朴素贝叶斯分类器的独立性假设研究

范金金[1],刘 鹏[2]
FAN Jin-jin[1],LIU Peng[2]

1.上海财经大学 信息管理与工程学院,上海 200433
2.上海财经大学 人事处,上海 200433
1.School of Information Management and Engineering,Shanghai University of Finance and Economics,Shanghai 200433,China
2.Department of Human Resources,Shanghai University of Finance and Economics,Shanghai 200433,China
E-mail:fanjinjin2004@163.com

FAN Jin-jin,LIU Peng.Research on Naïve Bayesian Classifier's independence assumption.Computer Engineering and Applications,2008,44(34):139-141.

Abstract: Naïve Bayesian Classifier(NBC) is a simple and effective classification model.In this paper,after the introduction of the basic principle of the NBC model,the independence assumption of the model is analyzed.Based on the summary of the current research of the independence assumption,It is concluded that the satisfaction of the independence assumption is not a necessary condition for the NBC model's efficiency,through the demonstration from an example and some experimental analysis.
Key words: data mining;Naïve Bayesian Classifier(NBC) model;independence assumption

摘 要:朴素贝叶斯分类器(NBC)是一种简洁而有效的分类模型。介绍了NBC模型的基本原理,并着重分析了该模型的独立性假设条件。在总结现有独立性假设研究的基础上,通过例子和实验分析得出结论:NBC模型的表现和独立性假设是否满足没有必然联系。
关键词:数据挖掘;朴素贝叶斯分类器;独立性假设
DOI:10.3778/j.issn.1002-8331.2008.34.043 **文章编号**:1002-8331(2008)34-0139-03 **文献标识码**:A **中图分类号**:TP181

1 引言

分类是将一个未知样本分到几个预先已知类的过程。在数据挖掘的研究和应用中,分类问题一直受到很大地关注。在众多的分类模型中,应用最为广泛的两种分类模型是决策树和朴素贝叶斯分类器(Naïve Bayesian Classifier,NBC,以下称之为NBC模型)。

NBC模型是从贝叶斯理论发展而来的, 贝叶斯理论中最核心的部分是贝叶斯公式。假设 M 维样本变量 $X=(X_1,X_2,\cdots,X_M)$,x 为 X 的一个样本,类标签为 $t(t=1,2,\cdots,T)$。贝叶斯公式可以表示如下:

$$P(t|X=x)=\frac{P(t)*P(X=x|t)}{P(X=x)} \tag{1}$$

这里,$P(X=x)$对于所有的类来说都一样,而 $P(X=x|t)$和 $P(t)$可以由训练数据集得出。所以,对于每个样本 x 来说,不需要计算 $P(t|X=x)$的精确值,只需要求出使 $P(t)*P(X=x|t)$值最大的那个类 t,就可以预测出该样本 x 所在的类了。

然而,计算 $P(X=x|t)$是有难度的,特别是在数据集很小的时候。有鉴于此,学者们提出了NBC模型,它最重要的假设是:给定类标签 t,样本的各个属性之间是相互独立的。于是就有:

$$P(X=x|t)=\prod_{m=1}^{M}P(X_m=x_m|t) \tag{2}$$

现在贝叶斯公式变为:

$$P(t|X=x)=\frac{P(t)*\prod_{m=1}^{M}P(X_m=x_m|t)}{P(X=x)} \tag{3}$$

这就是NBC模型,它是目前公认的一个简单而有效的分类器,其性能可与决策树,神经网络等分类器相竞争[1]。

在现实中,NBC模型的独立性假设很少满足,因为数据集的属性是很少相互独立的。然而,在一些属性之间存在明显关联关系的数据集中,NBC模型也有优异表现。本文通过例子和实验分析得出结论:NBC模型的表现和独立性假设是否满足没有必然联系。

2 现有的独立性假设研究

Hand和Yu的研究指出:在分类场合,人们所关心的是若干概率的排序结果,而不是其具体的值。最佳的分类效果是:只要 $P(t_1|X=x)>P(t_2|X=x)$,就可以提供一个 $P'(t_1|X=x)>P'(t_2|X=x)$,这里 P 是实际的概率,而 P' 是 P 的估计值。在大多数情况下,NBC模型可以满足这个要求。虽然样本属性之间存在的关联关系使得 $P'(t|X=x)$和 $P(t|X=x)$之间误差很大,但是所有类的估计值都比真实值来得小[2]。当概率的排列次序没有发生变化时,分类结果就不会受到影响。

