基于 Bayes 定理的分类规则研究*

吕安民[1,2]，牛晓太[3]，郭建忠[4]

(1. 清华大学 水沙科学教育部重点实验室，北京 100084；2. 华北水利水电学院，河南 郑州 450011；3. 郑州航空工业管理学院，河南 郑州 450005；4. 郑州测绘学院，河南 郑州 450052)

摘 要：研究了利用 Bayes 定理发现分类规则的方法，用 Bayes 定理可以发现分类规则，然后用分类规则进行数据分类。结合实例针对概念性数据集及包含数值性属性和概念性属性的数据集两种情况进行讨论。通过实例说明 Bayes 定理是数据挖掘中一种有效的数据分类方法。

关键词：数据挖掘；Bayes 定理；分类规则

中图法分类号：TP311 **文献标识码**：A **文章编号**：1001-3695(2006)02-0024-02

An Investigation of Classfication Rule Based on Bayes Theorem

LV An-min[1,2], NIU Xiao-tai[3], GUO Jian-zhong[4]

(1. *The Key Laboratory of Hydraulic & Hydropower Engineering, Ministry of Education, Tsinghua University, Beijing* 100084, *China;* 2. *North China Institute of Water Conservancy & Hydroelectric Power, Zhengzhou Henan* 450011, *China;* 3. *Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou Henan* 450005, *China;* 4. *Zhengzhou Surveying & Mapping Institute, Zhengzhou Henan* 450052, *China*)

Abstract: Investigates a method of classification rule with Bayes theorem. We can found classification rules using Bayes theorem, then use the classification rules to classify new data set. The paper discusses two instancees: concept attribute, concept attribute and numerical attribute. Bayes theorem is a effective method of classification in data mining through examples.

Key words: Data Mining; Bayes Theorem; Classification Rule

从大规模数据库中分析大量的数据，从中发现隐含的、潜在的、有意义有价值的规则知识，即基于数据库的知识发现近年来在理论模型、发现方法和实际应用等方面都得到了很大的发展。数据库中隐含的规则知识主要有分类规则、特征规则、函数依赖、相互关系等。发现知识的方法主要有两大类：机器学习方法和统计分析方法。前者主要是定性分析，即从知识基表中提取所有或大多数元组满足的规则；后者主要是定量分析，提取属性之间的函数依赖与相关关系等。这些发现方法中使用的主要算法有归纳学习、神经网络、基于集合理论的方法、基于范例的推理(CBR)、遗传算法和统计算法等。

数据分类(Classification)作为数据挖掘的一个重要的内容，在统计学、机器学习、神经网络和专家系统中得到了较早的研究，但只是近些年来，人们才将它与数据库技术结合起来解决实际问题。数据分类实际上就是从数据库对象中发现共性，并将数据对象分成不同几类的一个过程。在数据分类中，一个样本数据库被当作一个训练集，训练集中的每个元组都保持总数据库的所有属性，并且都有一个类的标识符与之相联系。分类的目标首先是对训练数据进行分析，使用数据的某些特征属性，给出每个类的准确描述，然后使用这些描述，对总数据库中的其他数据进行分类，或者是为每个类产生更好的描述，也即分类规则。数据分类的方法很多[2,3]，包括决策树方法、统计学方法、神经网络方法、最近邻居方法等等。在统计学方法中Bayes 定理是一种有效的数据分类方法。

收稿日期：2004-12-07；修返日期：2005-04-06

基金项目：国家计委 2000 年高科技应用项目；中国博士后二等科学基金项目(2004035020)；2005 年河南省杰出青年科学基金项目

1 Bayes 定理

定义：设(Ω, F, P)为概率空间，如果$A_i \Omega (i=1,2,\ldots,n)$，$A_i \cap A_j = \quad (i \neq j)$，且$\bigcup_{i=1}^{n} A_i = \Omega$

则称$A_1, A_2, \ldots, A_n$为Ω的一个有限剖分。如果$A_i \Omega (i=1,2,\ldots)$，$A_i \cap A_j = \quad$，且$\bigcup_{i=1}^{\infty} A_i = \Omega$

则称$A_1, A_2, \ldots, A_n$为Ω的一个有列无穷剖分。图 1 是$n=8$的剖分。

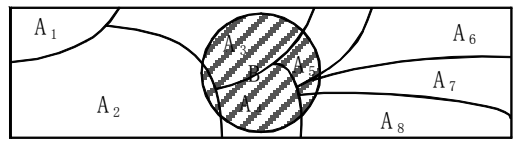

图 1 n=8 的剖分

Bayes 定理：设(Ω, F, P)为概率空间，$A_i \in F (i=1,2,\ldots,n)$为$\Omega$的一个有限剖分且$P(A_i) > 0 (i=1,2,\ldots,n)$，则对任意$B \in F$且$P(B) > 0$，有

