文章编号:1001 - 9081(2005)12 - 2882 - 03

基于贝叶斯方法的决策树分类算法

樊建聪[1],张问银[2],梁永全[1]
(1. 山东科技大学 信息科学与工程学院,山东 青岛 266510;
2. 临沂师范学院 计算机系,山东 临沂 276005)
(howdoyoudo07@ yahoo. com. cn)

摘 要:针对数据挖掘的特点和本质,充分利用贝叶斯方法和决策树分类的优点,将贝叶斯的先验信息方法与决策树分类的信息增益方法相结合,提出了一种新的数据挖掘分类算法(BD1.0 算法),并对此算法进行了设计和分析。实验分析表明,该算法可以处理不一致或者不完整数据等"脏数据",比单纯使用贝叶斯方法或决策树方法具有更高的准确率,而且与 C4.5 算法具有近似的时间复杂度。

关键词:数据挖掘;分类;贝叶斯原理;决策树
中图分类号: TP301.6 **文献标识码**:A

Decision tree classification algorithm based on Bayesian method

FAN Jian-cong[1], ZHANG Wen-yin[2], LIANG Yong-quan[1]
(1. *College of Information Science and Technology, Shandong University of Science and Technology, Qingdao Shandong* 266510, *China*;
2. *Department of Computer, Linyi Normal University, Linyi Shandong* 276005, *China*)

Abstract: According to the characteristic and essence of data mining and taking advantage of Bayesian method, a new classification method named BD1.0 algorithm was presented. This method combined the prior information and information gain method of decision tree. The design and analysis of the algorithm was introduced too. The experiment results show that the algorithm can deal with dirty data such as incomplete data or inconsistent data, and it is more accurate than only useing Bayesian method or decision tree. It has approximate time complexity with C4.5 algorithm.

Key words: data mining; classification; Bayesian principle; decision tree

1 分类方法及算法概述

分类是数据挖掘的任务之一。分类就是找出一个类别的概念描述,它代表了这类数据的整体信息,即该类的内涵描述,并用这种描述来构造模型,一般用规则或决策树模式表示。分类利用训练数据集通过一定的算法而求得分类规则,可被用于规则描述和预测。分类的目的是通过一定的学习获取一个分类函数或分类模型(也常称作分类器),该模型能把数据集中的数据项映射到给定类别中的某一个[1]。

要构造分类器,需要有一个训练样本数据集作为输入。训练集由一组数据库记录或元组构成,每个元组是一个由有关字段(又称属性或特征)值组成的特征向量。此外,训练样本还有一个类别标记。一个具体样本的形式可为 $(v_1, v_2, \cdots, v_n; c)$,其中 v_i 表示属性值,c 表示类别。

目前,分类器的构造方法有多种,应用比较广泛的有决策树方法、统计方法、机器学习方法、神经网络方法、遗传算法、粗糙集以及模糊逻辑等[2]。不同的分类方法有不同的特点。通常,有三种对分类方法评价或比较的尺度:1)分类结果的准确度。它是用得最多的一种比较尺度,特别是对于预测型分类任务;2)分类计算的速度。计算速度依赖于具体的实现细节和硬件环境,在数据挖掘中,由于操作对象是海量数据,因此空间和时间的复杂度问题将是非常重要的一个环节;3)分类器对各种类型数据集的适应度。由于所分析的数据对象中,可能会存在不完整数据、噪声数据或不一致数据,或者数据的分布是稀疏的,因此一个好的分类器能够对各种类型的数据集有较强的适应能力。

统计方法主要有最近邻归类、基于事例的学习等,这些方法本质上是基于某种距离进行相应变换,得到具有另外一些参数的分类公式。统计学上主要用的基本距离公式有绝对值距离、欧氏距离、明斯基距离等。无论哪种基于距离的分类器,都需要对数据集中的所有数据进行两两检测,以确定一条记录是否与另一条记录在同一类别中,如果数据量非常大,这种检测将很耗时,因为每个数据都必须进行两次计算。

神经网络方法对数据集中的噪声数据有很好的处理性能,而且即使数据未经训练,也能发现对数据的分类模式。但是神经网络在运行时需要大量参数,这必然会增加人为地干预,致使分类速度和分类器的适应度差。

遗传算法、粗糙集等方法是基于规则的方法,这种方法都有一个共同的问题,就是对于连续数据会形成明显的截断面。这种截断面可能会把一些属于同一类别 A 的数据分到两个不同的类别 A 和 B 中,其中 B 是与 A 相邻近的类别。

决策树[3]分类算法中应用比较广泛的是 ID 算法和 C4.5、C5.0 算法[4,5]等,基于决策树的分类算法的一个最大的优点,同时也是最大的缺点是它在学习过程中不需要使用

收稿日期:2005 - 06 - 06;**修订日期**:2005 - 08 - 14

作者简介:樊建聪(1977 -),男,山东青岛人,助教,硕士,主要研究方向:人工智能应用、数据挖掘算法; 张问银(1972 -),男,山东临沂人,讲师,博士,主要研究方向:图像信息安全; 梁永全(1967 -),男,山东聊城人,教授,博士生导师,主要研究方向:人工智能理论及应用、数据库、多媒体.

