# 基于机器学习算法的分类知识发现及其在文本分析中的应用

祁瑞华 著

清华大学出版社

# 基于机器学习算法的分类知识发现及其在文本分析中的应用

祁瑞华 著

清华大学出版社
北京

## 内 容 简 介

随着数据获取技术的不断发展和电子商务的广泛应用，各种信息正以前所未有的速度日益积累，高效率地分析信息海洋中的大量数据已经成为商业领域、工程领域和科学领域的共同需要。文本挖掘是数据挖掘领域的一个分支，与数据挖掘假设数据源是结构化数据集相比，文本挖掘的对象是非结构化或是半结构化的文本集合，需要从以文件形式存储的文本中提取和分析特征。不完整数据处理是现实世界中分类知识挖掘必须认真考虑和对待的重要问题。本书探讨了不完整数据分类算法的改进及其在文体风格识别中的应用，并基于缺失补偿策略最大熵模型对文本分类算法改进进行了探索性的研究。

本书既可以作为数据挖掘或文本分析领域的研究人员及相关专业的研究生开展文本分析与处理研究的教科书，也可以作为政府相关部门产品研发人员的参考书。



**图书在版编目(CIP)数据**

基于机器学习算法的分类知识发现及其在文本分析中的应用/祁瑞华著. --北京：清华大学出版社，2015
ISBN 978-7-302-41576-3

Ⅰ. ①基… Ⅱ. ①祁… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2015)第 220443 号

**责任编辑**：付弘宇　柴文强
**封面设计**：何凤霞
**责任校对**：焦丽丽
**责任印制**：刘海龙

**出版发行**：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦 A 座　　邮　编：100084
社总机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课　件　下　载：http://www.tup.com.cn，010-62795954
**印 装 者**：三河市金元印装有限公司
**经　　销**：全国新华书店
**开　　本**：170mm×230mm　　**印　张**：11.75　　**字　数**：205 千字
**版　　次**：2015 年 12 月第 1 版　　**印　次**：2015 年 12 月第1 次印刷
**印　　数**：1～1000
**定　　价**：39.00 元

产品编号：063144-01

# 前　言

各行业数据集普遍存在数据不完整的情况。据统计，在机器学习和数据挖掘应用过程中，不完整数据的预处理花费大量的时间和精力。不完整数据处理是现实世界中分类知识挖掘必须认真考虑和对待的重要问题。有效地处理不完整数据有助于更加充分地利用已经搜集到的数据，从而提高机器学习和数据挖掘的效率，探索不完整数据的分类知识挖掘具有重要的现实意义。本书探讨了不完整数据分类算法的改进策略，第1章为不完整数据知识发现研究背景概述，第2章针对朴素信念分类算法忽略了属性变量的投票权重，提出了基于相关系数的加权保守推理规则；第3章针对目前半监督分类算法中未考虑缺失属性数据项隐含信息和算法复杂度高的情况，提出两阶段半监督加权朴素信念分类模型；第4章针对朴素信念分类算法明确分类样本比例低的情况，提出基于放松区间优势的不完整数据分类模型。并均在国际公开标准数据集上进行了对比实验，验证了提出模型在不完整数据上进行分类知识发现的性能。

典籍英译本的文体风格识别在对外作品推荐、匿名作者识别和促进中外文化交流方面具有重要的意义。本书第5章选择典籍英译作品作为研究对象，进行基于不完整数据分类算法的文体风格识别应用研究，进一步验证了本书提出的模型方法的有效性和性能。

本书第6章尝试在最大熵文本分类模型中使用高斯平滑进行特征补偿，并提出混合的特征选择方法对传统的特征选择方法进行改进。实验结果显示，基于特征缺失补偿最大熵模型的分类器的综合性能较好。

本书第7章基于微博客的网络舆情指标体系，分析了基于关键字的微博客舆情传播规律，进行了基于关键字的网络舆情个案研究；同时探讨网络文本的多语言特性，分析了网络文本情感分析粒度、基本问题、前沿问题和研

究框架。

本书可以作为数据挖掘或文本分析领域的研究人员及相关专业的研究生开展文本分析与处理研究的教科书，也可以作为政府相关部门产品研发人员的参考书。

本书能够尽快完成出版，首先要感谢我的同事霍跃红老师，本书的研究思想的起源来自与霍跃红老师的探讨合作，她无私提供了典籍英译文本语料；感谢刘彩虹老师、郭旭老师等，以及参与数据收集和整理的同学们，本书的若干专题研究都与他们有深入的讨论。还要感谢清华大学出版社的员工，是他们的鼓励和细致工作使得本书得以顺利出版。最后感谢在本书中所引用参考文献的作者们和公开语料库的开发者们，本书的写作从他们的研究成果中获取了很多营养，正是他们勤奋和分享的科研精神引领和启发我完成本书的写作。

