资助 教育部新世纪优秀人才支持计划（NCET-12-0953）
武汉市青年科技晨光计划（2015070404010202）
国家青年科学基金项目（61203287）

贝叶斯网络分类器：算法与应用

BEIYESI WANGLUO FENLEIQI: SUANFA YU YINGYONG

蒋良孝 李超群 编著

中国地质大学出版社
ZHONGGUO DIZHI DAXUE CHUBANSHE

资助 教育部新世纪优秀人才支持计划(NCET-12-0953)
武汉市青年科技晨光计划(2015070404010202)
国家青年科学基金项目(61203287)

贝叶斯网络分类器:算法与应用

BEIYESI WANGLUO FENLEIQI:SUANFA YU YINGYONG

蒋良孝　李超群　编著

中国地质大学出版社
ZHONGGUO DIZHI DAXUE CHUBANSHE

图书在版编目(CIP)数据

贝叶斯网络分类器:算法与应用/蒋良孝,李超群编著. —武汉:中国地质大学出版社,2015.12

ISBN 978-7-5625-3780-9

Ⅰ.①贝…
Ⅱ.①蒋… ②李…
Ⅲ.①贝叶斯理论-应用-数据采集
Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 307024 号

贝叶斯网络分类器:算法与应用 蒋良孝 李超群 **编著**

责任编辑:张 琰 党梅梅 组稿编辑:张 琰 责任校对:周 旭

出版发行:中国地质大学出版社(武汉市洪山区鲁磨路 388 号) 邮政编码:430074

电 话:(027)67883511 传 真:67883580 E-mail:cbb@cug.edu.cn

经 销:全国新华书店 http://www.cugp.cug.edu.cn

开本:787 毫米×1092 毫米 1/16 字数:190 千字 印张:7.5

版次:2015 年 12 月第 1 版 印次:2015 年 12 月第 1 次印刷

印刷:武汉市籍缘印刷厂 印数:1—500 册

ISBN 978-7-5625-3780-9 定价:35.00 元

前　言

分类问题是数据挖掘与机器学习领域研究的重点问题，而贝叶斯网络分类算法作为一种简单有效的概率学习算法，近年来已得到广泛的关注和应用。国内外很多学者在贝叶斯网络分类算法上进行了大量理论和应用研究，但是，目前在国内还没有一本系统介绍贝叶斯网络分类算法方面的中文书籍。这对该类算法在国内的进一步研究有一定的阻碍，同时，也不便于该类算法被更多没有贝叶斯网络研究背景的计算机及相关专业工程人员使用。

为此，本书在综述了基本贝叶斯网络分类算法和当前贝叶斯网络分类算法研究进展的基础上，对作者近十年来在贝叶斯网络分类算法方面的理论与应用研究成果进行了系统的介绍，并对该类算法今后可能的研究方向进行了简述。本书不仅能为贝叶斯网络分类算法研究人员提供参考，同时也能为计算机及相关专业工程应用人员进行分类算法的选择提供方便。

本书共分为三个部分：第一部分（第1—7章）介绍了贝叶斯网络分类器的理论学习算法；第二部分（第8章）介绍了贝叶斯网络分类器在文本分类中的应用研究成果；第三部分（第9章）介绍了贝叶斯网络分类器在距离度量中的应用研究成果。三个部分的写作思路都是在给出相关基础知识和基本模型之后，从削弱属性条件独立假设的五个不同方向详细介绍了基本模型的改进思路和算法。目前，本书涉及的核心研究成果大部分已公开发表在中国计算机学会推荐的国际重要期刊和会议论文集上，书中相关内容更详细的介绍和实验结果可参阅编著者近年来发表的学术论文。

本书第一部分（第1—7章）主要由蒋良孝教授负责编写，第二部分（第8章）和第三部分（第9章）主要由李超群副教授负责编写。最后由蒋良孝教授和李超群副教授共同完成对本书的统稿和校读。在本书的编写过程中，得到了三位担任蒋良孝教授助研工作的硕士研究生的大力支持和帮助：邱晨同学负责了第3章和第7章部分内容的文字录入与整理工作；孔刚刚同学负责了第4章、第5章以及第6章部分内容的文字录入与整理工作；张伦干同学负责了第8章部分内容的文字录入与整理工作。在此，对三位同学的大力支持和帮助表示衷心的感谢。