者了解很多背景知识,只要训练样本能够用属性—结论式的方式表达出来即可。而在某些情况下,某些数据在分类时,需要考虑与它们相关的背景知识,尤其是一些类属不是很明显的数据,这对于分类的精度和准确度是很重要的。

2　贝叶斯方法与决策树方法

2.1　贝叶斯方法

贝叶斯方法的关键是使用概率表示各种形式的不确定性。在选择某事件面临不确定性时,在某一时刻假定此事件会发生的概率,然后根据不断获取的新的信息修正此概率。修正之前的概率称为先验概率,修正之后的概率称为后验概率。贝叶斯原理就是根据新的信息从先验概率得到后验概率的一种方法。通常用下面的式子表示贝叶斯原理[5]:

$$P(\theta_k \mid a) = \frac{P(a \mid \theta_k)P(\theta_k)}{\sum_{j=1}^{n} P(a \mid \theta_j)P(\theta_j)}$$

其中,θ_j 表示一个特定的事件;a 表示某一行动;$P(\theta_k)$ $(k = 1,2,\cdots,n)$ 是先验概率;先验概率通过对条件概率 $P(a \mid \theta_k)$ 加权平均的计算后,得到后验概率 $P(\theta_k \mid a)$。设某样本数据集中每个变量的特征向量为 $X = (x_1,x_2,\cdots,x_n)$,另外假设类别向量为 $C = (c_1,c_2,\cdots,c_l)$。分类的目的[6]是把特征向量 X 归入到某个类别 $c_i(i \in \{1,2,\cdots,l\})$ 中。于是可以选择后验概率最大的类别,即 $P(c_i \mid X) > P(c_j \mid X)$,其中 $i,j \in \{1,2,\cdots,l\}$。

2.2　决策树

决策树分类是以实例为基础的归纳分类算法,它主要从一组无次序、无规则的事例中推理出决策树表示形式的分类规则,并采用自顶向下的递归方式,在决策树的内部节点之间进行属性值的比较,在节点内部进行属性的选择,根据不同的属性值判断从节点向下的分支,在决策树的叶节点得出结论。如图 1 所示是一棵决策树。

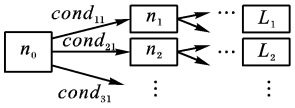

图 1　决策树

图 1 中,叶节点 L_i 表示类别,n_i 表示对某个属性值的测试,$cond_{ij}$ 表示测试条件。每一个类别可表示为节点的合取,所有类别可表示为叶节点的析取,即:

$L_i = n_{i1} \wedge n_{i2} \wedge \cdots \wedge n_{ik}, i = 1,2,\cdots$

$C = L_1 \vee L_2 \vee \cdots \vee L_t, t = 1,2,\cdots$

2.3　贝叶斯决策树

定义　贝叶斯决策树

在原有决策树 T 的基础上,在 T 中加入新的节点,此节点位于 T 的两个属性测试节点之间,能够根据贝叶斯原理进行函数计算,称之为贝叶斯节点,具有这样节点的决策树称作贝叶斯决策树。

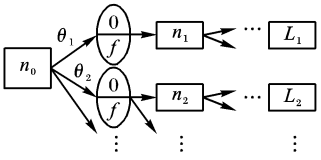

图 2　贝叶斯决策树

贝叶斯决策树结构如图 2 所示。贝叶斯节点分为两部分,分别是 0 值和 f 值。0 值表示此节点不进行任何计算,直接根据条件 θ 转向下一属性测试节点;f 值表示需要计算函数 f 的值,这里的函数 f 可以是朴素贝叶斯公式,也可以是其他贝叶斯公式,针对具体情况而定。也就是说,如果贝叶斯节点需要 f 值,则下一个属性节点的选择依赖于两点:

(1) 属性测试条件 θ;

(2) 函数 f 的值。

这两部分进行下一属性节点的选取时,都采用 IF…THEN… 的形式,即:

IF θ THEN …

IF f 取某个值 THEN…

3　算法的设计及其分析

3.1　算法设计

根据上面对分类问题的介绍与分析,可以设计使用贝叶斯方法的一种分类算法 BD1.0。此算法的基本思想是:

(1) 对于能够用信息增益方法确切选择某个属性的分支,选取贝叶斯节点的 0 或 f 值。其中信息增益的计算方法[2]为:

