信念网络的学习可以被分解成结构学习和概率分布学习(也称为参数学习)两部分。

贝叶斯信念网络的学习通过学习将描述系统行为的联合概率分布转化为一组条件概率的乘积。贝叶斯方法具有综合先验知识的增量学习特性。贝叶斯信念网络本身没有输入和输出的概念，各节点的计算是独立的，因此，贝叶斯网络学习既可以由上级结点向下级结点推理，也可以由下级结点向上级结点推理。

2.4 本章小结

本章对贝叶斯分类技术的原理和方法作了全面的介绍:首先给出了贝叶斯分类的原理，即贝叶斯定理，贝叶斯定理是贝叶斯分类技术的理论基础，所有的贝叶斯分类模型都是基于贝叶斯定理的。接着重点讨论了朴素贝叶斯分类模型，给出了朴素贝叶斯分类模型的一般原理，并在此基础上的一系列改进；最后，简要介绍了贝叶斯信念网络。

第三章　遗传算法的原理与方法

3.1 遗传算法的基本思想

遗传算法是从代表问题可能潜在解集的一个种群(population)开始的，而一个种群则由经过基因(gene)编码(coding)的一定数目的个体(individual)组成。每个个体实际上是染色体(chromosome)带有特征的实体。染色体作为遗传物质的主要载体，即多个基因的集合，其内部表现(基因型)是某种基因组合，它决定了个体的性状的外部表现，如黑头发的特征是山染色体中控制这一特征的某种基因组合决定的。因此，在一开始需要实现从表现型到基因型的映射即编码土作。由于仿照基因编码的工作很复杂，我们往往进行简化，如二进制编码、浮点数编码。初代种群产生后，按照适者生存和优胜劣汰的原理，逐代(generation)演化产生出越来越好的近似解。在每一代，根据问题域中个体的适应度(fitness)大小挑选(selection)个体，并借助于自然遗传学的遗传算子(genetic operators)进行组合交叉(crossover)和变异(mutation)，产生出代表新的解集的种群。这个过程将导致种群象自然进化一样的后代种群比前代更加适应于环境，末代种群中的最优个体经过解码(decoding)，可以作为问题近似最优解[34]。

3.2 遗传算法的操作

3.2.1 编码

编码是应用遗传算法时所要解决的首要问题，也是设计遗传算法时的一个关键步骤。在遗传算法执行过程中，对不同的具体问题进行编码，编码的好坏直接影响选择、交叉、变异等遗传运算。把一个问题的可行解从其解空间转换到遗传算法搜索空间的转换方法就称为编码，而由遗传算法解空间向问题空间的转换称为解码(或译码)。

遗传算法的编码就是解的遗传表示，它是应用遗传算法求解问题的第一步。编码的准则为：(1)有意义的积木块编码规则，即所定编码应当易于生成与所求问题相关的短距和低阶的积木块；(2)最小字符集编码规则，即所定编码应采用最小字符集以使问题得到自然的表示和描述。针对一个具体应用问题，如何设计一种完美的编码方案一直是遗传算法应用的难点之一，也是遗传算法的一个重要研究方向。由于遗传算法应用的广泛性，迄今为止人们已经提出了许多种不同的编码方法，总的来说，可以分为三大类：二进制编码方法、符号编码方法

和浮点数编码方法。

一、二进制编码方法

二进制编码是遗传算法中最主要的一种编码方法，它使用的编码符号集是由二进制符号 0 和 1 所组成的二值符号集{0，1}，它所构成的个体基因型是一个二进制编码符号串，二进制编码符号串的长度与问题所要求的求解精度有关。

二进制编码有如下优点：

(1)编码、解码操作简单易行。

(2)交叉、变异等遗传操作便于实现。

(3)符合最小字符集编码原则。

(4)便于利用模式定理对算法进行理论分析，因为模式定理是以二进制编码为基础的。

二进制编码有以下缺点：

首先，二进制编码存在着连续变量离散化时的映射误差。个体编码串的长度较短时，可能达不到精度要求，而个体编码串的长度较大时，虽能提高编码精度，但却会使遗传算法的搜索空间急剧扩大。其次它不能直接反映所求问题本身的结构特征，这样也就不便于开发针对问题的专门的遗传运算算子，很难满足积木块编码原则。

二、浮点数编码方案

对于一些多维、高精度要求的连续函数优化问题，使用二进制编码来表示个体时将会有一些不利之处。所谓浮点数编码方法，是指个体的每个基因值用某一范围内的一个浮点数来表示，个体的编码长度等于其决策变量的个数。因为这种编码方法使用的是决策变量的真实值，所以浮点数编码方法也叫做真值编码方法。

浮点数编码方法有以下几个优点：

（1）适合于在遗传算法中表示范围较大的数。

（2）适合于精度要求较高的遗传算法。

（3）便于较大空间的遗传搜索。

（4）改善了遗传算法的计算复杂性，提高了运算效率。

（5）便于遗传算法与经典优化方法的混合使用。

（6）便于设计针对问题专门知识的知识型遗传算子。

（7）便于处理复杂的决策变量约束条件。

三、多参数级联编码

一般常见的优化问题中往往含有多个决策变量。对这种含有多个变量的个体进行编码的方法就称为多参数编码方法。多参数编码的一种最常用的和最基本的方法是：将各个参数分别以某种编码方法进行编码，然后再将它们的编码按照一定的顺序连接在一起就组成了表示全部参数的个体编码。这种编码方法称

为多参数级联编码方法。在进行多参数级联编码时，每个参数的编码方式可以是二进制编码方法，格雷码、浮点数编码或符号编码方式中的一种，每个参数可以具有不同的上下界，也可以有不同的编码长度或编码精度。

其他常用的编码技术有格雷码，多参数交叉编码、一维染色体编码、二维染色体编码、可变染色体编码和树结构编码。

3.2.2 选择

选择(Selection)又称为复制(Reproduction)，是在群体中选择生命力强的个体产生新的群体过程。遗传算法使用选择算子(又称为复制算子，Reproduction Operator)来对群体中个体进行优胜劣汰操作:根据每个个体的适应度大小做出选择，适应度较高的个体被遗传到下一代群体的概率较大，适应度较低的个体被遗传到下一代群体中的概率较小。这样就可以使群体中个体的适应度值不断接近最优解。选择操作建立在对个体的适应度进行评价的基础之上。选择操作的主要目的是为了避免丢失有用的遗传信息，提高全局收敛性和计算效率。选择算子的好坏，直接影响到遗传算法的计算结果。选择算子确定不当，会造成群体中相似度值相近的个体增加，使得后代个体与父代个体相近，导致进化停止不前;或者使适应度值偏大的个体误导群体的发展方向，使遗传失去多样性，产生早熟问题。

选择操作就是确定从父代群体中选取哪些个体遗传到下一代群体中的一种遗传运算，选择操作不但确定用来交叉(或重组)的个体，还确定被选个体产生后代个体的个数。选择操作的策略与编码方式无关。选择操作算子包括:轮盘赌选择、随机竞争选择、最佳保留选择、均匀排序选择、随机联赛选择、最优保存策略、无回放随即选择、确定性选择和排挤选择等。下面介绍几种常用的选择算子:

一、轮盘赌选择

轮盘赌选择方法(Roulette Wheel Selection)是一种回放式随机采样方法，所有选择是从当前种群中根据个体的适应度值，按某种准则挑选出好的个体进入下一代种群。选择算子有多种，经典遗传算法中常采用的是轮盘赌(Wheel，或是比例选择 Proportional Selection)选择方法，每个个体进入下一代的概率就等于它的适应度值与整个种群中个体适应度值总和的比例。适应度值越高，被选中的可能性就越大，进入下一代的概率就越大。每个个体就像圆盘中的一个扇形部分，扇面的角度和个体适应度值成正比，随机拨动圆盘，当圆盘停止转动时指针所在的扇面对应的个体被选中，轮盘赌式的选择方法由此得名。由于群体规模有限和随机操作等原因，使得个体实际被选中的次数与它应该被选中的期望值之间可能存在着一定的误差,因此这种选择方法的选择误差比较大，有时甚至连适应度较高的个体也选不中。

二、随机竞争选择

随机竞争(Stochastic Tournament)选择与轮盘赌选择基本一样，在随即竞争选择中，每次按轮盘赌选择机制选取一对个体，然后让这两个个体进行竞争，适应度高的被选中，如此反复，直到选满为止。