本书研究获得大连外国语大学学术专著出版资助，2014 年大连外国语大学学科建设专项经费资助，特此表示感谢。

虽然我始终以认真严谨的态度对待本书的撰写工作，但很多研究尚属于探索阶段，书中难免有不足之处，恳请广大读者批评指正！

祁瑞华

2015 年 12 月

# 目　录


**第 1 章　概述** …… 1

1.1　分类知识发现 …… 1

1.1.1　知识发现的概念和过程 …… 1

1.1.2　数据挖掘中的知识表示模式 …… 4

1.1.3　分类知识发现主要算法 …… 7

1.1.4　不完整数据分类知识发现 …… 15

1.2　文本挖掘 …… 17

1.3　本书内容组织 …… 21

**第 2 章　不完整数据分类算法研究** …… 23

2.1　不完整数据分类知识发现 …… 24

2.1.1　不完整数据的类型 …… 24

2.1.2　不完整数据的处理 …… 25

2.1.3　不完整数据分类算法 …… 29

2.1.4　健壮贝叶斯分类 …… 30

2.1.5　朴素信念分类 …… 32

2.2　对现有方法的思考 …… 34

2.2.1　朴素信念分类算法的权重假设简单 …… 34

2.2.2　缺乏属性数据和类标记同时缺失情况下分类知识发现的研究 …… 35

2.2.3　半监督算法的效率问题 …… 35

2.3　不完整数据加权朴素信念分类算法 …… 36

2.3.1　相关分析及相关系数 …… 37

2.3.2 加权保守推理规则 …… 39
2.3.3 加权朴素信念算法分类过程 …… 41
2.4 标准数据集 UCI 上的对比实验 …… 44
2.4.1 实验数据集及实验设计 …… 44
2.4.2 实验结果分析 …… 45
2.5 本章小结 …… 48

**第 3 章 两阶段半监督加权朴素信念分类算法研究** …… 49
3.1 半监督分类知识发现研究现状 …… 49
3.2 问题分析 …… 52
3.2.1 未标记样本在分类学习中的作用 …… 52
3.2.2 现有半监督分类方法分析 …… 54
3.3 两阶段分类方法相关思路 …… 57
3.3.1 基于规则模型的两阶段分类 …… 58
3.3.2 两阶段半监督文本分类 …… 59
3.4 两阶段半监督加权朴素信念分类 …… 59
3.4.1 TSS-WNC 分类主要过程 …… 60
3.4.2 时间复杂度分析 …… 63
3.5 在标准数据集 UCI 上的实验 …… 64
3.5.1 分类对比实验 …… 64
3.5.2 实验结果及分析 …… 64
3.5 本章小结 …… 65

**第 4 章 放松区间优势的朴素信念分类算法研究** …… 66
4.1 问题分析 …… 66
4.2 区间优势比较 …… 69
4.3 基于放松区间优势推理规则的不完整数据分类 …… 73
4.3.1 放松的区间优势 …… 73

4.3.2　放松的区间优势推理规则 …… 74
4.3.3　基于放松区间优势推理规则的分类过程 …… 78
4.4　在标准数据集 UCI 上的实验 …… 78
4.4.1　RCIR-NCC 分类对比实验 …… 78
4.4.2　实验结果分析 …… 82
4.5　本章小结 …… 84

**第 5 章　典籍英译文体风格识别研究** …… 85
5.1　文体风格特征 …… 85
5.2　文体风格识别算法 …… 87
5.3　典籍英译文体风格向量空间模型 …… 89
5.3.1　典籍英译语料特点 …… 89
5.3.2　典籍英译多层面文体风格模型 …… 90
5.4　文体风格特征选择 …… 95
5.4.1　信息增益 …… 95
5.4.2　$\chi^2$ 统计量 …… 97
5.4.3　典籍英译文体风格识别特征选择 …… 97
5.5　特征数据项缺失文体识别实验 …… 99
5.5.1　加权朴素信念文体风格识别实验 …… 102
5.5.2　两阶段半监督文体风格识别实验 …… 106
5.5.3　放松区间优势朴素信念文体风格识别实验 …… 114
5.5.4　类别不平衡文体识别实验 …… 116
5.6　本章小结 …… 123