设集合 S,要把 S 中的数据样本分到 m 个不同类别 C_i $(i = 1,2,\cdots,m)$ 中,且 s_i 是类别 C_i 中的样本个数,则一个给定样本集的期望值是:

$$I(s_1,s_2,\cdots,s_m) = -\sum p_i \log_2(p_i) \tag{1}$$

其中 $p_i = s_i / |S|$。

设属性 A 具有 v 个不同的值 $\{a_1,a_2,\cdots,a_v\}$,用属性 A 将 S 划分为 v 个子集 $\{S_1,S_2,\cdots,S_v\}$,其中 S_j 包含 S 中的这样一些样本,它们在 A 上具有值 a_j。设 s_{ij} 是子集 S_j 中类 C_i 的样本数,则由 A 划分成子集的期望信息值是:

$$E(A) = \sum ((s_{1j} + \cdots + s_{mj}) / |S|) \cdot I(s_{1j},\cdots,s_{mj}) \tag{2}$$

其中,$I(s_{1j},\cdots,s_{mj}) = -\sum p_{ij} \log_2(p_{ij})$,$p_{ij} = s_{ij} / |S_j|$。

由式(1)和(2)可得信息增益值 $Gain(A)$ 为:

$Gain(A) = I(s_1,s_2,\cdots,s_m) - E(A)$

(2) 对于无法确定其分类类别的数据对象,或者属性值丢失的属性,选取 f 值。函数 f 的选取主要依据经验知识或者以前的实验结果确定其先验概率,根据概率判断先将其分到哪些类中,然后利用贝叶斯方法进行处理,确定后验概率,选取后验概率最大的那一类,此类即为数据对象所属的类别。根据上述分析,BD1.0 算法描述如下:

算法:使用贝叶斯方法的 BD1.0 算法

Input:给定一个数据集 $\{X_1,X_2,\cdots,X_n,\cdots\}$,其中每个 X_i 具有一个或多个属性 $X_{ij}(i,j = 1,2,\cdots,n,\cdots)$;

Output:对输入的数据集 $\{X_1,X_2,\cdots,X_n,\cdots\}$ 已划分到各相关的类别 L_j 中。显示或打印出各类数据。

(1) 确定要生成的类的数目 k 和各类别 $L_j(j = 1,2,\cdots,k)$,L_j 的确定依据事先给定的类的特征或属性;

(2) 使用信息增益方法首先确定要对哪个属性先进行判断,确定要进行分类的数据 $X_i(i = 1,2,\cdots)$ 的某个或某些属性,属性值与相应的类相关;

(3) IF 属性选择无二义性 and 数据分类无二义性

THEN 贝叶斯节点选取 0 值

ELSE 转(4);

(4) 按某种原则对 X_i 进行分类,若 X_i 确定对应某一类别 L_j,则将 X_i 划分到此类;否则,若 X_i 不能确定分到某一类别,而是与某些类都相关,则根据先验信息 $P(L_j)$ 先把它置入某一类,然后根据计算出的 $P(X_i \mid L_j)$ 和 $P(X_i)$ 来计算后验概率。若是根据 X_i 的 m 个属性进行分类,并且属性之间是独立的,则将 X_i 划分为 $X_{i1}, X_{i2}, \cdots, X_{im}$,于是 $P(X_i \mid L_j)$ 可表示为如下的乘积公式:

$$P(X_{i1} \mid L_j) \times P(X_{i2} \mid L_j) \times \cdots \times P(X_{im} \mid L_j)$$

从而可计算出后验概率:

$$P(L_j \mid X_i) = \frac{P(L_j)}{P(X_i)} \prod_{j=1}^{m} P(X_{ij} \mid L_j)$$

其中 $P(L_j)$ 表示 X_i 属于 L_j 类的概率,表示先验信息。

选取根据此式计算出的后验概率值最大的点划分到类别 L_j 中;

(5) 选取贝叶斯节点的 f 值,且 $f = P(L_j \mid X_i)$;

(6) 转到(3)。

3.2 算法分析

BD1.0 算法具有与其他决策树分类算法相似的优点:

(1) 产生的分类规则易于理解。决策树每个分支的节点的合取都对应一个分类规则,决策树分类算法最后可输出一个类别规则集,即树叶节点的析取式。

(2) 分类速度相对较快,最坏时间复杂度为 $O(n^3)$。BD1.0 算法主要进行两项工作,即判断是否要计算 f 值和是否要计算属性的后验概率值,而在最坏情况下,根据算法第(3)步的判断,需要计算所有数据的后验概率值。设共有 n 个数据,每个数据有 m 个属性,需要把这些数据分配到 k 个类别中,并且假设计算一个数据的后验概率值需要时间 τ_1,计算一次信息增益值需要时间 τ_2,则最坏情况下此算法的计算时间为:

$$(\tau_1 + m\tau_2) \cdot n \cdot k = nk\tau_1 + mn\tau_2$$

当 $m = n = k$ 时,计算时间为 $n^2\tau_1 + n^3\tau_2$,即最坏时间复杂度为 $O(n^3)$。

BD1.0 算法独具的优点:

