一种基于强属性限定的贝叶斯分类模型

王 峻[1,2]

(1.合肥工业大学,安徽 合肥 230009;2.淮南师范学院,安徽 淮南 232001)

摘 要 朴素贝叶斯分类模型一种简单而高效的分类模型,但它的条件独立性假设使其无法将属性间的依赖表达出来,影响了它分类的正确率。属性间的依赖关系与属性本身的特性有关,有些属性的特性决定了其他属性必然依赖于它,即强属性。文中通过分析属性相关性的度量和贝叶斯定理的变形公式,介绍了强属性的选择方法,通过在强弱属性之间添加增强弧以弱化朴素贝叶斯的独立性假设,扩展了朴素贝叶斯分类模型的结构。在此基础上提出一种基于强属性限定的贝叶斯分类模型SANBC。实验结果表明,与朴素贝叶斯分类模型相比,SANBC分类模型具有较高的分类正确率。

关键词 朴素贝叶斯;贝叶斯定理;属性相关性

中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)02-0205-03

A Restricted Bayesian Classification Model Based on Strong Attributes

WANG Jun[1,2]

(1.Hefei University of Technology,Hefei 230009,China;
2.Huainan Normal University,Huainan 232001,China)

Abstract:Naive Bayesian classification model is a simple and effective classification model, but its attribute independence assumption makes it unable to express the dependence among attributes, and affects its classification accuracy. The inter-dependence between attributes is closely related to their features, i.e. the features of some entail the others' dependence upon them-strong attributes. The paper presents SANBC(A Restricted Bayesian Classification Mmodel Based on Strong Attributes) following the extension of structure of naive Bayesian classification model, through the analysis of a variant of Bayes theorem, the evaluation of condition attribute with correlation, and the instruction of the selection of strong attributes and the attribute independence assumption that naive Bayesian classification model can be weakened through the adding of highlighting lines between strong and weak attribute. Compared with Bayesian classification model, experimental results show SANBC has higher accuracy.

Key words:naive Bayes;Bayes theorem;attribute with correlation

0 引 言

朴素贝叶斯分类模型是一种最简单的、有效的并且在实际使用中很成功的分类模型,其性能可以与神经网络、决策树相媲美。朴素贝叶斯分类模型基于假定特征向量的各分量间相对于决策变量是相对独立的,即条件独立性假设。但是这个限制过于严格,在实际的应用中,通常各属性变量之间常常具有明显的依赖性。属性间的依赖关系与属性本身的特性有关,有些属性本身所具有的特性决定了其他属性必然依赖于它,因此通过对属性空间的搜索,找出一些对其他属性具有较强影响的属性,通过在强属性与其他属性之间添加增强弧反映强属性与其他属性间的依赖关系。

文中从分析特征属性间的相关性出发,找出对其他属性有较强影响的属性,提出一种新的基于强属性限定的贝叶斯分类模型SANBC(A Restricted Bayesian Classification Model Based on Strong Attributes),给出构造SANBC的算法,并通过实验与朴素贝叶斯分类器进行比较。

1 朴素贝叶斯分类模型

贝叶斯分类模型是一种基于统计方法的分类模型,贝叶斯定理是贝叶斯学习方法的理论基础。朴素贝叶斯分类模型在贝叶斯定理的基础上,通过条件独立性假设,降低计算开销,预测未知数据样本属于最高后验概率的类。

令 $S=\{A_1,A_2,\cdots,A_n,C\}$ 为离散变量的有限集,其中 $A_1,A_2,\cdots,A_n$ 是属性变量,$C=\{c_1,c_2,\cdots,c_m\}$ 是类变量,a_i 是属性 A_i 的取值,实例 $x_i=(a_1,a_2,\cdots,$

收稿日期:2006-05-12

作者简介:王 峻(1967-),男,安徽淮南人,硕士,讲师,研究方向为数据挖掘。

a_n) 属于类 c_j 的概率,由贝叶斯定理可表示为:

$$P(c_j \mid a_1,a_2,\cdots,a_n) = \frac{P(a_1,a_2,\cdots,a_n \mid c_j)\cdot P(c_j)}{P(a_1,a_2,\cdots,a_n)} = \alpha \cdot P(c_j)\cdot P(a_1,a_2,\cdots,a_n \mid c_j) = \alpha \cdot P(c_j)\cdot \prod_{i=1}^{n} P(a_i \mid a_1,a_2,\cdots,a_{i-1},c_j)$$