本书的第一作者蒋良孝教授得到了教育部新世纪优秀人才支持计划和武汉市青年科技晨光计划的资助；本书的第二作者李超群副教授得到了国家自然科学基金青年科学基金项目和中国地质大学（武汉）摇篮计划的资助。本书编写过程中还得到了中国地质大学（武汉）计算机学院各位老师的关心和支持，在此一并表示感谢！

由于编著者水平有限，不当之处敬请批评指正。

蒋良孝　李超群

2015年10月

目 录

第1章 引 言

1.1 本书选题的背景和意义

分类是数据挖掘与机器学习中一项非常重要的任务，在现实生活中有着广泛的应用。比如文本分类、岩爆预测、欺诈检测，等等。构造分类器的方法很多，常见的有贝叶斯网络、决策树、基于实例的学习、人工神经网络、支持向量机、演化算法、粗糙集和模糊集，等等。其中，贝叶斯网络以其独特的不确定性知识表达形式、丰富的概率表达能力、综合先验知识的增量学习特性等成为众多方法中最为流行的方法之一。

鉴于学习最优的贝叶斯分类器如同学习贝叶斯网络一样是一个 NP - hard 问题(Chickering,1994)，从而学习朴素贝叶斯分类器得到了广大学者的重视。然而，朴素贝叶斯分类器基于一个简单而不现实的假设：在给定类标记时属性值之间相互条件独立。这在一定程度上影响了朴素贝叶斯分类器的分类性能。为此，学者们提出了许多改进朴素贝叶斯分类器的方法，概括起来主要分为五类：①结构扩展，这一类方法用有向边来表达属性之间的依赖关系；②属性选择，这一类方法从原始的属性空间中搜索出一个最佳的属性归约子集；③属性加权，这一类方法给每个属性赋予不同的权值；④局部学习，又称为实例选择，这一类方法利用局部学习原理在测试实例的邻域构建朴素贝叶斯分类器；⑤实例加权，这一类方法给每个训练实例赋予不同的权值。

本书在简单介绍贝叶斯网络理论的基础上，以朴素贝叶斯网络分类器的研究为出发点，以改进朴素贝叶斯网络分类器的五类方法为主线，对现有的贝叶斯网络分类器学习算法进行了较为全面的概括与综述。在此基础之上，作者利用增加隐含父亲结点、演化属性选择、动态邻域、判别学习、随机组合学习等新方法，设计了一系列新颖的贝叶斯网络分类器学习算法，并探讨了这些新算法在文本分类和距离度量两个实际问题中的应用。在研究过程中通过实验的方法证明了各算法的有效性。目前，相关研究成果大部分已公开发表在中国计算机学会推荐的国际重要期刊和会议论文集上。

据作者所知，在国内，还未曾发现任何教材和专著专门介绍各种贝叶斯网络分类器学习与应用算法。关于贝叶斯网络分类器方面的内容大都以章节的形式零散地出现在数据挖掘与知识发现、机器学习、模式识别和人工智能等教材中。在国外，有两本关于贝叶斯网络方面的专著：*Probabilistic reasoning in intelligent systems - networks of plausible inference*(Pearl,1988)和 *Bayesian Methods*(Leonard 和 Hsu,2001)，但它们都侧重于介绍基于贝叶斯网络的概率推理。另外，由于这两本专著都是英文书籍，限制了其在国内的有效普及。综上所述，本书为目前国内第一本用中文全面介绍贝叶斯网络分类器的书籍。

本书主要内容均为编著者近十年对贝叶斯网络分类器研究中取得的核心成果的详细描

述,其中包括改进算法的动机、改进算法的设计与分析、改进算法的应用以及今后可能的研究方向,等等。这些对贝叶斯网络分类器的研究人员具有很好的借鉴作用。同时,书中关于算法的优缺点分析部分对应用贝叶斯网络分类器解决分类问题给出了一些指导性的建议,便于计算机及相关专业工程应用人员使用。