(1) 分类的精度更高。对于用信息增益计算不能确定的属性选取,可通过贝叶斯方法解决。分类一般按照数据的某个或某些属性进行,假设某数据有两个需要计算信息增益值的属性 A_1 和 A_2,如果其增益值 $Gain(A_1)$ 和 $Gain(A_2)$,则属性的选择出现二义性,如果大量的数据具有这种二义性,则必然会影响数据的分类精度和准确率。BD1.0 算法通过贝叶斯方法利用先验信息,可以对这种情况作很好地处理。

(2) 分类的鲁棒性更强。BD1.0 算法能够更好地处理不一致、不完整和噪声等干扰数据。数据挖掘处理的是海量数据,由于主观和客观原因,在这些数据中不可避免的存在非正常数据。解决这类问题可以使用数据预处理方法[2],但这种解决方法十分耗时;也可以使用贝叶斯方法,根据对数据历史信息和专家的经验来消除不一致数据,平滑不完整数据,排除噪声数据等。

4 算法的验证

取一组平面扫描数据,利用 BD1.0 算法从这些数据中找出平面上的点和异常点。部分数据列于表 1 中。表 1 中的数据有 5 列,其中 ID 是数据序号,每一个数据有 4 个属性,分别用 *Attri*1、*Attri*2、*Attri*3、*Attri*4 表示,其中 *Attri*2 列中存在空值和缺失值。

表 1 部分实验数据表

ID	*Attri*1	*Attri*2	*Attri*3	*Attri*4
1	1.330	18	10.1	1720088
2	1.789	22	10.5	1720093
3	2.331	28	10.9	1720098
4	3.400	30	11.3	1720112
		…		
160	23.504	NULL	119.3	1721291
161	23.456		119.7	1721295
165	23.672	126	121.3	1721310
166	23.640	82	121.7	1721314
170	20.400	33	123.3	1721329
		…		

首先将所有数据分为两大类,即平面数据集类和异常数据集类,然后计算属性的信息增益值,得:

$Gain(Attri1) = 0.35$, $Gain(Attri2) = 0.5$, $Gain(Attri3) = 0.72$, $Gain(Attri4) = 0.11$

由于 *Attri*3 的信息增益值最大,先按 *Attri*3 进行分类。*Attri*3 中有不完全的属性数据值,则按 Bayes 方法进行计算,再进行分类。这样将这组数据分类后的结果如图 3 所示。

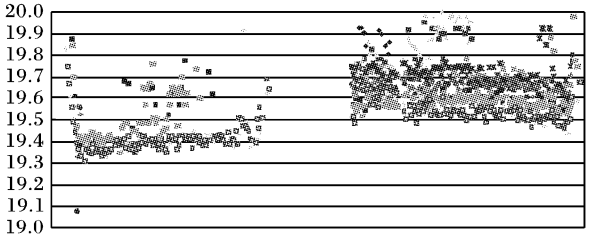

图 3 对数据分类后的结果

5 结语

使用贝叶斯方法最有争议之处就是先验信息的使用。先验信息来源于经验或者以前的实验结论,没有确定的理论依据作支持,因此在很多方面颇有争议。但是大量事实表明,对于大型数据集,贝叶斯分类表现出高准确率和运算的高速度。

参考文献:

[1] 高洪深. 决策支持系统理论·方法·案例[M]. 第 2 版. 北京: 清华大学出版社, 2000.56 - 89.

[2] HAN J, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 等译. 北京: 机械工业出版社, 2001. 70 - 95, 185 - 219.

[3] QUINLAN JR. Induction of decision trees [J]. Machine Learning, 1986, 1(1): 81 - 106.

[4] QUINLAN JR. Simplifying decision trees[J]. International Journal of Man Machine Studies, 1987, 27(3): 221 - 234.

[5] QUINLAN JR. Learning efficient classification procedures and their application to chess and games[A]. Machine Learning: an artificial intelligence approach [C]. San Mateo, CA: Morgan Kaufmann, 1983. 463 - 482.

[6] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002. 65 - 122.

[7] CODD EF. Providing OLAP(On-Line Analytical Processing) to User-Analysts: An IT Mandate[R]. Technical Report, Technical Report, IBM, San Jose, CA, 1993.

[8] 杨炳儒, 黄绍君. 知识发现系统的研究进展与成果综述[J]. 人工智能进展, 2000, 1(1): 334 - 340.

[9] 茆诗松. 贝叶斯统计[M]. 北京: 中国统计出版社, 1999. 12 - 83.