**第 6 章　基于特征缺失补偿最大熵模型的文本分类** …… 124
6.1　最大熵模型 …… 124
6.2　基于 Gaussian 先验平滑特征补偿的最大熵模型 …… 125

6.3 混合特征选择算法 …… 126
6.4 基于特征缺失补偿最大熵模型的文本分类 …… 127
6.5 本章小结 …… 130

**第 7 章 基于文本分析的网络舆情研究** …… 131
7.1 基于微博客的网络舆情指标体系 …… 131
7.1.1 网络舆情指标体系 …… 132
7.1.2 基于微博客的网络舆情指标体系 …… 136
7.1.3 微博客舆情预警对策 …… 140
7.2 基于关键字的微博客舆情传播规律 …… 141
7.2.1 网络舆情传播规律 …… 141
7.2.2 微博客网络舆情传播规律和对策 …… 143
7.3 基于关键字的网络舆情个案研究 …… 144
7.3.1 个案研究环境及实验数据 …… 144
7.3.2 大连地区抢盐潮个案分析 …… 145
7.4 微博客舆情的跨语言特征 …… 148
7.4.1 跨语言微博客特征表示 …… 150
7.4.2 跨语言微博客舆情预警研究框架 …… 153
7.5 网络文本情感倾向 …… 154
7.5.1 网络文本情感分析粒度 …… 154
7.5.2 网络文本情感分析基本问题 …… 158
7.5.3 网络文本情感分析前沿问题 …… 161
7.5.4 网络文本情感分析研究框架 …… 162
7.6 本章小结 …… 164

**参考文献** …… 165

# 第1章 概 述

本章介绍本书的研究背景和国内外研究现状，阐述本书的研究目的、方法及主要研究工作并说明本书的整体结构与组成。

## 1.1 分类知识发现

### 1.1.1 知识发现的概念和过程

随着数据获取技术的不断发展和电子商务的广泛应用，各种信息正以前所未有的速度日益积累，高效率地分析信息海洋中的大量数据已经成为商业领域、工程领域和科学领域的共同需要，从大规模乃至海量数据集中发现潜在规律并获取有用知识的智能信息处理技术——数据库知识发现(Knowledge Discovery in Databases，KDD)应运而生。KDD 概念的首次提出是在 1989 年于美国底特律市召开的第十一届国际人工智能会议举行的以 KDD 为主题的学术讨论会上。知识发现是以统计学、决策支持系统和数据库等多种学科为基础，综合机器学习、信息论和可视化等技术而形成的多领域交叉学科。KDD 不仅能够学习利用现有的知识，更重要的功能是发现新的对用户有价值的知识使其“显式”表达，并将这些知识应用在决策支持、查询响应、过程控制和信息管理等方面，便于人们理解和应用。因此，从 KDD 出现就一直受到国内外学者和相关机构的广泛重视，目前已经成为计算机科学、管理科学与工程等相关领域研究的重要课题之一，并被普遍认为是能够为商业领域带来丰厚回报的热点研究方向。

众多的学者根据自身对知识发现的理解和认识给出了 KDD 的定义，其

中比较全面和准确并具有共识性的是由 Fayyad 在 1996 年知识发现和数据挖掘国际学术会议论文中给出的定义[1]：

数据库知识发现是从数据集中识别出有效的、新颖的、潜在有用的并且能够最终被人们理解的模式的高级处理过程。

其中，数据集是指一个有关事实 $K$ 的集合(例如银行客户的基本情况和信用好坏的记录)，这个集合用来描述事物各方面的有用信息，是进一步发现知识的原始资料。有效性是指从数据集中发现的模式必须具有一定的可信度和正确性；新颖是指从数据中发现的知识必须是人们以前未知的、未注意到的或者没有期望得到的新的规则模式。模式的新颖与否通常通过两个途径衡量：一是将得到的数据与以前得到的数据或期望得到的数据相比较是否具有一定的新颖性；二是发现的新模式与以后模式是否有一定的关系。如果用函数 $N(E,K)$ 表示模式的新颖程度，其返回值可以是逻辑值或者是模式 $E$ 的新颖程度数值；潜在有用性是指发现的知识能够为人们的决策行为提供支持，是有意义的；可理解性是指从数据中提取的隐含知识应该用人们容易理解的形式表达出来，从而帮助人们更好地理解数据中包含的信息。KDD 不同于以往知识获取技术的一个特点是其发现的知识应该是人们可以理解的信息。模式是指可以用语言 $L$ 描述集合 $K$ 中数据的特性；高级过程是指 KDD 不仅是对数据进行简单运算或者查询，而是对数据进行更深层次处理，并在一定程度上智能化和自动化的过程。