其中 α 是正则化因子,$P(c_j)$ 是类 c_j 的先验概率,$P(c_j \mid a_1,a_2,\cdots,a_n)$ 是类 c_j 的后验概率。

根据贝叶斯最大后验规则,对于给定某一实例 $x_i = (a_1,a_2,\cdots,a_n)$,朴素贝叶斯分类器选择使后验概率 $P(c_j \mid a_1,a_2,\cdots,a_n)$ 最大的类作为该实例的类标签。

2 SANBC 分类模型

2.1 贝叶斯定理的变形公式

令 S_1 和 S_2 是属性集 $\{A_1,A_2,\cdots,A_n\}$ 的一个划分,s_1 和 s_2 分别是属性集 S_1 和 S_2 的取值,实例 $\{a_1,a_2,\cdots,a_n\}$(或表示为(s_1,s_2)) 属于类 c_j 的概率,可由贝叶斯定理的变形公式[1~3] 表示为:

$$P(c_j \mid s_1,s_2) = \frac{P(s_2 \mid c_j,s_1)}{P(s_2 \mid s_1)}\cdot P(c_j \mid s_1) = \beta \cdot P(s_2 \mid c_j,s_1)\cdot P(c_j \mid s_1) \quad (1)$$

其中 β 是一个正则化因子,假设 $s_1 = \{ak_1,ak_2,\cdots,ak_m\}$,$s_2 = \{al_1,al_2,\cdots,al_{n-m}\}$,并且在给定 c_j 和 s_1 时,s_2 中的各属性是条件独立的,则式(1) 可表示为

$$P(c_j \mid s_1,s_2) = \beta\cdot P(c_j \mid ak_1,ak_2,\cdots,ak_m)\cdot P(s_2 \mid c_j,ak_1,ak_2,\cdots,ak_m) = \beta \cdot P(c_j \mid ak_1,ak_2,\cdots,ak_m)\cdot \prod_{i=1}^{n-m} P(al_i \mid c_j,ak_1,ak_2,\cdots,ak_m) \quad (2)$$

式(2) 是假设在较少的属性(即 $al_1,al_2,\cdots,al_{n-m}$) 之间是条件独立的,该假设比朴素贝叶斯独立性假设要弱,从而通过贝叶斯变形公式弱化了朴素贝叶斯独立性假设。这个假设的强弱取决于强属性集 $S_1 = \{Ak_1,Ak_2,\cdots,Ak_m\}$ 中属性的个数,强属性集 S_1 中属性的个数越多,条件独立性假设就越弱。又由于

$$P(c_j \mid ak_1,ak_2,\cdots,ak_m) = \gamma\cdot P(c_j)\cdot \prod_{i=1}^{m} P(ak_i \mid c_j,ak_1,ak_2,\cdots,ak_{i-1}) \quad (3)$$

其中 γ 是一个正则化因子,将式(3) 代入式(2) 等号的右侧,得

$$P(c_j \mid s_1,s_2) = \beta\cdot\gamma\cdot P(c_j)\prod_{i=1}^{m} P(ak_i \mid c_j,ak_1,ak_2,\cdots,ak_{i-1})\cdot \prod_{z=1}^{n-m} P(al_z \mid c_j,ak_1,ak_2,\cdots,ak_m) = \beta\cdot\gamma\cdot P(c_j)\cdot \prod_{i=1}^{m} P(ak_i \mid c_j,K(ak_i))\cdot \prod_{z=1}^{n-m} P(al_z \mid c_j,K(al_z)) = \beta\cdot\gamma\cdot P(c_j)\cdot \prod_{i=1}^{n} P(a_i \mid c_j,K(a_i)) \propto P(c_j)\cdot \prod_{i=1}^{n} P(a_i \mid c_j,K(a_i))$$

其中 $K(a_i)$ 是 A_i 的非父类结点集 $K(A_i)$ 的取值,若 $A_i \in S_1$,则 $K(A_i) \subseteq \{Ak_1,Ak_2,\cdots,Ak_m\}$;若 $A_i \subseteq S_2$,则 $K(A_i) = \{Al_1,Al_2,\cdots,Al_{n-m}\}$。