作者简介:范金金(1984-),男,硕士研究生,主要研究领域为数据挖掘;刘鹏(1962-),男,博士,教授,主要研究领域为数据挖掘。
收稿日期:2008-05-07 **修回日期**:2008-08-11

Pedro Domingos 和 Michael Pazzni 在 28 个数据集上进行了一次实证研究,它们衡量了数据集中各个属性之间的关联程度,同时也比较了 NBC 模型和其他模型的表现。它们得出了一个重要的观测结果:NBC 模型在许多属性实质上是相互关联的数据集中,比其他更为复杂的模型有更好的表现,所以它表现较好的原因并不是属性之间的独立关系[3]。

Irina Rish 的实证研究得出了一个“令人惊奇”的结论,NBC 模型的预测准确率不是由属性之间的关联程度确定的。Irina Rish 由此联想到:当假设属性之间是条件独立的时候,原本能够使得 NBC 模型取得更优表现的信息已经丢失了[1]。

由于人们认为 NBC 模型的主要缺陷是其不现实的独立性假设,所以目前的许多改进策略都是从这方面入手,即放松或改进 NBC 模型的独立性假设或者调整算法来适应独立性假设的需要,从而提高它的分类准确性。Pedro Domingos 和 Michael Pazzni 对这一类研究作出了总结性论述,他们发现这些改进策略对 NBC 模型的提高程度都相当有限[3]。这也从反面说明了 NBC 模型的表现和独立性假设是否满足没有必然联系。

除此之外,Pedro Domingos 和 Michael Pazzni 举了一个简单例子,在近乎苛刻的假设条件下得出结论:使属性满足独立性假设并不能优化 NBC 模型。本文对这个例子进行了适当的扩展,使上述结论更具普遍性。

3 一个简单例子

考虑这样一个例子,数据集有 3 个属性 A、B 和 C,假设 A 和 C 相互独立,B 和 A 完全一样,类标签的值有两个+和−。值得说明的是:在 Pedro Domingos 和 Michael Pazzni 的例子中,$P(+)$和 $P(-)$是等概率的,即 $P(+)=P(-)=0.5$。而本文没有这个约束条件,以使结论更具普遍性。

按照以往对 NBC 模型的理解和处理,由于 A 和 C 相互独立,B 和 A 完全一样,B 应该被忽略以建立一个只包含属性 A 和 C 的独立 NBC 模型。对独立 NBC 模型,应该被归为+的样本的最佳分类条件是:

$$P(+)*P(A|+)*P(C|+)>P(-)*P(A|-)*P(C|-) \tag{4}$$

然而,即使 B 和 A 完全一样,NBC 模型也会假设属性 A、B 和 C 条件独立,然后使用 3 个属性建立模型,这等于把属性 A 计算两次。因此,对 NBC 模型,应该被归为+的样本的最佳分类条件是:

$$P(+)*P(A|+)^2*P(C|+)>P(-)*P(A|-)^2*P(C|-) \tag{5}$$

由贝叶斯公式可以得出:$P(A|+)=P(A)*P(+|A)/P(+)$,其余的概率也作类似转换。于是,独立 NBC 模型的最佳分类条件就等价于:

$$\frac{P(+|A)*P(+|C)}{P(+)}>\frac{P(-|A)*P(-|C)}{P(-)} \tag{6}$$

而对于 NBC 模型,其最佳分类条件等价于:

$$\frac{P(+|A)^2*P(+|C)}{P(+)}>\frac{P(-|A)^2*P(-|C)}{P(-)} \tag{7}$$

为方便起见,设 $P(+|A)$ 为 p,$P(+|C)$ 为 q,$P(+)$ 为 r,p、q 和 r 均属于区间(0,1)。于是式(6)和式(7)分别转换为式(8)和式(9):

$$\frac{p*q}{r}>\frac{(1-p)*(1-q)}{1-r} \tag{8}$$

$$\frac{p^2*q}{r}>\frac{(1-p)^2*(1-q)}{1-r} \tag{9}$$

把式(8)和式(9)都看作是以 p 为自变量,q 为因变量及 r 为常量的一元函数不等式,于是式(8)和式(9)分别转换为式(10)和式(11):

$$q>\frac{(1-p)*r}{(1-p)*r+(1-r)*p} \tag{10}$$

$$q>\frac{(1-p)^2*r^2}{(1-p)^2*r^2+(1-r)^2*p^2} \tag{11}$$

如果某点(p,q)满足不等式(10),那么这一点代表独立 NBC 模型的一个最佳分类条件,所有这类点将组成一个最佳分类条件域,而该区域的大小就代表了此时独立 NBC 模型的分类能力。独立 NBC 模型的最佳分类条件域的面积大小是:

$$S_1=1-\int_0^1\frac{(1-p)*r}{(1-p)*r+(1-r)*p}dp=1-\frac{r}{(2*r-1)}-\frac{r*(1-r)}{(2*r-1)^2}*\ln(\frac{1-r}{r}) \tag{12}$$