三、最佳保留选择

首先按轮盘赌选择方法执行选择操作，然后将当前群体中适应度最高的个体结构完整地复制到下一代群体中。其主要优点是能保证遗传算法终止时得到的最后结果是历代出现过的最高适应度的个体。

四、均匀排序选择

排序(Ranking)选择算法的思想是:对群体中的所有个体按其适应度大小进行排序，基于这个排序来分配各个个体被选中的概率。其具体操作过程是:

(1)对群体中的所有个体按其适应度大小进行降序排序。

(2)根据具体求解问题，设计一个概率分配表，将各个概率值按上述排列次序分配给各个个体。

(3)以各个个体所分配到的概率值作为其能够被遗传到下一代的概率,基于这些概率值用比例选择的方法来产生下一代群体。各个个体被选中的概率只与个体适应度之间的大小次序有关，而与其适应度的具体数值无直接关系。

五、最优保持策略

在遗传算法中，通过对个体进行交叉、变异等遗传操作而不断产生出新的个体。虽然随着群体的进化过程会产生出越来越多的优良个体，但由于选择、交叉、变异等操作的随机性，它们也有可能破坏掉当前群体中适应度最好的个体。这将会降低群体的平均适应度，并且对遗传算法的运行效率、收敛性都有不利的影响，所以，我们希望适应度最好的个体要尽可能保留到下一代中。为达到这个目的，可以使用最优保持策略进化模型来进行优胜劣汰操作，即当前群体中适应度最高的个体不参与交叉和变异运算，而是用它来替换掉本代群体中经过交叉、变异操作后产生的适应度最低的个体。

六、随机联赛选择

联赛选择方法的基本思想是:每次选取几个个体中适应度最高的一个个体遗传到下一代群体中，每次进行适应度大小比较的个体数目称为联赛规模。一般情况下，联赛规模 N 的取值为 2。具体操作过程是:

(1)从群体中随机选取 N 个个体进行适应度大小的比较,将其中适应度高的个体遗传到下一代群体中。

(2)将上述过程重复 M 次，就可得到下一代群体中的 M 个个体。

3.2.3 交叉

遗传算法中使用交叉算子来产生新的个体。交叉(Crossover)又称重组

(Recombination)，是指对两个相互配对的染色体按某种方式相互交换器部分基因，从而形成两个新的个体。交叉运算是遗传算法区别与其他进化算法的重要特征，它在遗传算法中起着关键作用，是产生新个体的主要方法。

遗传算法中，在交叉运算之前还必须对群体中的个体进行配对。目前常用的配对算法策略是随机配对，即将群体中的 M 个个体以随机的方式组成[M/2]对配对个体组，交叉操作是在这些配对个体组中的两个个体之间进行的。交叉操作算子包括:单点交叉、两点交叉、均匀交叉、算术交叉、顺序交叉和循环交叉等。下面介绍几种最常用的交叉算子:

一、单点交叉

单点交叉(One-point Crossover)又称为简单交叉，是指在个体编码串中只随机设置一个交叉点，然后在该点相互交换两个配对个体的部分染色体。单点交叉的具体执行过程如下:

(1)个体进行两两随机配对，若群体大小为 M，则共有[M/2]对相互配对的个体组。

(2)对每一对相互配对的个体，随机设置某一基因座之后的位置为交叉点，若染色体的长度为 N，则共有 N-1 个可能交叉点位置。

(3)对每一对相互配对的个体，依设定的交叉概率在其交叉点处相互交换两个个体的部分染色体，从而产生出两个新的个体。

二、两点交叉

两点交叉(Two-point Crossover)是指在个体编码串中随机设置了两个交叉点，然后再进行部分基因交换。具体操作过程如下:

(1)在相互配对的两个个体编码串中随机设置两个交叉点。

(2)交换两个个体在所设定的两个交叉点之间的部分染色体。

三、算术交叉

算术交叉(Arithmetic Crossover)是指由两个个体的线性组合而产生出两个新的个体。为了能够进行线性组合运算，算术交叉的操作对象一般是由浮点数编码所表示的个体。假设在两个个体之间进行算术交叉，则交叉运算后所产生的两个新个体为:

$$\begin{cases} X_A^{t+1} = \alpha X_B^t + (1-\alpha) X_A^t \\ X_B^{t+1} = \alpha X_A^t + (1-\alpha) X_B^t \end{cases} \tag{3.1}$$

其中，α 为一个参数，α 可以是一个常数(此时所进行的交叉运算称为均匀算术运算)，α 也可以是一个由进化迭代次数所决定的变量(此时所进行的交叉运算称为非均匀算术运算)。算术交叉的主要操作过程为:

(1)确定两个个体进行线性组合时的系数 α。

(2)根据公式(3.1)生成两个新个体。

遗传算法的收敛性主要取决于其核心操作交叉算子的收敛性。由交叉算子的搜索能力，可以得出结论:只在交叉算子的作用下，随着演化代数的增加，模式内部的各基因将趋于独立，并且只要组成模式的各基因都存在，则该模式一定能被搜索到,此时模式的极限概率等于组成该模式各基因的初始概率的乘积，并且与模式的定义距无关，从而说明了交叉算子能使群体分布扩充的特性。

3.2.4 变异

遗传算法中所谓的变异运算，是指将个体染色体编码串中的某些基因座上的基因值用该基因座的其他等位基因来替换，从而形成一个新个体。从遗传运算过程中产生新个体的能力方面来说，变异本身是一种随机算法，但与选择和交叉算子结合后，能够避免由于选择和交叉运算而造成的某些信息丢失，保证遗传算法的有效性。交叉运算是产生新个体的主要方法，它决定了遗传算法的全局搜索能力;而变异运算只是产生新个体的辅助方法,但它也是必不可少的一个步骤，因为它决定了遗传算法的局部搜索能力。交叉算子和变异算子相互配合，共同完成对搜索空间的全局搜索和局部搜索，从而使得遗传算法能够以良好的搜索性能完成最优化问题的寻优过程。使用变异算子的主要目的有两个:(1)改善遗传算法的局部搜索能力;(2)维持群体的多样性，防止出现早熟现象。变异算子包括:基本位突变、均匀变异、边界变异、非均匀变异和高斯近似变异等。最常用的变异算子如下:

一、基本位变异

基本位变异(Simple Mutation)操作是指对个体编码串中以变异概率随机制定的某一位或某几位基因座上的值作变异运算，操作过程如下:

(1)对个体的每一个基因座，以变异概率指定其为便一点。

(2)对每一个指定的变异点，对其基因值作取反运算或其他等位基因值来代替，从而产生出新一代的个体。

二、均匀变异

均匀变异(Uniform Mutation)操作是指分别用符合某一范围内均匀分布的随机数，以某一较小的概率来替换个体编码串中各个基因座上的原有基因值。操作过程如下:

(1)依次指定个体编码串中的每个基因座为变异点。

(2)对每一个变异点，以变异概率从对应基因的取值范围内取一随机数来替代原有值。

均匀变异操作特别适合应用于遗传算法的初级运行阶段，它使得搜索点可以在整个搜索空间内自由地移动，从而可以增加群体的多样性，使算法处理更多的模式。

三、高斯近似变异

高斯变异(Gaussian Mutation)是改进遗传算法对重点搜索区域的局部搜索性能的另一种变异操作方法。所谓高斯变异操作，是指进行变异操作使用符合正态分布的一个随机数来替换原有的基因值。由正态分布的特性可知，高斯变异也是重点搜索原个体附近某个局部区域。高斯变异的具体操作过程与均匀变异类似。

3.2.5 适应度函数

适应度函数也称为评价函数，是根据目标函数确定的用于区分群体中个体好坏的标准，是算法演化过程的驱动力，也是进行自然选择的唯一依据。适应度函数总是非负的，任何情况下都希望其值越大越好。评价个体适应度的一般过程为：

（1）对个体编码串进行解码处理后，可得到个体的表现型。

（2）由个体的表现型可计算出对应个体的目标函数值。

（3）根据最优化问题的类型，由目标函数值按一定的转换规则求出个体的适应度。

3.3 遗传算法的表示

要利用遗传算法解决问题，就必须将其表示成便于处理的形式，这里介绍一下基本遗传算法(Simple GA, SGA)的表示。SGA 可定义为一个 8 元组：

SGA=(C, E, p0, M, Φ, Γ, Ψ, T)