从功能上看，KDD 技术能够帮助人们获得决策所需的知识。当用户不知道数据集中存在哪些有价值的信息和知识的情况下，KDD 技术可以搜索并允许用户参与和指导发现新的、多层次、有价值的模式及知识的挖掘过程。

随着对 KDD 的深入研究及其在众多领域的广泛应用，已经开发出大量具有商业价值的知识发现产品投入市场，并在客户关系管理、零售业和金融等环节成功应用，例如：Rough Enough、Rose、KDD-R、LERS、KEFIR、Rosetta、Recon 等。

知识发现过程是一个多步骤相互连接，不断反馈的人机交互过程，

Fayyad 给出的 KDD 知识发现处理过程[2]是公认的通用知识发现过程定义，如图 1.1 所示。

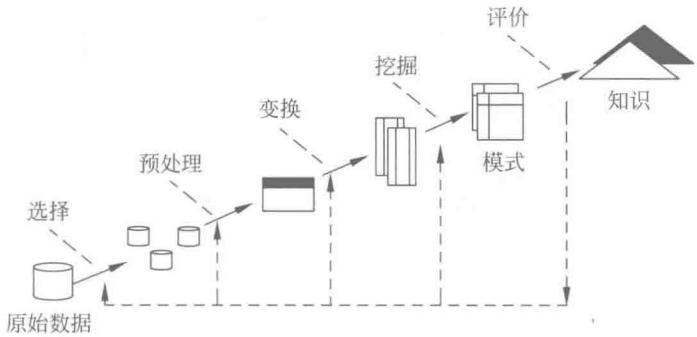


图 1.1　Fayyad 定义的知识发现过程模型[2]

知识发现的过程是一个反复迭代的过程，其基本步骤包括数据选取、数据预处理、数据变换、数据挖掘和模式解释与评估。其中，数据挖掘是知识发现过程的核心步骤，这一过程的任务是运用各种算法从数据中识别、发现和抽取潜在的模式、规则或最终可理解的知识的过程。由于很多数据挖掘算法的功能已经超越了数据挖掘原本的范围，目前数据挖掘往往被当作知识发现的同义词，但严格地说数据挖掘是知识发现的一个核心步骤。

知识发现过程各个基本阶段的任务如下所述：

(1) 数据选取、预处理和数据变换

高质量的数据集是获得准确数据挖掘的前提，这部分过程的目的是通过保证数据集与挖掘任务的相关性来提高数据挖掘的效率和准确性。数据选取是指准确界定选取的数据源和抽取原则，根据用户的需要从原始数据库中抽取数据。数据预处理是指去除或推导计算数据源中不完整和不一致的数据、消除噪声、删除重复记录、转换数据类型转换的一系列过程。数据变换的主要任务是消减数据集的初始特征维数，去除冗余的、可以忽略的特征，从中选取有用的特征以提高数据挖掘的效率，如果需要则对数据结构进行变换，并将数据集整理为能够被数据挖掘算法直接使用的存储形式。

(2) 数据挖掘

数据挖掘是知识发现的重要步骤，在这个阶段将建立挖掘模型并利用

相应的数据挖掘算法对指定的数据集进行知识提取。相关的模型和算法包括分类[3][4][5][6]、聚类[7][8][9][10][11]、关联规则发现[12][13][14]、序列模式发现[15][16][17]和决策树[18][19]等。算法的选择依据主要是考虑数据集自身的特点和用户实际应用的需求，例如：有的用户希望获取预测准确率尽可能高的预测性知识，有的用户希望获取的知识是描述型易于理解的，有的用户希望快速地建立模型而只要求达到一般的预测准确率。数据挖掘质量主要取决于两个主要因素：数据挖掘技术的有效性、用于数据挖掘的数据集的质量和样本数量。

(3) 模式评估

任何模型都无法适应所有的数据集，因此模型的验证和评估显得尤为重要。在模式评估阶段将对上一阶段的挖掘结果与预期目标进行对比，进行一致性和合理性的检查。在验证过程中可以采用样本学习的方式，将数据集用交叉验证或者按比例的方式分为两部分：训练集和测试集。使用训练集建立模型，用测试集验证模型的准确性。经过验证和评估，可以删除冗余和无关的模式。如果结果与预期目标有较大偏差，则将知识发现过程回退到前面的某个阶段。因此知识发现过程的整个发现过程是一个反复迭代的过程。