1.2 如何阅读本书

本书主要介绍了编著者十余年来对贝叶斯网络分类器学习算法及其在文本分类和距离度量中的应用的核心研究成果。编著者尽量做到各章的写作与读者的阅读顺序无关,然而各章的相互依赖性不可能完全避免。建议读者在阅读本书的具体内容之前,快速浏览整本书的目录结构,以帮助读者对本书研究内容和写作思路有整体把握。通过对目录结构的阅读,读者不难发现,整本书分为三个部分:第一部分(1～7 章)主要讲解了贝叶斯网络分类器的理论学习算法,第二、三部分(8、9 章)分别介绍了贝叶斯网络分类器在文本分类和距离度量中的实际应用价值。三个部分的写作思路都是在给出相关基础知识和基本模型之后,从五个不同的方向详细介绍基本模型的改进。因此,建议读者先阅读前面相关基础知识和基本模型的介绍部分,然后可以以任意顺序独立阅读感兴趣的章节。下面简要介绍各章的内容。

第 2 章介绍了贝叶斯网络分类器学习相关的理论基础知识,包括分类的定义、贝叶斯网络的定义、贝叶斯公式、极大后验假设以及贝叶斯网络分类器和朴素贝叶斯的定义。

第 3 章介绍了朴素贝叶斯分类器的结构扩展方法,包括结构扩展学习的算法框架、研究综述以及编著者的核心研究成果。

第 4 章介绍了朴素贝叶斯分类器的属性选择方法,包括属性选择学习的算法框架、研究综述以及编著者的核心研究成果。

第 5 章介绍了朴素贝叶斯分类器的属性加权方法,包括属性加权学习的算法框架、研究综述以及编著者的核心研究成果。

第 6 章介绍了朴素贝叶斯分类器的局部学习方法,包括局部学习的算法框架、研究综述以及编著者的核心研究成果。

第 7 章介绍了朴素贝叶斯分类器的实例加权方法,包括实例加权学习的算法框架、研究综述以及编著者的核心研究成果。

第 8 章介绍了贝叶斯网络分类器在文本分类中的应用,包括文本分类简介、朴素贝叶斯文本分类器简介以及编著者在五个方向上取得的核心研究成果。

第 9 章介绍了贝叶斯网络分类器在距离度量中的应用,包括距离度量简介、距离度量与贝叶斯网络分类器、基本的值差度量以及编著者在五个方向上取得的核心研究成果。

第 2 章　贝叶斯网络分类器基础知识

2.1　分类的定义

分类是数据挖掘与机器学习中一项非常重要的任务，在现实生活中有着广泛的应用。例如，根据某天的天气状况判断该天是否适合打网球，根据病人各项检查指标判断其是否患有某种疾病，根据电子邮件的标题和内容判断其是否为垃圾邮件，根据动物的各项特征判断动物的类别，等等。

分类是一个两步走的过程。第一步，用类标记已知的实例集构建分类器。这一步一般发生在训练阶段（或叫学习阶段）。用来构建分类器的已知实例集称作训练实例集，训练实例集中的每一个实例称作训练实例。由于训练实例的类标记是已知的，所以分类器的构建过程是有导师的学习过程。相比较而言，在无导师的学习过程中，训练实例的类标记是未知的，有时候甚至连要学习的类别数也可能是未知的，比如聚类。

第二步，使用构建好的分类器分类类标记未知的实例。这一步一般发生在测试阶段（或叫工作阶段）。需要分类的未知实例称作测试实例。一般在分类器被用来预测之前，需要对它的分类精度进行评估。只有分类准确率达到要求的分类器才可以用来对测试实例进行分类。

评估分类器分类精度的方法很多，主要有：交叉法（Cross - Validation）、分裂法（Percentage Split）、补充集法（Supplied Test Set）、留一法（Leave One Out）和回代法（Using Training Set）。其中，交叉法和分裂法最为常用；留一法可以看成是交叉法的一种特例；回代法一般不推荐使用，因为它评估过的分类器倾向于过度拟和训练实例集，而导致评估的分类精度过于乐观；补充集法只在已知条件中有两个数据集的时候才使用。

2.2　贝叶斯网络的定义

贝叶斯网络是结合概率论和图论的概率模型。作为概率模型，一个基本的假设就是兴趣的数量取决于概率分布，同时最优决策可以通过对它们的概率与观测数据推理得到。贝叶斯网络不仅有严密的概率基础还有直观上很有吸引力的界面，所以在数据挖掘与机器学习的算法设计与分析中起着越来越重要的作用。近年来，贝叶斯网络被广泛应用于故障诊断、医疗专家系统和软件调试等应用领域。