式中各元素代表的意义：

C—个体的编码方法，SGA 使用固定长度二进制符号串的编码方法；

E—个体的适应度评价函数；

p0—初始种群；

M—群体大小，一般取 20～100，根据具体情况会有所变化；

Φ—选择算子，SGA 使用比例选择算子；

Γ—交叉算子 SGA 使用单点交叉算子；

Ψ—变异算 r，SGA 使用基本位变异算子；

T—算法中止条件，一般中止进化代数为 100～500 。

随着遗传算法研究的不断深入，遗传算法的应用范围越来越广，形式也越来越多样，它们使用的编码方法、适应度评价函数、初始种群、群体大小、中止条件以及遗传算子选择、交叉和变异都可能不一样，但基本都是 SGA 表示方法的变形。

3.4 基本遗传算法过程

Procedure 基本遗传算法:

Begin

```
    Initialize   P(0);
    t=0;
    while(t<=T) do
        for i=1 to M do
            Evaluate   fitness to P(t);
        End for
        For i=1 to M
            Select  operation to P(t)
        End for
        For I=1 to M/2  do
            Crossover operation to P(t);
        End for
        For i=1 to M   do
            Mutation operation to P(t);
        End for
        For i=1 to M   do
            P(t+1)=P(t);
        End   for
        t=t+1;
    end while
end
```

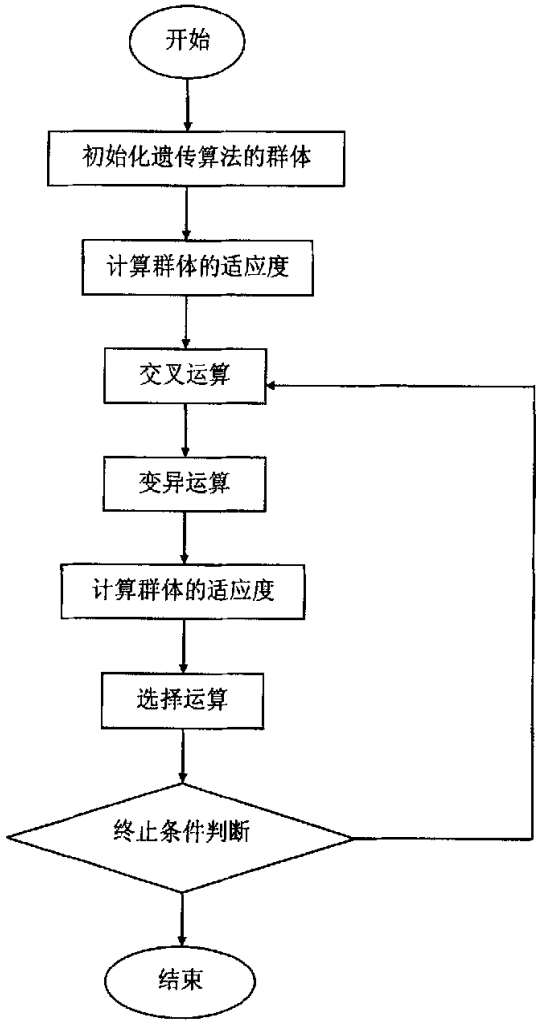

图 3.1　基本遗传算法的流程图

3.5　自适应遗传算法

遗传算法的参数中交叉概率 P_c 和变异概率 P_m 的选择是影响遗传算法行为和性能的关键所在，直接影响算法的收敛性，P_c 越大，新个体产生的速度就越快。然而，P_c 过大时遗传模式被破坏的可能性也越大，使得具有高适应度的个体结构很快就会被破坏；但是如果 P_c 过小，会使搜索过程缓慢，以至停滞不前。对于变异概率 P_m，如果 P_m 过小，就不易产生新的个体结构；如果 P_m 取值过大，那么遗传算法就变成了纯粹的随机搜索算法。针对不同的优化问题，需要反复实验来确定 P_c 和 P_m，这是一件繁琐的工作，而且很难找到适应于每个问题的最佳

值。Srinvivas 等提出一种自适应遗传算法（Adaptive GA，AGA），P_c和P_m能够随适应度自动改变。当种群各个体适应度趋于一致或者趋于局部最优时，使P_c和P_m增加，而当群体适应度比较分散时，使P_c和P_m减少。同时，对于适应度高于群体平均适应度的个体，对应于较低的p_c和p_m，使该解得以保护进入下一代；而低于平均适应度的个体，相对应于较高的p_c和p_m，使该解被淘汰掉。因此，自适应的P_c和P_m能够提供相对某个解的最佳P_c和P_m。自适应遗传算法在保持群体多样性的同时，保证遗传算法的收敛性。

在自适应遗传算法中，P_c和P_m按如下公式进行自适应调整：

$$P_c=\begin{cases}\dfrac{k_1(f_{\max}-f')}{f_{\max}-f_{avg}}, & f\geq f_{avg}\\ k_2, & f<f_{avg}\end{cases} \tag{3.2}$$

$$P_m=\begin{cases}\dfrac{k_3(f_{\max}-f)}{f_{\max}-f_{avg}}, & f\geq f_{avg}\\ k_4, & f<f_{avg}\end{cases} \tag{3.3}$$

式中，$f_{\max}$——群体中最大的适应度值；

f_{avg}——每代群体的平均适应度值；

f'——要交叉的两个体中较大的适应度值；

f——要变异个体的适应度值。

这里，只要设定k_1，k_2，k_3，k_4取（0，1）区间的值，就可以自适应调整了。

当适应度值低于平均适应度值时，说明该个体是性能不好的个体，对它就采用较大的交叉率和变异率；如果适应度值高于平均适应度值，说明该个体性能优良，对它就根据其适应度值取相应的交叉率和变异率。可以看出，当适应度值接近最大适应度值时，交叉率和变异率就越小；当等于最大适应度值时，交叉率和变异率的值为零。这种调整方法对于群体处于进化后期比较合适，但对于进化初期不利，因为进化初期群体中的较优的个体几乎处于一种不发生变化的状态，而此时的优良个体不一定是优化的全局最优解，这容易使进化走向局部最优解的可能性增加。为此，可以做进一步的改进，使群体中最大适应度值的个体的交叉率和变异率不为零。分别提高到P_{c2}和P_{m2}，这就相应地提高了群体中表现优良的个体的交叉率和变异率，使得它们不会处于一种近似停滞不前的状态。为了保证每一代的优良个体不被破坏，采用精英选择策略，使它们直接复制到下一代中。

经过上述改进，P_c和P_m计算表达式如下：

$$p_c = \begin{cases} p_{c1} - \dfrac{(p_{c1} - p_{c2})(f' - f_{avg})}{f_{\max} - f_{avg}}, & f' \geq f_{avg} \\ p_{c1}, & f' < f_{avg} \end{cases} \tag{3.4}$$

$$p_m = \begin{cases} p_{m1} - \dfrac{(p_{m1} - p_{m2})(f - f_{avg})}{f_{\max} - f_{avg}}, & f \geq f_{avg} \\ p_{m1}, & f < f_{avg} \end{cases} \tag{3.5}$$

上式中 P_{c1}=0.9, P_{c2}=0.6, P_{m1}=0.1, P_{m2}=0.001。

3.6 本章小结

遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法，具有简单、通用、稳健等特性。本章详细阐述了遗传算法的基本思想，重点讨论了遗传算法的操作。在分析基本遗传算法表示和过程的基础上，描述了一种改进的遗传算法即自适应遗传算法，这种算法对遗传算法中的交叉算子和变异算子作了改进，使种群中最大适应度个体的交叉概率和变异概率均不为零，使它们在遗传操作中不会处于一种近似停滞不前的状态。

第四章　基于遗传算法的朴素贝叶斯分类研究

本章主要针对第二章提出的朴素贝叶斯分类模型的改进方法作了进一步的讨论，给出了具体的实现方法。此外，对算法进行了性能分析，并给出了相应的实验结果。

4.1　问题的提出

通过前面的研究，发现朴素贝叶斯分类模型在实际应用中常常会碰到以下两个问题：一是为了减小计算规模，朴素贝叶斯分类器是基于条件独立性假设的，但是这个限制过于严格，在实际应用中常常难以满足。二是朴素贝叶斯分类方法所选训练集的条件属性集在预处理时需要进行属性约简，否则即为原始数据库的完全属性集，由于一些属性与分类无关，可能会降低分类能力。而属性约简的好坏会直接影响分类效果。