在实际应用中，数据挖掘的处理对象是大量数据，数据挖掘阶段还有可能在知识发现的迭代过程中多次反复，因此时间复杂度是验证阶段需要考虑的重要因素。

### 1.1.2　数据挖掘中的知识表示模式

KDD是涉及不同领域的新兴的交叉性学科，自诞生以来出现了很多相关的术语和名称。例如："知识抽取"(Information Extraction)、"信息收获"(Information Discovery)、"信息发现"(Information Discovery)等。从名称来看，KDD更强调与数据库的联系。原则上KDD的研究对象可以是以各种存储方式的信息，本文所述知识发现的研究对象主要是指数据库、数据仓库中的知识发现。

数据挖掘是KDD过程中特定的模式抽取阶段，是知识发现最重要的步骤。这一阶段主要采用的知识表示模式和方法主要包括[20]：

(1) 广义知识

广义知识是指描述类别特征的概括性知识，广义数据挖掘是指反映同类事物共同性质，对含有大量数据的数据集合进行数据概括、精炼和抽象的过程。源数据中体现的通常是细节性数据，通过对这些数据进行不同层次上的泛化可以找出其中隐含的概念或逻辑，提供从较高层次上观察和处理数据的途径。

KDD能够根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识，这一过程所发现的知识类型称为概念描述。获取概念描述的技术主要有：

① 联机分析处理(OLAP)

OLAP的概念最早是由关系数据库之父E. F. Codd于1993年提出的，他同时提出了关于OLAP的12条准则。OLAP通过计数、求和、平均、最大值、钻取等操作，实现多维数据库中的数据分析，能够快速、灵活地提供不同角度和不同抽象层次上的数据视图，可以应用于决策支持、知识发现或其他领域。

② 面向属性的归约方法

面向属性的归约方法以类语言表示数据挖掘查询，收集数据库中的相关数据集，然后在相关数据集上应用一系列数据推广技术(包括属性删除、概念树提升、属性闭值控制、计数及聚集函数传播等)，进行数据推广。

(2) 关联分析

关联分析是从大量数据中挖掘出频繁出现的、有价值的描述数据项之间相互联系的知识的过程。数据库中数据之间的相互联系是现实世界中事物联系的表现。如果数据的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，从而有助于进行有关的商业决策。关联规则的发现可分为两步：第一步是迭代识别所有的频繁项目集，要

求频繁项目集的支持率不低于用户设定的最低值；第二步是从频繁项目集中构造可信度不低于用户设定的最低值的规则。其中识别或发现所有频繁项目集是关联规则发现算法的核心，也是计算量最大的部分。

(3) 分类知识

分类模型是数据挖掘的基本模型之一，也是知识发现中最重要的目标和任务之一。在分类模型中数据集的属性通常分为预测属性和目标属性，也称为条件属性和决策属性，待分类的数据分为训练集和测试集。分类知识挖掘的任务是对训练集进行学习，抽取其蕴涵的一般性分类知识预测规则，并用这些规则对待分类数据进行分类预测，即根据条件属性值来预测决策属性值，其性能可以通过测试集来进行测试。

构造分类器需要有一个训练集作为输入，训练集中的每个样本描述为由若干个属性变量和一个类变量组成的向量。分类一般包括两步：首先是学习阶段，在此阶段用已知的训练集构建分类器，训练集中的每个样本称作训练样本，通常训练样本的类标记是已知的。第二步是测试阶段，在此阶段使用构建好的分类器预测测试集中待分类样本的类别。

分类知识是通过对源数据中类知识的过滤、抽取、压缩和概念提取来获取的，源数据集中每个训练样本已经有类标记，因此从机器学习的角度，分类模型的学习是一种有监督的学习(Supervised Learning)。

具有代表性的分类模型构造方法主要包括：贝叶斯方法、支持向量机、决策树模型、近邻法、人工神经网络、基于实例的学习方法、多元判别分析模型、基于粗糙集的方法和基于模糊集的方法等等。

(4) 聚类分析

聚类分析是将数据集中没有类别标记的数据按照相似性归成若干类别，标注类标记的过程。聚类学习的依据是使属于同一类别样本差别尽可能的小，不同类别样本间的差别尽可能的大。聚类是一种无监督学习，在统计方法、机器学习和神经网络等方法的基础上，通过对数据的分析比较生成新的类标记，展示数据中蕴含的类知识。主要的聚类分析方法有基于划分的方法、层次聚类法、密度方法、网格方法和模型方法[21]。