2.2 基于 x^2 统计的属性相关性度量

对于两个基本属性 A,B,分别有值 $a_i(i = 1,2,\cdots,m)$,$b_j(j = 1,2,\cdots,n)$,其频数的列表如表 1 所示。

表 1 两个属性的频度

	b_1	b_2	…	b_n	SUM
a_1	f_{11}	f_{12}	…	f_{1n}	$A_1 = \sum f_{1j}$
a_2	f_{21}	f_{22}	…	f_{2n}	$A_2 = \sum f_{2j}$
…	…	…	…	…	…
a_m	f_{m1}	f_{m2}		f_{mn}	$A_m = \sum f_{mj}$
SUM	$B_1 = \sum f_{i1}$	$B_2 = \sum f_{i2}$		$B = \sum f_{in}$	

为检验行列变量的相关性,使用 x^2 统计量

$$x^2 = \sum_{i,j}^{m,n} \frac{(f_{ij} - A_iB_j/f)^2}{A_iB_j/f}$$

其中 f_{ij} 表示 a_i,b_j 同时出现的频度,A_i 表示 a_i 出现的频度,B_j 表示 b_j 出现的频度,f 为样本容量。由 x^2 统计量,可以得到 $m * n$ 列表数据中行列变量属性相关性的度量:

$$\Psi = \begin{cases} \dfrac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{A_1A_2B_1B_2}} & m = n = 2 \\ \sqrt{x^2/f} & \text{其他} \end{cases}$$

据有关统计学理论,x^2 提供有无关联性的证据,而 Ψ 反映出关联性的强弱,Ψ 的绝对值越大,属性相关性越强,其绝对值接近于 0 时属性相关性较弱。该方法同样适用于各属性变量与类变量相关性的度量[4,5]。

2.3 强属性的选择方法

令 S_1 和 S_2 是条件属性集 $\{A_1,A_2,\cdots,A_n\}$ 的一个划分,S_1 是强属性集,S_2 是弱属性集。按照上文介绍的属性相关性的分析方法,分别计算出每个条件属性与其他条件属性的相关性度量,然后计算每个条件属性与其他条件属性相关性度量的平均值,并按平均值的大小对条件属性进行排序,平均值大的条件属性显然对其他属性具有较强影响,可以作为强属性处理。具体算法描述如下:

(1) 令强属性集 S_1 为空,弱属性集 $S_2 = \{A_1,A_2,\cdots,A_n\}$;

(2) 设最大强属性个数为 m;

(3) 计算所有属性之间的属性相关性度量 $\Psi(A_i,$

A_j)；

(4) 计算每个属性的属性相关性度量的平均值 $E\Psi(A_i)$；

(5) 根据平均值 $E\Psi(A_i)$ 将所有属性降序排列 DescendSorted($E\Psi(A_t)$)；

(6) 根据设定的 m 值，将强属性放入强属性集 S_1，并从弱属性集 S_2 中删除。

算法复杂度分析：

算法的第一部分是计算所有条件属性之间的相关度，每次循环的复杂度是 $O((\text{AttriSet}-1)^2)$，算法的第二部分排序的复杂度为 $O(n\log_2 n)$。

2.4　基于强属性的朴素贝叶斯分类模型 SANBC

构造 SANBC 模型的关键是确定强属性集 $S_1=\{Ak_1,Ak_2,\cdots,Ak_m\}$，$S_1$ 中的任意两个属性可以有依赖关系，给定 S_1 和 C，S_2 中任意两个属性是条件独立的，S_1 中的属性可以 S_2 中每个属性为非类父结点，因此可以在属性之间添加增强弧以弱化朴素贝叶斯的独立性假设，从而构造出 SANBC 模型。

对于给定某一实例 $\{a_1,a_2,\cdots,a_n\}$，计算 $V_{\text{SANBC}}=\arg\max P(c_j)\prod_{i=1}^{n}P(a_i\mid c_j,K(a_i))$ 的最大类 c_j 作为该实例的类标签。