一个贝叶斯网络由一个结构模型和一组条件概率组成。其中结构模型是一个有向无环图，图中的结点代表随机变量，有向边代表变量间的信息或者因果依赖关系。这种依赖关系用

网络中每个结点在给定其父亲结点前提下的条件概率来量化。下面是贝叶斯网络的正式定义(Zhang,2002)。

定义 1：一个贝叶斯网络是由一个有向无环图 $G=<N,E>$ 和一组概率分布 P 组成，其中 $N=<A_1,A_2,\cdots,A_n>$ 是结点的集合，E 是边的集合，P 是每一个结点 A_i 的局部条件分布的集合。A_i 的局部条件分布用 $P(A_i|pa_i)$ 表示，其中 pa_i 表示 A_i 的父亲。

随机变量分两种类型：名词(或离散)变量和数字(或连续)变量。名词(或离散)变量取值于一个有限的集合，而数字(或连续)变量取值于一组连续的实数。因此，贝叶斯网络也相应地分为两类：离散贝叶斯网络和连续贝叶斯网络。本书仅仅讨论离散贝叶斯网络。图 2-1 展示了一个离散贝叶斯网络的例子。

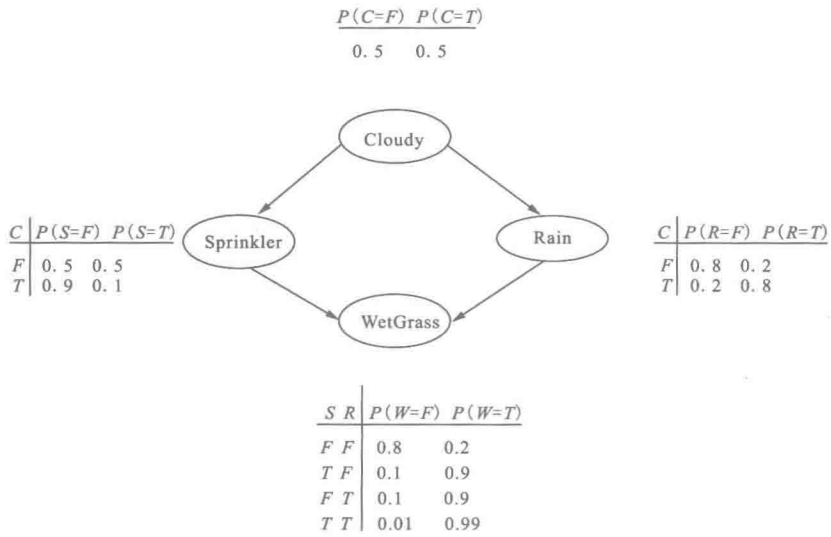

图 2-1 一个离散贝叶斯网络的例子

从概率论的角度来看，一个贝叶斯网络表示一组随机变量的联合分布。根据链式法则，这个联合分布可以用方程(2.1)来表示。

$$P(A_1,A_2,\cdots,A_n)=\prod_{i=1}^{n}P(A_i|A_{i+1},\cdots,A_n) \tag{2.1}$$

式中，$A_1,A_2,\cdots,A_n$ 是随机变量。

通常直接计算一个联合分布是非常困难的，因为随机变量的组合数目是呈指数增长的。然而，一个贝叶斯网络提供了一组关于随机变量的条件独立假设，以至于一个联合分布可以用因式分解的方法来简洁表示，即一个联合分布可以被分解成每一个变量在给定其父亲前提下的局部条件分布。贝叶斯网络结构表示的独立关系可以由马尔科夫条件给定。下面是马尔科夫条件的定义。