(5) 特异型知识

特异型知识是对差异和极端特例的描述，目的是揭示了事物偏离常规的异常规律。通过从数据集中检测特异型知识能够发现标准类别之外的特例或聚类之外的离群值，可以应用于如商业欺诈行为、网络非法入侵行为自动检测等领域。

### 1.1.3　分类知识发现主要算法

分类知识发现的目的是构造一个能够描述数据概念集的模型，然后运用这个模型对新的数据进行分类，是机器学习、模式识别、统计学等研究领域的基础问题。目前已经开发出很多分类知识发现算法并广泛应用于实践，主要有：基于统计学的贝叶斯方法、最近邻法、基于范例的推理方法以及神经网络、决策树、粗糙集方法等。

(1) 贝叶斯学习[22]

贝叶斯学习方法基于概率统计理论，具有稳固的数学基础，一直是分类知识发现研究的重要方向。在构造分类模型时，贝叶斯方法以概率表示各种形式的不确定性，基于贝叶斯定理预测待分类样本属于某个特定类别的可能性，能够很好地处理不确定性问题。

贝叶斯定理是由英国学者贝叶斯在 18 世纪提出的关于随机事件的先验概率和后验概率关系的一则定理，令 $P(A)$为事件 $A$ 的先验概率，$P(B)$为 $B$ 的先验概率，$P(A|B)$为已知 $B$ 发生条件下 $A$ 发生的后验概率，$P(B|A)$为已知 $A$ 发生条件下 $B$ 发生的后验概率，贝叶斯定理可以表示为：

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \tag{1.1}$$

如果影响事件 $A$ 的因素包括 $B_1, B_2, \cdots, B_n$，贝叶斯定理可以表示为：

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)} \tag{1.2}$$

基于贝叶斯的分类算法的基本原理就是通过特定实例已知的先验概率，利用贝叶斯定理计算此实例属于各个类别的后验概率，其中具有最大后

验概率的类别就是贝叶斯分类算法预测的分类结果。研究较多的贝叶斯分类算法有普通贝叶斯网络分类器(General Bayesian Network Classifier, GAN)[23]、朴素贝叶斯分类器(Naïve Bayes Classifier, NBC)[24]、树扩展朴素贝叶斯(Tree-Augmented Naïve Bayes, TAN)[22]、贝叶斯网络扩展朴素贝叶斯(Bayesian network Augmented Naïve Bayes, BAN)等。

普通贝叶斯网络分类器是将直观的知识表示形式与概率理论有机结合的基于概率推理的有向无环网络,其中以节点表示随机变量,以节点之间的有向边表示节点之间的关系,以条件概率表示节点之间关系的强度。贝叶斯网络不仅能处理定量信息而且可以处理定性信息,可以处理含有数值型和非数值型信息的数据集,可以直接处理多类别分类问题,还能够处理层次型结构或层次型类别分类问题,并且具有对不精确数值的鲁棒性。

其局限主要在于,贝叶斯网络的建立需要领域专家的参与,建立起网络的节点关系后,通过对贝叶斯网络的训练来获得节点关系的概率参数。贝叶斯网的推理已经被证明是 NP 难题,对于规模较大的贝叶斯网来说,即使是近似推理也是一项非常费时的工作。一般的贝叶斯模型受节点数量和节点间关系复杂性的限制,难以对含有大量特征项的高维数据集进行分类。

朴素贝叶斯分类器假定在给定类标记时,数据集的各个属性变量之间相互条件独立。对于属性为 $a_1, a_2, \cdots, a_n$ 的数据集,如果将类变量表示为 $c \in C$,独立性假设可以表示如下:

$$P(a_1, a_2, \cdots, a_n \mid c) = \prod_{i=1}^{n} P(a_i \mid c) \tag{1.3}$$

朴素贝叶斯分类器的分类公式表示如下:

$$c(x) = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(a_i \mid c) \tag{1.4}$$

研究表明,如果朴素贝叶斯分类器的属性独立性假设没有改变真实情况下类别后验概率值的排列顺序,朴素贝叶斯分类器是高效率和高准确率的[25]。反之,其分类准确率会明显降低。

为避免朴素贝叶斯独立性假设的负面影响,围绕改进朴素贝叶斯分类器开展了一系列研究,具有代表性的是 Friedman 提出的树扩展朴素贝叶斯(Tree-Augmented Naïve Bayes, TAN)[22]和贝叶斯网络扩展朴素贝叶斯。