另外，一个有效的分类器，应当既有很高的分类精度，又使其误差分布在输入空间的不同部分。这就要求在构造分类器时，不但要考虑分类精度，还应考虑分类误差在实例空间中的分布程度，即差异度。

遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法，具有较强的鲁棒性，其思想简单，应用广泛。本章结合遗传算法，提出一种基于遗传算法的朴素贝叶斯分类算法，在训练集上通过随机属性选取生成若干属性子集，并以这些子集构建相应的朴素贝叶斯分类器，进而采用遗传算法进行优选；从而避免了属性约简的好坏对分类精度的影响。

4.2　基于遗传算法的朴素贝叶斯分类方法（G_NBC）

基于以上的研究和分析，本文提出了基于遗传算法的朴素贝叶斯分类算法，算法基本思想如下：

首先考虑给定的信息系统是否完备，如果是不完备信息系统，利用统计学原理对空值属性做出处理，把不完备信息系统完备化；

采用分层随机取样方法将数据库分成训练集和验证集，随机生成S个随机属性子集，并对每一个属性子集构建一个相应的朴素贝叶斯分类器，将S个随机属性子集对应的NBC作为初始种群，采用遗传算法优选。这样，遗传操作的最后一代的最优解即为要求的朴素贝叶斯分类器。

下面我们从理论与实践来分析 G_NBC 算法。

4.2.1 数据预处理

数据预处理是知识获取系统中不可避免的问题，包括原始数据的采样、收集和整理。通常取得的原始数据不一定适合直接用于知识获取，需要进行预处理加工，对于遗失的数据，需要补充，对于值域为实数值的数据，还需要进行离散化处理，下面将对补齐数据和离散化问题进行介绍。

信息系统不确定性的一个主要原因是由于待处理的信息表并不是一个完备的信息表，表中的某些属性值是空值，我们无从知道其原始值。对于这种情况，目前主要通过以下方法来对信息表中的遗漏数据进行补齐。

将存在空属性值的实例记录删除，从而得到一个完备的信息表。

将空属性值作为一种特殊的属性值来处理，不同于其他任何属性值。

采用统计学原理，根据信息表中其余实例在该属性上的取值的分布情况来对一个遗漏属性值进行估计补齐，如Mean Completer算法和Combinatorial Completer算法等。

离散化的本质是利用选取的断点来对条件属性构成的空间进行划分，把n维空间划分为有限个区域，使得每个区域中的对象的决策值相同。假设某个属性有m个属性值，则在此属性上就有m-1个断点可取，随着属性个数的增加，可取的断点数将随着属性值的个数呈几何增长。在选取断点的同时进行属性值的合并，减小问题的复杂度。

目前常用的离散方法有：等距离划分算法、等频率划分算法、适应离散法等等。但总体效果不理想。还有采用结合 Rough 集理论的方法来解决离散化问题，如:Nguyen H. S 和 Skowron 提出的布尔逻辑和 Rough 集理论相结合的离散化算法、基于属性重要性的离散化算法等等，实用效果良好。

4.2.2 分类器差异度的定义

一个有效的分类器，应当既有很高的分类精度，还应考虑分类误差在实例空间中的分布程度，即差异度。差异度的定义如下：

分类精度为R，对数据集中m个类的分类精度分别为R_1，R_2，…，R_m；设数据集的记录个数为P，每一类的记录数分别为P_1，P_2，…P_m；显然，

$$P=P_1+P_2+\cdots+P_m \tag{4.1}$$

$$R=\frac{P_1R_1+P_2R_2+\cdots+P_mR_m}{P} \tag{4.2}$$

则分类器的差异度 D 可以定义为：$D=\frac{R_1R_2\cdots R_m}{R^m}$，由此可知，D 的值介于 0 到 1 之间，值越大，差异性越好。

4.2.3 G_NBC 编码方式

采用传统的二进制编码方式，每条染色体由一组二进制位构成，长度为数据库中随机属性的个数，每个二进制位依次与数据库中的一个属性相对应。若某个二进制位为1，则表示数据库对应的属性参与构建朴素贝叶斯分类器。这样，每个染色体事实上就对应着一个朴素贝叶斯分类器。

4.2.4 G_NBC 适应度函数

适应度通常用来度量群体中各个个体在优化计算中有可能达到或接近于找到最优解的优良程度。适应度函数是用来评估个体的适应度，即区分群体中个体好坏的标准。衡量朴素贝叶斯分类器分类效果除了分类精度要高之外，还应考虑分类误差在实例空间中的分布程度，即差异度。本文适应度函数设为

$$F = R + \lambda D \tag{4.3}$$

式中，R为 NBC 在验证集上的分类精度，D为 NBC 在验证集上的差异度，λ为决定差异度影响的系数。

4.2.5 遗传操作

一、选择操作

在群体中选取优胜的个体，淘汰劣质的个体的操作称作选择。根据染色体适应度值的大小选择适应性更强的染色体生成新的种群。因此适应度值越大，被选中的概率就越大。本算法采用按适应度比例的轮盘赌选择法，其中每个个体被选择的期望数量与其适应值和群体平均适应值的比例有关。首先计算每个个体的适应值，然后计算出此适应值在群体适应值总和中所占的比例，作为该个体在选择过程中被选中的概率。具体方法如下：

对于一个规模为 n 的群体，其中每个个体$a_i(i=1,2,\cdots,n)$的适应度为$f(a_i)$，则个体a_i的选择概率为$\frac{f(a_i)}{\sum_{i=1}^{n} f(a_i)}$。轮盘赌选择的具体实施过程为，将上述得到的个体选择概率按由高到低排序，然后计算他们的累积概率，并产生一个[0,1]之间的随机数，当累积概率大于随机数时，就得到了被选择的个体，这种方法充分体现了“优胜劣汰”的原则。

二、交叉操作和变异操作

交叉运算是指对两个相互配对的染色体按某种方式相互交换部分基因，从

而形成两个新的个体。本算法采用两点交叉，即交换父本两个基因位间的部分，产生相应的后代变异模拟了生物进化过程中的基因突变现象，变异算子是以一定的概率改变遗传基因的操作。对个体进行变异，可以保持群体的多样性，增加了自然选择的余地，并使遗传算法跳出局部极值点。为了避免早熟现象，采用自适应方法动态调节交叉概率和变异概率，使得交叉概率 p_c 变异概率 p_m 能够随适应度自动改变。当种群各个体适应度趋于一致或者趋于局部最优时，使 p_c 和 p_m 增加；而当群体适应度比较分散时，使 p_c 和 p_m 减少。同时，对于适应度高于群体平均适应度的个体，对应于较低的 p_c 和 p_m，使该解得以保护进入下一代；而低于平均适应度的个体，相对应于较高的 p_c 和 p_m，使该解被淘汰掉。选取的交叉概率 p_c、变异概率 p_m 计算表达式如下：

$$p_c = \begin{cases} p_{c1} - \dfrac{(p_{c1} - p_{c2})(f' - f_{avg})}{f_{\max} - f_{avg}}, & f' \geq f_{avg} \\ p_{c1}, & f' < f_{avg} \end{cases} \tag{4.4}$$

$$p_m = \begin{cases} p_{m1} - \dfrac{(p_{m1} - p_{m2})(f_{\max} - f)}{f_{\max} - f_{avg}}, & f \geq f_{avg} \\ p_{m1}, & f < f_{avg} \end{cases} \tag{4.5}$$

式中，$p_{c1} = 0.9$，$p_{c2} = 0.6$，$p_{m1} = 0.1$，$p_{m2} = 0.001$，f' 为要交叉的两个个体中较大的适应度值，f 为要变异个体的适应度值。

三、终止条件

遗传算法迭代过程终止方法一般有以下几种：

1)设定最大代数。该方法简单易行但不准确，需要人工的多次试验、修正。

2)根据群体的收敛程度来判断，即通过计算种群中所有基因位的相似性程度进行控制。

3)在采用精英保留选择策略的情况下，按每代最佳个体的适应值的变化情况确定。

4.2.6 算法描述

算法主要步骤如下：

输入：D={X_1，X_2，…，X_n}，X_i=(A_1^i，A_2^i，…，A_n^i，C_i)，其中，A_1，A_2，…，A_3 为原始属性集，C={C_1，C_2，…，C_m}为类别属性；