下面以强属性个数为 1 时，对 $P(a_i\mid c_j,K(a_i))$ 算法进行介绍：

假定有 m 个类 $\{c_1,c_2,\cdots,c_m\}$，强属性为 Ak，其属性取值为 $\{ak_1,ak_2,\cdots,ak_j\}$，SANBC 算法的先验条件概率分别为：

$P(C=c_1,Ak=ak_1)\quad P(C=c_1,Ak=ak_2)\cdots P(C=c_1,Ak=ak_j)$

$P(C=c_2,Ak=ak_1)\quad P(C=c_2,Ak=ak_2)\cdots P(C=c_2,Ak=ak_j)$

……

$P(C=c_m,Ak=ak_1)\quad P(C=c_m,Ak=ak_2)\cdots P(C=c_m,Ak=ak_j)$

$P(a_i\mid c_j,Ak)$ 的算法为：

$P(a_i\mid c_1,ak_1)=P(x=a_i,C=c_1,Ak=ak_1)/P(C=c_1,Ak=ak_1)$

$P(a_i\mid c_1,ak_2)=P(x=a_i,C=c_1,Ak=ak_2)/P(C=c_1,Ak=ak_2)$

……

$P(a_i\mid c_1,ak_j)=P(x=a_i,C=c_1,Ak=ak_j)/P(C=c_1,Ak=ak_j)$

$P(a_i\mid c_2,ak_1)=P(x=a_i,C=c_2,Ak=ak_1)/P(C=c_2,Ak=ak_1)$

$P(a_i\mid c_2,ak_2)=P(x=a_i,C=c_2,Ak=ak_2)/P(C=c_2,Ak=ak_2)$

$P(a_i\mid c_2,ak_j)=P(x=a_i,C=c_2,Ak=ak_j)/P(C=c_2,Ak=ak_j)$

……

$P(a_i\mid c_m,ak_1)=P(x=a_i,C=c_m,Ak=ak_1)/P(C=c_m,Ak=ak_1)$

$P(a_i\mid c_m,ak_2)=P(x=a_i,C=c_m,Ak=ak_2)/P(C=c_m,Ak=ak_2)$

……

$P(a_i\mid c_m,ak_j)=P(x=a_i,C=c_m,Ak=ak_j)/P(C=c_m,Ak=ak_j)$

通过比较，SANBC 算法在引入强属性后，其条件概率受到一定的限定，从而限定了条件概率的计算范围，尽管该算法的复杂度较朴素贝叶斯分类模型有所增加，但分类的正确率有很大提高。

3　实验结果及分析

3.1　实验结果

本实验所用的数据来自 UCI 机器学习数据库，从中选择 3 个数据集，分别为：Vote，Tic－Tac－Toe，Postoperative－Patient。首先使用 SANBC 分类模型对 3 个实验数据集进行分类正确率测试，然后使用 Weka[6] 系统中朴素贝叶斯分类工具对 3 个实验数据集进行分类正确率测试，分别得到两种算法分类的正确率，最后将得到的两种实验结果进行比较，实验结果如表 2 所示（m 为强属性个数）：

表 2　两种算法分类正确率比较

数据集	实例数	属性个数	分类正确率		
			NBC	SANBC	
				$m=1$	$m=2$
Vote	435	16	75.6%	92.8358%	95.5223%
Tic－Tac－Toe	958	9	69.8015%	77.9519%	81.1912%
Postoperative－patient	90	8	74.7126%	75.862%	83.908%

3.2　实验结果分析

实验的主要目的是对 SANBC 分类模型与 NBC 分类模型在 3 个数据集上的分类正确率进行比较，每个分类正确率是在测试集上成功预测的实例占总实例的百分比。实验结果明显表明，SANBC 分类模型在每个实验数据集上均取得了较好的分类性能，其分类性能明显优于 NBC 分类模型。从实验的过程分析，当强属性的个数为 1 时，前两个数据集分类的正确率有较大幅度提高；当强属性个数为 2 时，第三个数据集分类的正确率有较大幅度提高；继续增加强属性的个数，分类