定义 2：贝叶斯网络中的任何结点在给定其父亲结点的前提下条件独立于它的非儿子结点。

马尔科夫条件允许随机变量 $A_1,A_2,\cdots,A_n$ 的联合概率分布被因式分解成如方程(2.2)所示的乘积。

$$P(A_1,A_2,\cdots,A_n)=\prod_{i=1}^{n}P(A_i\mid pa_i) \tag{2.2}$$

式中，pa_i 是 A_i 所有父亲的集合。

对于图 2-1 给出的例子，根据概率的链式法则，所有结点的联合分布可以用方程(2.3)来表示(分别用单词的首字母来表示每个变量)。

$$P(C,S,R,W)=P(C)\times P(S|C)\times P(R|C,S)\times P(W|C,S,R) \tag{2.3}$$

但通过应用马尔科夫条件，方程(2.3)可以被改写成方程(2.4)。

$$P(C,S,R,W)=P(C)\times P(S|C)\times P(R|C)\times P(W|S,R) \tag{2.4}$$

假如每个结点的父亲结点数量是有限的，那么所需参数数量随网络大小呈线性增长，但联合分布本身却呈指数增长。

根据马尔科夫条件，贝叶斯网络中的一个结点只会被它的马尔科夫毯中的结点所影响。下面是马尔科夫毯的定义。

定义 3：设 A 是贝叶斯网络 G 中的一个结点，那么 A 的马尔科夫毯是由 A 的父亲结点、A 的儿子结点以及 A 的儿子结点的其他父亲结点组成的集合，标记为 $MB(A)$。比如，如图 2-2 所示的例子，A_5 的马尔科夫毯是 $\{A_2,A_3,A_4,A_7\}$。

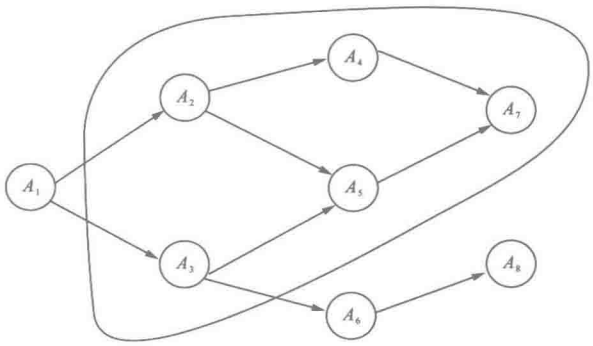

图 2-2　一个马尔科夫毯的例子

简而言之，贝叶斯网络提供了一种利用随机变量间条件独立性来简化联合分布的方法，这种方法使得直接计算和操作概率成为可能实用。通过这种方式，在给定任何其他子集作为证据的前提下，贝叶斯网络支持变量的任何子集的概率计算，这就叫贝叶斯推理。

贝叶斯推理的目标就是，在给定其他观察变量取值的前提下，导出任何目标变量的取值。在推理过程中，贝叶斯规则起着至关重要的作用。下面一节将详细介绍贝叶斯规则。

2.3　贝叶斯规则

在数据挖掘与机器学习中，通常我们感兴趣的是：给定训练实例集 D，假设空间 H 中最有可能的假设是什么。贝叶斯规则就提供了一种直接计算这种最有可能假设的方法。更准确地讲，贝叶斯规则提供了一种计算假设概率的方法，它基于假设的先验概率、给定假设下观察到

的不同数据的概率以及观察数据本身的先验概率。贝叶斯规则的定义如下：

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)} \tag{2.5}$$

式中，$P(h)$代表还没有观察到训练实例集之前假设 h 拥有的初始概率，即 h 的先验概率，它反映了所拥有的关于 h 是一正确假设的机会的背景知识，在没有这一先验知识的情况下，可以简单地将每一个候选假设赋予相同的先验概率；$P(D)$代表将要观察到的训练实例集 D 的先验概率，即在没有确定某一假设成立时 D 的概率；$P(D|h)$代表假设 h 成立的情形下观察到训练实例集 D 的条件概率；$P(h|D)$代表给定训练实例集 D 时 h 成立的条件概率，即 h 的后验概率，它反映了在看到训练实例集 D 后 h 成立的置信度。

可见，贝叶斯规则提供了用 $P(h)$、$P(D)$和 $P(D|h)$计算 $P(h|D)$的方法。另外，$P(h|D)$随着 $P(h)$和 $P(D|h)$的增加而增加，随 $P(D)$的增加而减少，因为如果 D 独立于 h 被观察到的可能性越大，那么 D 对 h 的支持度就越小。

2.4 极大后验假设

在大多数情况下，学习器需要发现给定实例集 D 时可能性最大的假设 $h\in H$（H 为候选假设的集合）。这种具有最大可能性的假设被称为极大后验（Maximum A Posteriori，简记为MAP）假设 h_{MAP}。即：

$$h_{MAP}=\underset{h\in H}{\operatorname{argmax}}P(h|D) \tag{2.6}$$

应用贝叶斯规则得到：

$$h_{MAP}=\underset{h\in H}{\operatorname{argmax}}\frac{P(D|h)P(h)}{P(D)} \tag{2.7}$$