输出：样本 X 的类别号；

步骤一：采用分层随机取样方法将数据库分成训练集和验证集；

步骤二：对给定属性集随机生成 S 个随机属性子集，生成相应的 S 个二进进制染色体。对于一个特定的属性子集来说，如果某一属性被选中，则对应的

二进制位为 1，否则为 0，染色体的长度和属性的个数相等；最后构建相应的 S 个朴素贝叶斯分类器，并在分类器和染色体建立一一对应关系；

步骤三：将这 S 个朴素贝叶斯分类器作为初始种群，计算它们的适应度值，然后采用上述的遗传算子对其相应的染色体进行优选；

步骤四：最后一代的最优染色体所对应的朴素贝叶斯分类器即是要求的分类器；

步骤五：将得到的分类器应用于样本 X，即可得到它的类标号；

步骤六：调整 λ 的值，重复步骤三～步骤五。

4.3 G_NBC 算法实验及结果分析

4.3.1 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/MLRepository. html)。我们从中选择了分别如下：breast-cancer, kr-vs-kp, mushroom，vote 等 4 个数据集，其中 mushroom 数据集包含 2480 个空属性值、breast-cancer 数据集包含 9 个空属性值、vote 数据集包含 382 个空值属性。下面对每个数据集进行简要介绍。

(1) breast-cancer

breast-cancer 数据来自于 the University Medical Centre, Yugoslavia. 总共有 286 个实例，9 个条件属性 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，数据集包含 9 个遗失数据。属性信息如下：

1. age {10-19','20-29','30-39','40-49','S0-59','60-69','70-79','80-89','90-99'}
2. menopause {'1t40','ge40','premeno'}
3. tumor-size{'0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39', '40-44','45-49','50-54','55-59' }
4. inv-nodes{'0-2','3-5','6-8','9-11','12-14','15-17','18-20','21-23','24-26','27-29','30-32','33-35','36-39'}
5. node-caps {'yes','no'}
6. deg-rnalig {'1','2','3'}
7. breast {'left','right'}
8. breast-quad{'left_up','left_low' ,'right_up',' right_low','central'}
9. 'irradiat' {'yes','no'}
10. 'Class' {'no-recurrence-events','recurrence-events'}

(2) kr-vs-kp

这是由 Rob Holte 提供的数据，总共有 3196 个实例，36 个条件属性 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，无缺失数据。这是一个下棋的数据，每个实例对应一个属性值序列，是对弈结束时的棋盘描述，前 36 个属性值描述的是棋盘的布局，最后一个属性描述的是白方的胜负情况。

1.' bkblk' {' t ', ' f '}　　2.' bknwy' {' t ', ' f '}
3.' bkon8' {' t ', ' f '}　　4.' bkona' {' t ', ' f '}
5.' bkspr' {' t ', ' f '}　　6.' bkxbq' {' t ', ' f '}
7.' bkxcr' {' t ', ' f '}　　8.' bkxwy' {' t ', ' f '}
9.' blxwp' {' t ', ' f '}　　10.' bxqsq' {' t ', ' f '}
11.' cntxt' {' t ', ' f '}　　12.' dsopp' {' t ', ' f '}
13.' dwipd' {' g ', ' l '}　　14.' hdchk' {' t ', ' f '}
15.' katri' {' b ', ' n ', ' w '}　　16.' mulch' {' t ', ' f '}
17.' qxmcq' {' t ', ' f '}　　18.' r2ar8' {' t ', ' f '}
19.' reskd' {' t ', 'f '}　　20.' reskr' {' t ', ' f '}
21.' rimmx' {' t ', ' f '}　　22.' rkswp' {' t ', ' f '}
23.' rxmsq' {' t ', ' f '}　　24.' simpl' {' t ', ' f '}
25.' skach' {' t ', ' f '}　　26.' skewr' {' t ', ' f '}
27.' skrxp' {' t ', ' f '}　　28.' spcop' {' t ', ' f '}
29.' stlmt' {' t ', ' f '}　　30.' thrsk' {' t ', ' f '}
31.' wkcti' {' t ', ' f '}　　32.' wkna8' {' t ', ' f '}
33.' wknck' {' t ', ' f '}　　34.' wkovl' {' t ', ' f '}
35.' wkpos' {' t ', ' f '}　　36.' wtoeg' {' n ', ' t ', ' f' }
37.' class' {' won ', ' nowin '}

(3) mushroom

这是由 Jeff Schlimmer 提供的数据，总共有 8124 个实例，22 个条件属性 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，有缺失数据。对于属性 11，有 2480 个属性值缺失，用?标记。

1) cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2) cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3) cap-surface: brown=n,buff=b,cinnamon=c,gray=gngreen=r, pink=p,purple=u,red=e,white=w,yellow=y
4) bruises: bruises=t,no=f
5) odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s

6) gill-attachment:	attached=a,descending=d,free=f,notched=n
7) gill-spacing:	close=c,crowded=w,distant=d
8) gill-size:	broad=b,narrow=n
9) gill-color:	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10) stalk-shape:	enlarging=e,tapering=t
11) stalk-root:	bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12) stalk-surface-aboue-ring:	ibrous=f,scaly=y,silky=k,smooth=s
13) stalk-surace-below-ring:	ibrous=f,scaly=y,silky=k,smooth=s
14) stalk-color-above-ring:	brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=,red=e,white=w,yellow=y
15) stalk-color-below-ring:	Brown=n,buff=b,cinnamon=c,gray=g,orange=o, Pink=p,red=e,white=w,yellow=y
16) veil-type:	Partial=p,universal=u
17) veil-color:	Brown=n,orange=o,white=w,yellow=y
18) ring-number:	None=n,one=o,two=t
19) ring-type:	Cobwebby=c,evanescent=e,flaring=f,large=l, None=n,pendant=p,sheathing=s,zone=z
20) spore-print-color:	Black=k,brown=n,buff=b,chocolate=h,green=r, Orange=o,purple=u,white=w,yellow=y
21) population:	Abundant=a,clustered=c,numerous=n, Scattered=s,several=v,solitary=y
22) habitat:	Grasses=g,leaves=l,meadows=m,paths=p, Urban=u,waste=w,woods=d
23) classes:	Edible=e,poisonous=p

(4) 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 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，包含 382 个缺失数据。属性信息如下：

1. Class Name:2(democrat,republican)
2. handicapped-infants:2(y,n)
3. water-project-cost-sharing:2(y,n)
4. adoption-of-the-budget-resolution:2(y,n)
5. physician-fee-freeze:2(y,n)
6. el-salvador-aid:2(y,n)
7. religious-groups-in-school:2(y,n)

8. anti-satelite-test-ban:2(y,n)
9. aid-to-nicaraguan-contras:2(y,n)
10. mx-missile:2(y,n)
11. immigration:2(y,n)
12. synfuels-corporation-cutback:2(y,n)
13. education-spending:2(y,n)
14. superfund-right-to-sue:2(y,n)
15. crime:2(y,n)
16. duty-free-exports:2(y,n)
17. export-administration-act-south-africa:2(y,n)

以上对我们选择的数据集进行了介绍，所有这些数据集的属性取值均符合G_NBC算法和NBC算法对数据集的要求。

4.3.2 实验过程

首先导入数据集，再利用原型系统进行数据补齐和离散化工作，然后将参数设置一下，就可以快速执行或单步执行原型系统得到朴素贝叶斯分类器分类的正确率；最后再将原数据集在不引入遗传算法时进行朴素贝叶斯分类，同样可得到分类的正确率；将这两种结果进行比较，完成实验。

4.3.3 实验结果分析

通过上述方法分别G_NBC和NBC对四个数据集进行测试获取分类器，记录相应的响应时间，并事先采用分层随机取样方法将数据库分成训练集（70%）和验证集（30%），对每个验证集进行测试，得到的正确率如下表所示（表4.1）

表4.1 G_NBC算法与NBC算法分类精度的比较

数据集	记录数	属性数	NBC	G_NBC		
				λ=0	λ=0.5	λ=1
Breast-cancer	286	9	67.84%	71.59%	70.07%	70.45%
Kr-vs-kp	3196	36	84.75%	85.36%	85.17%	84.65%
Mushroom	8124	22	89.49%	91.76%	90.57%	90.86%
vote	435	16	86.45%	87.57%	88.87%	84.98%