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

其中$i=1,2,\ldots,n$。符号$P(B)$表示事件B发生的可能性，符号$P(A_i|B)$表示B事件发生以后某一特定事件A_i发生的可能性。

Bayes 可以判断一个事件属于A_i的某一类，其判断的依据是根据这个事件属于某一类可能性的大小。既然每个可能性的分母都一样，可以在标准化时消去，因此计算时可以忽略。$P(A_i)$可以用A_i出现的总次数进行估计。下面的主要问题是

如何估计 $p(B|A_i)$。

如果假定 B 中的每个事件是独立的, 那么

$$p(B|A_i) = \prod_{k=1}^{l} P(B_k|A_i)$$

其中 $k = 1, 2..., l$, l 表示事件 B 的个数, $p(B_k|A_i)$ 的值可以通过在 A_i 类中出现的某一事件的次数来计算。

如果假定 B 中的每个事件服从正态分布, 那么

$$p(B|A_i) = \frac{1}{2\pi m} e^{-\frac{(B_j-\mu)^2}{2m^2}}$$

其中 m 是标准差, μ是平均数, $k = 1, 2..., l$, B_j 表示事件 B 的观测值。

我们把 $P(B|A_i)$ 称为"原因"概率, 而称 $P(A_i|B)$ 为事后概率。Bayes 定理告诉我们: "事后"概率可通过一系列的"原因"概率求得。

例如在诊病问题中, 若我们从病理或长期积累的经验中, 知道了有多种病因(这里假定这些病因满足定义剖分要求, 即它们之间的不相容性)会产生某症状, 并且知道这些"原因"概率, 我们可以用 Bayes 定理解决两类问题:

(1) 假若在一次诊病的病历中已经出现该症状, 问其最大可能的原因是什么?

(2) 用 Bayes 定理可以发现分类规则, 然后用分类规则去判断新的病例。

本文主要研究问题(2), 即如何用 Bayes 定理可以发现分类规则, 然后用分类规则去判断新的病例。这里的分类规则是指用已知的数据计算出当事件 A_i 出现时事件 B 出现的概率, 然后求出一个新事件属于事件 A_i 的可能性大小, 从而判断其属于 A_i 的某一类。

2　应用实例

2.1　概念性数据集

下面是应用 Bayes 定理从一个病例数据库中发现肺炎和肺结核两种疾病的分类规则的例子。每个病例都含有五种症状: 发烧、咳嗽、X 光所见阴影、血沉和听诊。肺炎和肺结核两种疾病的部分病例见表 1(病例数据来自文献[1])。

表 1　肺炎和肺结核病例集

病例号	发烧	咳嗽	X 光所见	血沉	听诊	诊断结果
1	高	剧烈	片状	正常	水泡音	肺炎
2	中	剧烈	片状	正常	水泡音	肺炎
3	低	轻微	点状	正常	干鸣音	肺炎
4	高	中度	片状	正常	水泡音	肺炎
5	中	轻微	片状	正常	水泡音	肺炎
6	无	轻微	索条状	正常	正常	肺结核
7	高	剧烈	空洞	快	干鸣音	肺结核
8	低	轻微	索条状	正常	正常	肺结核
9	中	轻微	点状	快	干鸣音	肺结核
10	低	中度	片状	快	正常	肺结核

我们假定每种症状都同等重要, 而且它们之间都彼此独立。由表 1 可得表 2。表 2 中的上半部分就是计算每一种属性出现的次数, 比如说发高烧的情况有三例, 其中二例属于肺炎, 一例属于肺结核。上半部分最后一列"诊断结果"是肺炎和肺结核出现的总次数。表 2 中的下半部分是由其上半部分计算而来的。比如在第一列中, 肺炎总共出现的次数为五, 发高烧时出现肺炎的次数为二, 所以表 2 中下半部分第一列"高"后面的数字是 2/5, 这就是 $p(B_k|A_i)$ 中的一个值。