因为 $P(D)$是一个不依赖于 h 的常量，所以方程(2.7)可以简化为：

$$h_{MAP}=\underset{h\in H}{\operatorname{argmax}}P(D|h)P(h) \tag{2.8}$$

比如，有这样一个医疗诊断问题（Mitchell，1997），其中有两个可选的假设：①病人有癌症；②病人无癌症。可用的数据来自于一个化验测试，它有两种可能的输出：⊕（正）和⊖（负）。我们有先验知识：在我国所有人口中只有 0.008 的人患有该疾病。另外，该化验测试只是对该病的一个不完全准确的预计。该测试针对确实有病的患者有 98%的可能正确返回⊕结果，而对无该病的患者有 97%的可能正确返回⊖结果。除此以外，测试返回的结果都是错误的。假定现有一新病人，化验测试返回了⊕结果。是否应将该病人断定为有癌症呢？

由题可知，实际上就是要计算极大后验假设 h_{MAP}。应用方程(2.8)有：

$P(\oplus|\text{cancer})P(\text{cancer})=(0.98)\times(0.008)\approx 0.0078$

$P(\oplus|\neg\text{cancer})P(\neg\text{cancer})=(0.03)\times(0.992)\approx 0.0298$

因此，$h_{MAP}=\neg\text{cancer}$。

确切的后验概率可将上面的结果归一化得到，即：

$$P(\text{cancer}|\oplus)=\frac{0.0078}{0.0078+0.0298}\approx 0.21$$

$$P(\neg\text{cancer}|\oplus)=\frac{0.0298}{0.0078+0.0298}\approx 0.79$$

该步骤的根据在于，贝叶斯公式说明后验概率就是上面的量除以数据 $P(\oplus)$。虽然 $P(\oplus)$没有作为问题陈述的一部分直接给出，但因为已知 $P(\text{cancer}|\oplus)$和 $P(\neg\text{cancer}|\oplus)$的和必为 1(即该病人要么有癌症，要么没有癌症)，因此可以进行归一化。

注意，虽然有癌症的后验概率(0.21)比先验概率(0.008)要大，但最可能的假设仍然为该病人没有癌症。同时可以看出，贝叶斯推理的结果很大程度上依赖于先验概率，要直接应用该方法必须先获取该值。

由此可见，贝叶斯方法提供了推理的一种概率手段。它假定待考查的变量遵循某种概率分布，且可根据这些概率及已观察到的数据进行推理，从而做出最优的决策。贝叶斯方法不仅能够计算显式的假设概率，还能为理解多数其他方法提供一种有效的手段。

贝叶斯方法的特点主要包括：增量式学习的特点；先验知识可以与观察到的实例一起决定假设的最终概率的特点；允许假设做出不确定性预测的特点；对新实例的分类可由多个假设以它们的概率为权重一起做出预测的特点，等等。

在实践中应用贝叶斯方法的困难在于：①它需要概率的初始知识。当这些概率预先未知的时候，可以基于背景知识、预先准备好的实例以及关于基准分布的假定来估计这些概率。②一般情况下确定贝叶斯最优假设的计算代价比较大。不过，在某些特定情形下，这种计算代价可以被大大降低。

2.5　朴素贝叶斯分类器

到目前为止，讨论的问题都是“给定训练实例集，最可能的假设是什么?”实际上，该问题通常与分类问题紧密相关，即“给定训练实例集，对新实例最有可能的分类是什么?”显然，分类问题可以直接用 MAP 假设来解决。

设每个实例 x 可由属性值的合取描述，而类标记 c 从某有限集合 C 中取值。现提供一训练实例集和一测试实例$<a_1,a_2,\cdots,a_m>$，然后预测它的类标记。

应用 MAP 假设分类新实例 x 的目标是在给定描述实例的属性值$<a_1,a_2,\cdots,a_m>$的情况下，得到最可能的类标记 $c(x)$。应用方程(2.8)得到：

$$c(x)=\underset{c\in C}{\operatorname{argmax}}P(a_1,a_2,\cdots,a_m|c)P(c) \tag{2.9}$$