为了更加直观，可以用图4.9表示，横坐标表示数据集，纵坐标表示分类精度，四种颜色的柱形分别表示NBC算法和G_NBC算法在λ为0，0.5，1时的分

类精度。

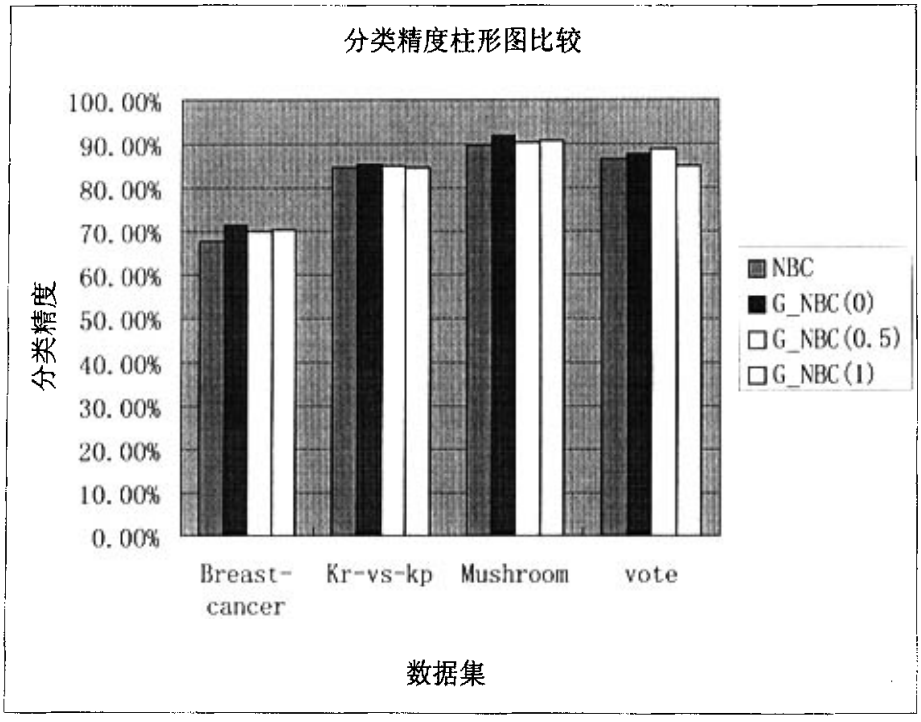

图4.1 分类精度柱形图比较

从图4.9可以看出，在分类精度方面，G_NBC算法普遍优于NBC算法，只有一种情况例外，就是vote数据集G_NBC算法在λ为1时分类精度要低于NBC算法，这与数据集本身的特点有关。总之，影响分类器准确度的因素是多方面的，除了算法本身以外，数据集自身的特点也是影响准确度的一个重要因素，实践已经证明，不存在某种方法能适于各种特点的数据。而且对于同一个数据集而言，在λ值不同时，分类精度也有所不同，比如：在数据集Breast-cancer上，λ值为0时，分类精度最高；在数据集vote上，λ值为0.5时，分类精度最高，这说明，适当地考虑差异度（即分类误差在实例空间中的分布程度）的影响有助于提高分类能力。

在实验中还发现，采用随机属性选择方法，对改进差异度也有较好的效果。

在算法运行时间上，由于G_NBC算法要进行行遗传操作优选，所以耗费进间要略高于NBC算法（如表4.2，时间数据单位：秒），但相对分类精度的提高，这点代价是值得的。

表 4.2 响应时间表

数据集 / 算法	Breast-cancer	Kr-vs-kp	Mushroom	vote
NBC	15	60	78	17
G_NBC(λ=0)	28	109	138	29
G_NBC(λ=0.5)	30	111	145	31
G_NBC(λ=1)	30	111	145	31

以上，通过对 UCI 四个数据集的测试，分析比较了 G_NBC 算法和 NBC 算法的性能，从分类精度来看，G_NBC 算法要普遍优于 NBC 算法；从响应时间来看，两者差异不大； G_NBC 算法还能够较好地改进分类的差异度。

4.4 本章小结

本章在前面几章的基础上，提出了基于遗传算法的朴素贝叶斯分类算法(G_NBC)，给出了 G_NBC 算法的具体描述和算法模型，通过 UCI 机器学习数据集对 G_NBC 算法和 NBC 算法进行了性能比较，从分类的准确度来看，G_NBC 算法比 NBC 算法要好；从响应时间来看，G_NBC 算法和 NBC 算法差别不大；而且 G_NBC 算法还能够较好地改进分类的差异度。

第五章　G_NBC 原型系统

5.1　实验环境

硬件环境：处理器 Intel pentium, 1.73GHZ，内存 512M

软件环境：操作系统 windows XP Professional,

实验工具：G_NBC 原型系统(VB6.0 开发)

5.2　系统功能结构图

首先导入数据集，再利用原型系统进行数据补齐和离散化工作，然后将参数设置一下，就可以快速执行或单步执行原型系统得到朴素贝叶斯分类器分类的正确率；最后再将原数据集在不引入遗传算法时进行朴素贝叶斯分类，同样可得到分类的正确率；将这两种结果进行比较，完成实验。G_NBC 原型系统的功能结构如图 5.1：