表 2　每种症状和病例各种可能性统计结果

	发烧		咳嗽		X 光所见		血沉		听诊		诊断结果	
	肺炎	肺结核	肺炎	肺结核	肺炎	肺结核	肺炎	肺结核	肺炎	肺结核	肺炎	肺结核
高	2	1	剧烈	2	1	片状	4	1	正常	5	2	水泡音	4	0	5	5
中	2	1	中度	1	1	点状	1	1	快	0	3	干鸣音	1	2		
低	1	2	轻微	2	3	索条状	0	2				正常	0	3		
无	0	1				空洞	0	1								
高	2/5	1/5	剧烈	2/5	1/5	片状	4/5	1/5	正常	5/5	2/5	水泡音	4/5	0/5	5/10	5/10
中	2/5	1/5	中度	1/5	1/5	点状	1/5	1/5	快	0/5	3/5	干鸣音	1/5	2/5		
低	1/5	2/5	轻微	2/5	3/5	索条状	0/5	2/5				正常	0/5	3/5		
无	0/5	1/5				空洞	0/5	1/5								

现在假定有一个新病人, 其五种病状的表现见表 3。

表 3　一个新病例

发烧	咳嗽	X 光所见	血沉	听诊	诊断结果
高	轻微	片状	正常	干鸣音	?

把 A_i 看成"诊断结果是肺炎"这一事件, 事件 B 是各个病状属性值的组合, 发烧 = 高, 咳嗽 = 轻微, X 光可见 = 片状, 血沉 = 正常, 听诊 = 干鸣音, 并把这五个事件分别用 B_1, B_2, B_3, B_4, B_5 表示, 假定这五个事件是相互无关的, 它们组合在一起后, 诊断结果为肺炎的可能性为

$$P(\text{肺炎}|B) = P(B_1|\text{肺炎})\,P(B_2|\text{肺炎})\,P(B_3|\text{肺炎})\,P(B_4|\text{肺炎})\,P(B_5|\text{肺炎})\,P(\text{肺炎})$$

在上式中 $P(\text{肺炎})$ 是不知道任何症状情况下肺炎出现的可能性, 在这个例子中 $P(\text{肺炎}) = 5/10$。所以对于表 3, 诊断结果是肺炎的可能性为

$$P(\text{肺炎}|B) = 2/5 \times 2/5 \times 4/5 \times 5/5 \times 1/5 \times 5/10 = 0.014\,4$$

同理, 诊断结果是肺结核的可能性为

$$1/5 \times 3/5 \times 1/5 \times 2/5 \times 2/5 \times 5/10 = 0.001\,92$$

将上面两个数字进行标准化可得

诊断结果是肺炎的可能性 = 0.014 4/(0.014 4 + 0.001 92) = 88.2%

诊断结果是肺结核的可能性 = 0.001 92/(0.014 4 + 0.001 92) = 11.8%

2.2　包含数值性属性和概念性属性的数据集

表 4 是包含数值性属性和概念性属性的数据集。对于概念性属性值, 其计算方法和表 2 所示的一样。对于数值性属性值, 要计算其中数和标准差, 其标准差计算公式为

$$m = \frac{[vv]}{n-1}$$

其中, n 代表所有病例数, 这里 $n = 10$, v 代表中数和每一个观测值之差, vv 表示每一个差数的平方, $[vv]$ 表示每一个差数的平方和。

表 4　肺炎和肺结核病例集

病例号	发烧	咳嗽	X 光所见	血沉	听诊	诊断结果
1	41	剧烈	片状	75	水泡音	肺炎
2	38.3	剧烈	片状	70	水泡音	肺炎
3	37.6	轻微	点状	68	干鸣音	肺炎
4	40	中度	片状	72	水泡音	肺炎
5	38.5	轻微	片状	65	水泡音	肺炎
6	36.7	轻微	索条状	74	正常	肺结核
7	39.6	剧烈	空洞	85	干鸣音	肺结核
8	37.4	轻微	索条状	76	正常	肺结核
9	38.4	轻微	点状	90	干鸣音	肺结核
10	37.5	中度	片状	92	正常	肺结核

当发烧为 40, 诊断结果是肺炎的可能性为

$$p(\text{发烧} = 40|\text{肺炎}) = \frac{1}{2\pi \times 1.4} e^{-\frac{(40-39.1)^2}{2\times 1.4^2}} = 0.231\,7$$

同样, 当血沉为 70, 诊断结果是肺炎的可能性为

$$f(\text{血沉} = 70|\text{肺炎}) = \frac{1}{2\pi \times 3.8} e^{-\frac{(70-70)^2}{2\times 3.8^2}} = 0.105$$

(下转第 72 页)