现在要做的就是基于该训练实例集，估计方程(2.9)中的两个概率值。估计每个 $P(c)$值很容易，只要计算每个类标记 c 出现在训练实例集中的频率就可以。然而，直接估计每个 $P(a_1,a_2,\cdots,a_m|c)$值非常困难，原因在于：首先，完整估计 $P(a_1,a_2,\cdots,a_m|c)$值的时间复杂度相当于学习一个贝叶斯网络，是一个 NP - hard 问题；其次，这些 $P(a_1,a_2,\cdots,a_m|c)$值的数量等于可能实例的数量乘以可能类的数量。因此，为获得合理的估计，实例空间中每个实例必须出现多次，这就要求训练实例集非常大。

为此，朴素贝叶斯(Naive Bayes，简记为 NB)分类器假定：在给定实例类标记的前提下，实例的属性值之间是相互条件独立的。也就是说，在给定实例类标记的情况下，观察到的联合概率正好是每个单独属性值概率的乘积。具体的数学表达形式如下：

$$P(a_1,a_2,\cdots,a_m|c)=\prod_{i=1}^{m}P(a_i|c)$$

将其代入方程(2.9)中，可得到朴素贝叶斯分类器的分类公式：

$$c(x)=\underset{c\in C}{\operatorname{argmax}}P(c)\prod_{i=1}^{m}P(a_i|c) \tag{2.10}$$

式中，a_i 为 x 的第 i 个属性值；概率 $P(c)$ 和 $P(a_i|c)$ 可以通过计算训练实例集中不同类标记和属性值组合的出现频率来简单估计，具体公式如下：

$$P(c)=\frac{\sum_{j=1}^{n}\delta(c_j,c)}{n} \tag{2.11}$$

$$P(a_i|c)=\frac{\sum_{j=1}^{n}\delta(a_{ji},a_i)\delta(c_j,c)}{\sum_{j=1}^{n}\delta(c_j,c)} \tag{2.12}$$

式中，n 为训练实例的个数；c_j 为第 j 个训练实例的类标记；a_{ji} 为第 j 个训练实例的第 i 个属性值；$\delta(c_j,c)$ 为一个二值函数，当 $c_j=c$ 时其值为 1，否则为 0。

在大多数情况下，上述这种基于频率比例的方法是对概率的一个良好估计，但当接近零频率属性值出现的时候，这种概率估计方法就会产生一个有偏差的过低估计(Under Estimate)概率。更极端的情况是，当零频率属性值出现的时候，就会使得某个概率值为 0，进而导致由式(2.10)计算的整个量为 0。为避免这些问题，Laplace 估计常常被用来平滑上述公式(2.11)、公式(2.12)得到的概率。重写公式(2.11)和公式(2.12)得到：

$$P(c)=\frac{\sum_{j=1}^{n}\delta(c_j,c)+1}{n+q} \tag{2.13}$$

$$P(a_i|c)=\frac{\sum_{j=1}^{n}\delta(a_{ji},a_i)\delta(c_j,c)+1}{\sum_{j=1}^{n}\delta(c_j,c)+n_i} \tag{2.14}$$

式中，q 为类标记的个数；n_i 为训练实例第 i 个属性的取值个数。

可见，学习朴素贝叶斯分类器具有非常低的时间复杂度，仅为 $O(nm)$。此外，不同 $P(a_i|c)$ 值的数量只是不同的属性值的数量乘以不同类标记的数量，这比上面提到的 $P(a_1,a_2,\cdots,a_m|c)$ 值的数量要小得多。因此，朴素贝叶斯分类器并不要求有非常大的训练实例集。实践也证明，它在小训练实例集上的分类性能仍然比较好。

不同于其他贝叶斯网络分类器，朴素贝叶斯分类器没有明确的假设空间搜索过程，可能假设的空间为可赋予不同的 $P(c)$ 和 $P(a_j|c)$ 项的可能值的集合。因此，只需要简单地计算训练实例中不同类和属性值组合的出现频率。而且，当所需的条件独立性假设能够被满足时，朴素贝叶斯分类就等于 MAP 分类。

比如，应用朴素贝叶斯分类器来解决这样一个分类问题(Mitchell，1997)：根据天气状况来判断某天是否适合于打网球。给定如表 2-1 所示的 14 个训练实例，其中每一天由属性 *Outlook*，*Temperature*，*Humidity* 和 *Wind* 来描述，类属性为 *PlayTennis*。