（下转第 211 页）

```
<param name="Rate" value="1">
<param name="SAMILang" value>
<param name="SAMIStyle" value>
<param name="SAMIFileName" value>
<param name="SelectionStart" value="-1">
<param name="SelectionEnd" value="-1">
<param name="SendOpenStateChangeEvents" value="-1"
>
<param name="SendWarningEvents" value="-1">
<param name="SendErrorEvents" value="-1">
<param name="SendKeyboardEvents" value="0">
<param name="SendMouseClickEvents" value="0">
<param name="SendMouseMoveEvents" value="0">
<param name="SendPlayStateChangeEvents" value="-1"
>
<param name="ShowCaptioning" value="0">
<param name="ShowControls" value="-1">
<param name="ShowAudioControls" value="-1">
<param name="ShowDisplay" value="0">
<param name="ShowGotoBar" value="0">
<param name="ShowPositionControls" value="0">
<param name="ShowStatusBar" value="-1">
<param name="ShowTracker" value="0">
<param name="TransparentAtStart" value="0">
<param name="VideoBorderWidth" value="0">
<param name="VideoBorderColor" value="0">
<param name="VideoBorder3D" value="0">
<param name="Volume" value="-130">
<param name="WindowlessVideo" value="0">
```

具体的参数控制含义见参数 name 中的英文含义。控制参数个数用户可根据自己实际需要任意组合。关键是 param name="Filename"参数项其值一定要设置成动态,如 value="Dataset1.spwj"它对应的是数据库中视频文件表的文件路径字段[6~8]。

第三步 编写相关控制程序。

在播放页面的加载事件中编写程序根据数据库中视频课件的文件格式决定播放控件的注册号。如果是文件扩展名为.rm 调用 Realplayer 控件,其注册号为 clsid:CFCDAA03-8BE4-11cf-B84B-0020AFBBCCFA。如果文件扩展名:.wma,.wme,.wms,.wmv,.wmx,.wmz 或 .wvx 调用 Windows Media 控件,其注册号为 clsid:22D6F312-B0F6-11D0-94AB-0080C74C7E95。

4 结束语

根据用户的实际情况可对该系统增加视频课件查询和搜索等功能,另外,本系统也可直接用于各种影视点播系统中。本系统作为安徽省教育厅自然科学基金资助项目基于校园网的课件制作与发布系统的一个子系统在安徽财经大学校园网上测试通过并且运行良好。

参考文献:

[1] Tanenbaum A S. 计算机网络[M]. 熊桂喜,王小虎,译. 北京:清华大学出版社,1998.

[2] 周森鑫. 基于校园网的学生成绩管理系统[J]. 计算机技术与发展,2006,16(2):35-37.

[3] 谢希仁. 计算机网络[M]. 第 3 版. 大连:大连理工大学出版社,2000.

[4] 启明工作室. ASP.NET + SQL Server 网络应用系统开发与实例[M]. 北京:人民邮电出版社,2005.

[5] Siyan K. Windows2000 TCP/IP 实用全书[M]. 张　锦,彭宗仁,等译. 北京:电子工业出版社,2001.

[6] 曾清国. Windows2000 + ASP + SQL Server 案例教程[M]. 北京:中科多媒体电子出版社,2001.

[7] 吕弘文. Dreamwaver Mx 2004 与 ASP.NET 动态网页设计[M]. 北京:机械工业出版社,2006.

[8] 王恩波. 网络数据库实用教程——SQL Server2000[M]. 北京:高等教育出版社,2004.

(上接第 207 页)

的正确率也许可能进一步提高。但对于不同的数据集,最佳的强属性的个数可能会有所不同,强属性的个数多少是最佳的,这还有待于进一步探讨。

参考文献:

[1] Lu R Q. Artificial Intelligence[M]. Beijing: Science Press, 1989:1134-1147.

[2] Zheng Z, Webb G I. Lazy learning of Bayesian rules[J]. Machine Learning, 2000,41(1):53-84.

[3] 石洪波. 一种限定性的双层贝叶斯分类模型[J]. 软件学报,2004,43(2):194-196.

[4] 王大玲. 一种基于关联性度量的决策树分类方法[J]. 东北大学学报,2001,22(5):482-483.

[5] 韩家新. 一种以相关性确定条件属性的决策树[J]. 微机发展,2003,13(5):38-39.

[6] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations [M]. Seattle: Morgan Kaufmann Publishers, 2000:265-314.