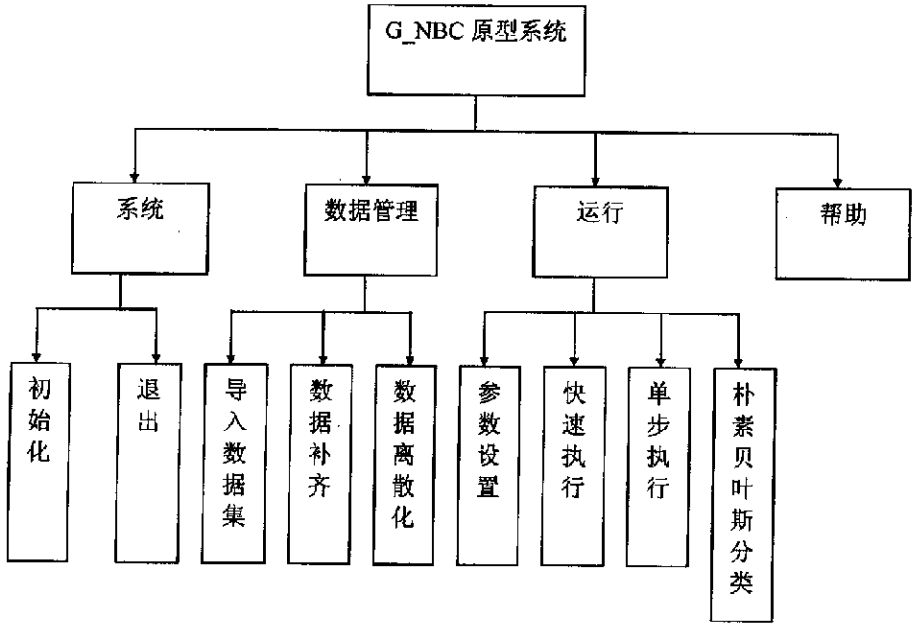

图 5.1 系统功能结构图

图 5.1 系统功能结构图

5.3 实验步骤

1. G_NBC 原型系统主界面

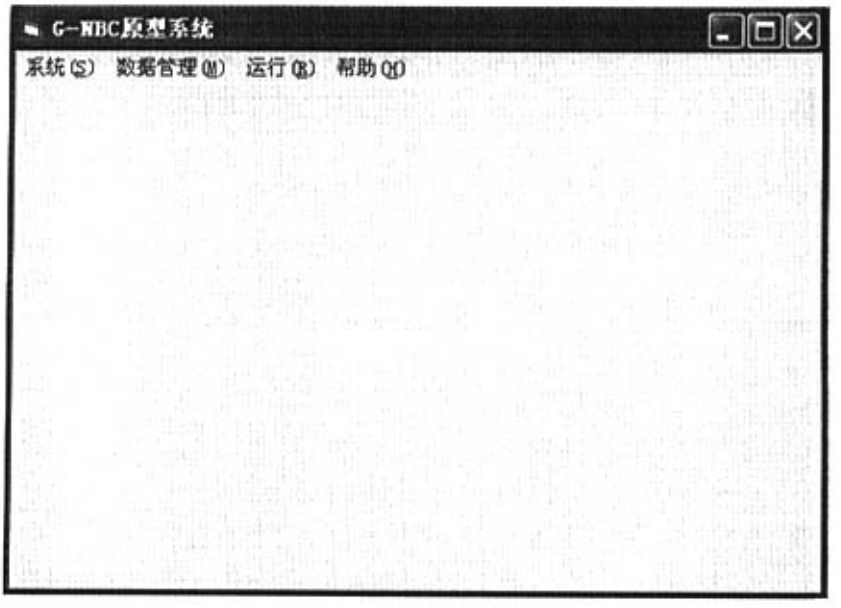

图 5.2 系统主界面

2. 数据预处理

数据预处理包括数据补齐和数据离散化，由于这部分工作无论引入或不引入遗传算法都是要做的，所以数据预处理所花的时间对这两种方法在时间性能上比较没有任何影响。

3. 执行

在系统运行之前先进行有关参数设置。

执行分快速执行、单步执行和朴素贝叶斯分类三种，前两种是引入遗传算法之后的系统运行，第三种是直接对数据集进行朴素贝叶斯分类。

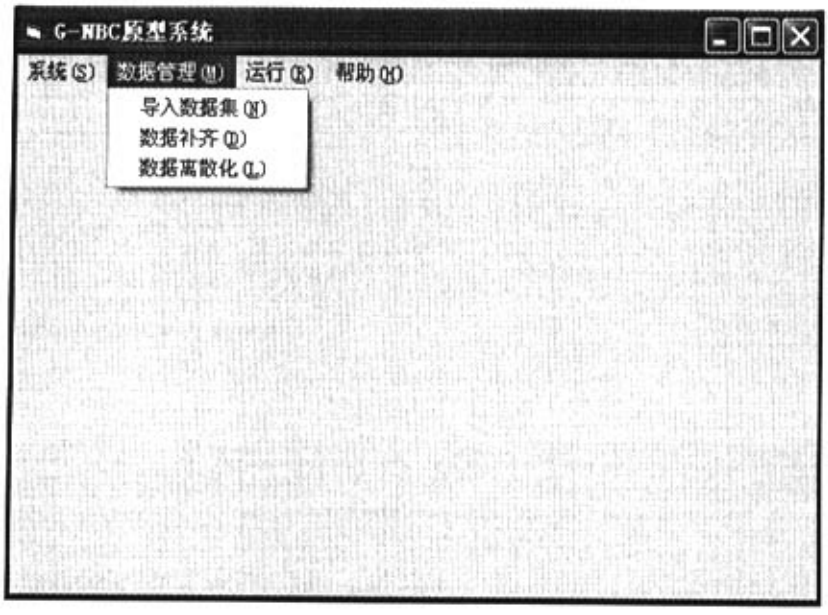

图 5.3 数据管理

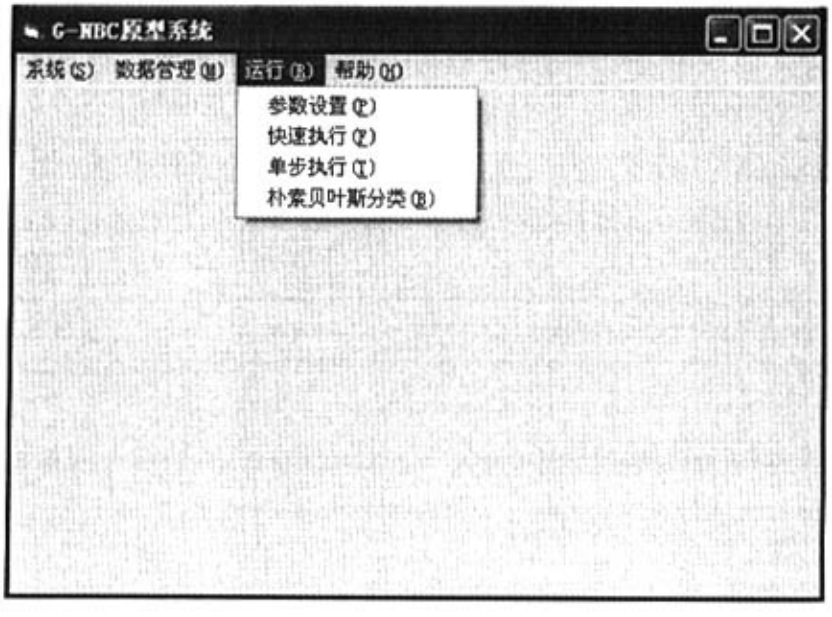

图 5.4 运行界面

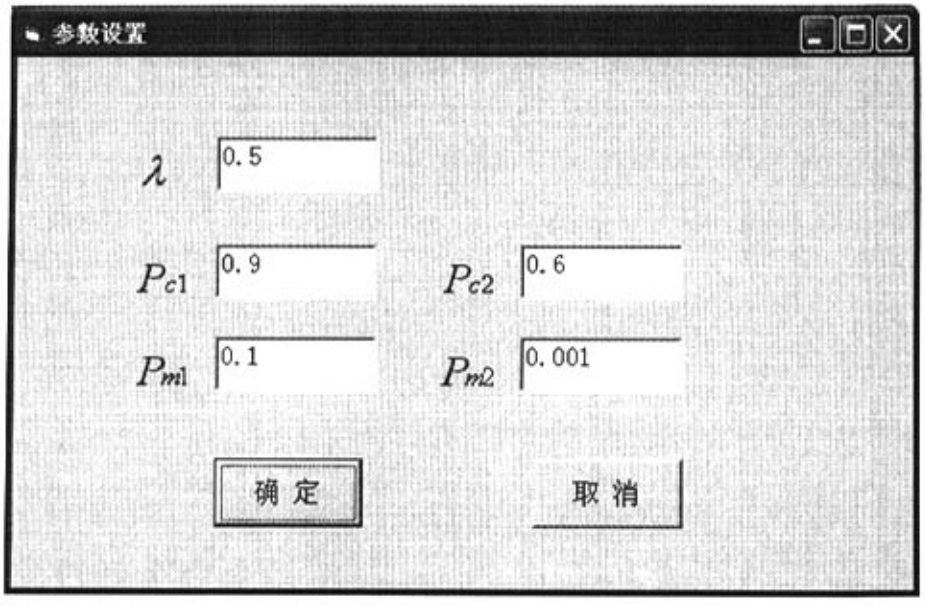

图 5.5 参数设置

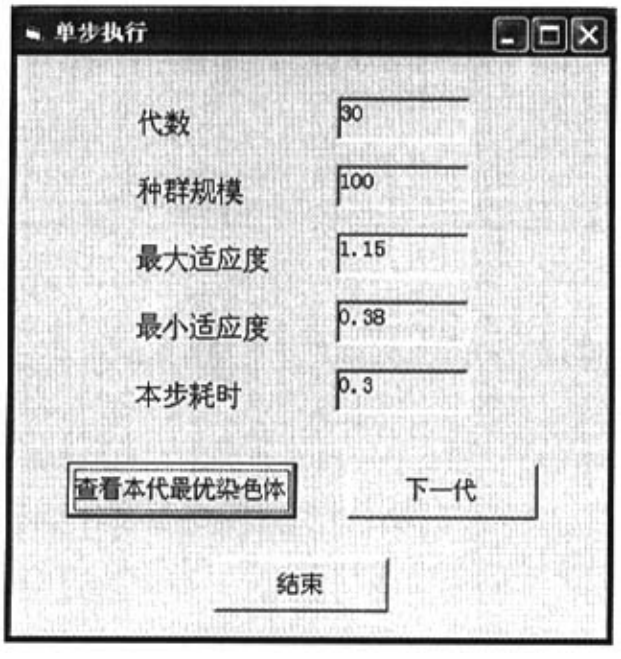

图 5.6 单步执行

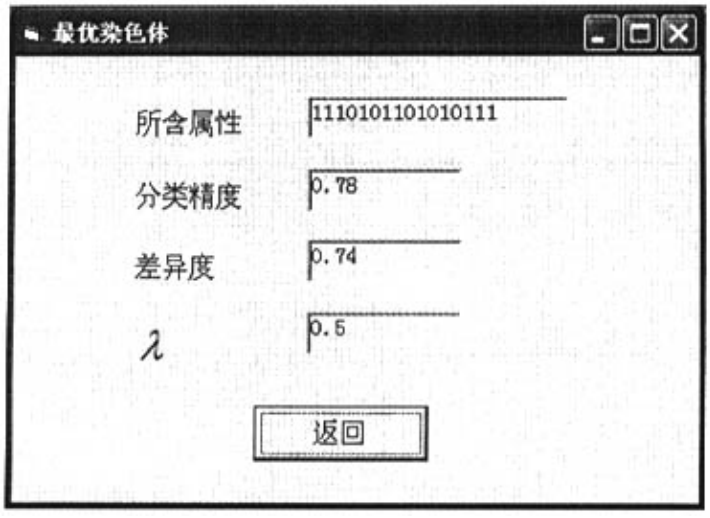

图 5.7 查看最优染色体

图 5.8 实验结果显示

5.4 本章小结

本章主要介绍了 G_NBC 原型系统的实验环境、功能结构图和主要运行界面。

第六章　　总结与展望

6.1 总结

本文所做的工作是对朴素贝叶斯分类算法的改进，提出了一种基于遗传算法和朴素贝叶斯的 G_NBC 算法。

贝叶斯分类器是建立在经典的贝叶斯概率理论的基础上的基于统计方法的分类模型。本文在简要介绍基本贝叶斯理论之后，在此基础之上重点介绍朴素贝叶斯分类器及其改进。

遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化概率搜索算法，来源于达尔文的自然进化理论和孟德尔的遗传变异理论，具有坚实的生物学基础；它所具有的简单、通用、稳健等特性使得它在困难、复杂实际问题中显示出巨大的优越性，而且能在概率意义下收敛到问题的全局最优解。本文正是将遗传算法引入到朴素贝叶斯分类中，进一步提高分类器的分类精度。

本文的创新性主要体现在两个方面：

1. 将自适应遗传算法引入到朴素贝叶斯分类中，提高朴素贝叶斯分类器的分类精度。

2. 在适应度函数中，引进了差异度的概念，有效地改进分类的差异度；而且在不同的数据集上，可进一步提高分类能力。

6.2 展望

下一步的研究方向应是：

1. 如何在一个给定领域里自动选取 λ 的优化值问题。

2. 先就对遗传算法进行改进，再和朴素贝叶斯相结合。改进方法主要包括：改进遗传算子；将遗传算法与其他优化算法相结合，构造混合遗传算法；使用并行遗传算法等方法。

3. 本文的算法是和传统的朴素贝叶斯分类算法作了比较，和其它一些分类技术相比，有没有优越性，性能如何，这也是下一步将要考虑的问题。

总之，数据挖掘是一个热门的研究课题，从数据中发现潜在的、有价值的信息和知识已经是时代大势所趋，但是同时它也面临着诸多的挑战。

参考文献

[1]Jiawei Han, and Micheline Kamber: Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[2]唐晓萍.数据挖掘与知识发现综述[[J].电脑开发与应用，2002,15(4):3 1 -32.

[3]Chen Metal.Data mining:An overview from a database perspective. IEEE Trans.onKnowledge and Data Engineering.1996,Vo1.8: 866883

[4]Agrawal Retal.Database mining:A performance perspective.IEEE Transactionsonknowledge and Data Engineering.1993,Vo1.5:914925.

[5]Fayyad Uetal.Knowledge discovery and data mining towards a uning framework.InKDD'96 Proc.2nd Int.Conf.on Knowledge Discovery&DataMining. AAAI Press,1996.

[6]刘刚. 数据挖掘技术与分类算法研究.解放军信息工程大学博士学位论文.2004

[7]戴稳胜等.数据挖掘中的关联规则[[J].统计研究，2002, 8:40-42.

[8]马国兵等.数据挖掘技术在运动目标轨迹预测中的应用[J].计算机工程与应用，2004, 40(11):210-211.

[9]刘红岩，陈剑，陈国青著.数据挖掘中的数据分类算法综述[J].清华大学学报(自然科学版),2002,42(6): 727-730.

[10]张瑜等.关于时序模式发现算法的研究[[J].河北科技大学学报，2004,25(2):27-29.

[11]MehtaM&AgrawalRetal. SLIQ:A fasts calable classifier for data mining. InProc.l996 Int.Conf.Extending DataBase Technology. Avignon, France. Mar.1996.

[12]罗海蛟等.数据挖掘中分类算法的研究及其应用[J].微机发展，2003, 13(z2):48-50.

[13]Quin lanJR. Bagging,boosting and C4.5. InProc. 13th Natl. Conf. Artificial Intelligence.Portland,OR.Aug.1996:725-730.

[14]ShaferJ&AgrawalRetal.SPRINT:A scalable parallel classifier for datamining.InProc.1996 Int.Conf VeryLarge Data Bases.Bombay, India. Sept. 1996:544555.