表 2-1　预测 *PlayTennis* 的训练实例

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*

现有一测试实例 x：<$Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong$>，问这一天是否适合于打网球。

显然，我们的任务就是要预测这个新实例的类属性 *PlayTennis* 的取值（*yes* 或 *no*）。为此，构建如图 2-3 所示的朴素贝叶斯分类器。图中的类结点 C 表示类属性 *PlayTennis*，其他 4 个结点 A_1, A_2, A_3, A_4 分别代表 4 个属性 *Outlook*, *Temprature*, *Humidity* 和 *Wind*，类结点 C 是所有属性结点的父亲结点，属性结点和属性结点之间没有任何的依赖关系。根据方程 (2.10)有：

$$c(x)=\underset{c\in\{yes,no\}}{\operatorname{argmax}}P(c)P(sunny|c)P(cool|c)P(high|c)P(strong|c)$$

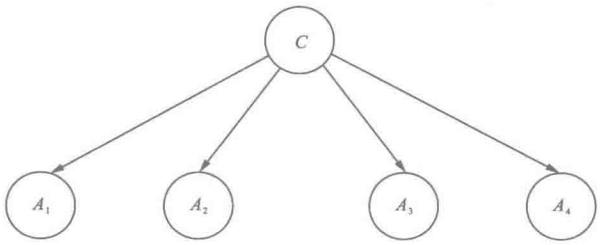

图 2-3　朴素贝叶斯分类器的结构

为计算 $c(x)$，需要从表 2-1 所示的 14 个训练实例中估计出概率：$P(yes)$, $P(sunny|yes)$, $P(cool|yes)$, $P(high|yes)$, $P(strong|yes)$, $P(no)$, $P(sunny|no)$, $P(cool|no)$, $P(high|$

no)和 $P(strong|no)$。根据公式(2.13)、公式(2.14),具体计算如下:

$P(yes)=(9+1)/(14+2)=10/16$

$P(sunny|yes)=(2+1)/(9+3)=3/12$

$P(cool|yes)=(3+1)/(9+3)=4/12$

$P(high|yes)=(3+1)/(9+2)=4/11$

$P(strong|yes)=(3+1)/(9+2)=4/11$

$P(no)=(5+1)/(14+2)=6/16$

$P(sunny|no)=(3+1)/(5+3)=4/8$

$P(cool|no)=(1+1)/(5+3)=2/8$

$P(high|no)=(4+1)/(5+2)=5/7$

$P(strong|no)=(3+1)/(5+2)=4/7$

所以有:

$P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)=0.0069$

$P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)=0.0191$

可见,朴素贝叶斯分类器将此实例分类为 *no*。将上述概率归一化,可得到朴素贝叶斯分类器分类此实例为 *no* 的概率是 0.0191/(0.0069+0.0191),即 0.7346。

又比如,应用朴素贝叶斯分类器来解决这样一个分类问题(Han,Kamber,2001):根据顾客的基本情况来判断其是否会买电脑。给定如表 2-2 所示的 14 个训练实例,其中每一个顾客由属性 *Age*,*Income*,*Student* 和 *Credit_rating* 来描述,类属性为 *Buys_computer*。

表 2-2 预测 *Buys_computer* 的训练实例

RID	*Age*	*Income*	*Student*	*Credit_rating*	*Buys_computer*
1	<=30	*high*	*no*	*fair*	*no*
2	<=30	*high*	*no*	*excellent*	*no*
3	30~40	*high*	*no*	*fair*	*yes*
4	>40	*medium*	*no*	*fair*	*yes*
5	>40	*low*	*yes*	*fair*	*yes*
6	>40	*low*	*yes*	*excellent*	*no*
7	31~40	*low*	*yes*	*excellent*	*yes*
8	<=30	*medium*	*no*	*fair*	*no*
9	<=30	*low*	*yes*	*fair*	*yes*
10	>40	*medium*	*yes*	*fair*	*yes*
11	<=30	*medium*	*yes*	*excellent*	*yes*
12	31~40	*medium*	*no*	*excellent*	*yes*
13	31~40	*high*	*yes*	*fair*	*yes*
14	>40	*medium*	*no*	*excellent*	*no*