[15]MichalskiRSetal.Machine learning and datamining:Methods and applications.NewYork:Jone Wiley and Sons.1998.

[16]GehrkeJetal.Rainforest:A framework for fast decision treec on struction of large datasets. InProc.1998Int.Conf.Very Large Data Bases. NewYork, USA. Aug.1998:416427.

[17]GehrkeJetal. BOAT: Optimistic decision tree construction. In Proc. 1999 ACM-SIGMOD Int .Conf Management of Data.Philadelnhia.USA.June 1999:169--180.

[18]DomingosPetaLBeyond independence:Conditions for the optimality of the simple bayesian classifier.InProc.l3thInt.Conf.Machine Learning.1996:105-112.

[19]Russell Setal.Local learning in probabilistic networks with hidden ariables.InProc.l4th JointInt..Conf.Artifiical Intelligence. Montreal, Canada.Aug.1995:1146-1152.

[20]BuntineAetal.Opretation for learning with graphical models. Journal of Artificial Intelligence Research.1994,Vo1.2:159-225.

[21]Lauritzen S Letal.The EM algorithm for graphical association models with missing data.Computation Statisticsand Data Analysis. 1995,Vo1.19:191-201.

[22]LuW&SetionoRetal.Neuro rule:A connection is tapproach to data mining.InProc.1995 Int.Conf.VeryLarg eData Bases. Zurich, Switzerlanf. Sept. 1995: 478-489

[23]Widrow Betal.Neural networks:Applications in industry,business and science.Communications of ACM.1994,Vo1.37:93-105.

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

[25]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

[26]史忠植．知识发现[M]．清华大学出版社．2002.

[27]朱明．数据挖掘[M]．中国科技大学出版社．2002.

[28]Kononenko I.Semi- Naïve Bayesian Classifiers[A].In:Proceedings of Europe an Conference on Artificial Intelligence [C]. Porto ,Portugal:Springer-Verlag,1991.206-219

[29]Langley P，Sage S. Induction of Selective Bayesian Classifiers[A]. In : Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence[C].Seattle, WA: Morgan Kaufmann Publishers , 1994.339-406.

[30]Kohavi R. Scaling up the Accuracy of Naive — Bayes Classifiers: A Decision-Tree Hybird[A].In:Simoudis E,Han J W ,Fayyad U M.Proceedings

of the Second International Conference on Knowledge Discovery and Data Mining[C].Menlo Park, CA:AAAI Press,1996.202-207.

[31]Zheng Z ，Webb G I. Lazy Learning of Bayesian Rules[J] .Machine Learning,2000,4 1:53-84.

[32]Wang Z H ,Webb G I .A Heuristic Lazy Bayesian Rule Algorithm[A] .In:Simoff S J ,Wiliams G J ,Hegland M .Proceedings of Australian Data Mining Workshop[C].Sydney，Australia: Sydney University of Technology Press,2002.57-63

[33]郭亚光，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

[34]王小平，曹立明.遗传算法.西安交通大学出版社，2002

攻读硕士学位期间发表的论文

发表的论文：

1. 胡为成，胡学钢. 基于遗传算法的朴素贝叶斯分类，计算机技术与发展，2007.1，已录用.
2. 胡为成，王本年，程转流. 基于模拟退火算法的遗传程序设计方法，计算机工程与设计，2007.2 ，已录用.
3. 胡为成，王本年，程转流. 基于 Rup 思想和 B/S 模式的考试系统，计算机技术与发展，2006.3.