特殊地,当 $r=0.5$ 时,$S_1=0.5$。

类似地,NBC 模型的最佳分类条件域的面积大小是:

$$S_1=1-\int_0^1\frac{(1-p)^2*r^2}{(1-p)^2*r^2+(1-r)^2*p^2}dp=1-(\frac{r^2}{2*r^2-2*r+1})-\frac{2*(1-r)^2*r^2}{(2*r^2-2*r+1)^2}*\ln(\frac{1-r}{r})+\frac{(1-r)*r*(2*r-1)}{(2*r^2-2*r+1)^2}*\frac{\pi}{2} \tag{13}$$

最后比较 S_1 和 S_2 的大小,从式(12)和式(13)可以看出,S_1 和 S_2 都和 r 相关,S_2-S_1 也和 r 相关。S_2-S_1 和 r 之间的数量关系如图 1 所示。

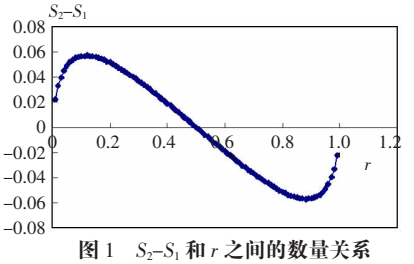

图 1 S_2-S_1 和 r 之间的数量关系

在图1 中,函数图像是关于(0.5,0)点对称的。由于 r 服从(0,1)的均匀分布,所以 S_1 和 S_2 在整体上是一样大小的。换句话说,独立 NBC 模型和 NBC 模型的最佳分类条件域在整体上是一样大小的,删除属性 B 并没有从实质上优化 NBC 模型。

这个简单例子在小范围内严格地证明:提高属性之间的独立性对改进 NBC 模型没有明显效果。为了验证这个结论在更大数据集上是否成立,将在下章中作出实验分析。

4 实验分析

4.1 数据集

从数据挖掘者常用的 UCI 数据集[4]中抽取了 7 个数据集作实验分析。7 个数据集的全部属性均为分类型。从原始数据集中随机抽取样本数的 2/3 作为训练数据集,把余下的 1/3 作为测试数据集。7 个数据集的具体信息见表 1。

4.2 属性相关性度量

张静[5]等人提出了一种简单有效的离散属性相关性度量。在此文中,两个属性的相关度被度量如下:

表 1　7 个数据集的介绍

No.	数据集名称	记录数	属性数（不含类标签）	分类数
1	breast-cancer	277	9	2
2	credita	671	9	2
3	cylinderband	352	14	2
4	primary-tumor	270	16	22
5	vote	232	16	2
6	soybean	562	35	19
7	german	1 000	13	2

$$R(A,B)=\frac{Card(Pos_A(B)\cup Pos_B(A))}{Card(U)} \tag{14}$$

U 是全体样本的总集。$Pos_A(B)$也是一些样本的集合，在这些样本上，当 A 取某个确定的属性值的时候，B 的取值不变，$Pos_B(A)$的含义类似。$Card$ 是集合中元素的个数。$R(A,B)$关于属性 A 和 B 对称，$R(A,B)\in[0,1]$。$R(A,B)$的值越大，属性 A 和 B 的相关性就越强。

4.3　实验设计

实验设计如下：

（1）计算每个属性和数据集中其他所有属性的关联程度，定义为 $R(A_i)$：

$$R(A_i)=\frac{\sum_{j=1}^{M}R(A_i,A_j)}{M-1},(j\neq i) \tag{15}$$

而后计算数据集属性之间的平均关联程度，定义为 R：

$$R=\frac{1}{M}\sum_{i=1}^{M}R(A_i) \tag{16}$$

R 值的大小反映了数据集中属性之间的独立性，R 值越小，独立性越大。

（2）删除 M 个属性中 $R(A_i)$值最大的属性，这是因为具有最大 $R(A_i)$值的属性和其他属性的关联度最高，删除该属性有助于提高剩余属性之间的独立性。然后使用剩余属性建立 NBC 模型，记录此时的预测准确率。

（3）重复以上两个步骤，直到数据集中只剩下一个属性为止，此时数据集的 R 值当然为 0。

4.4　实验结果

7 个数据集的实验结果都说明了这样一个事实：当具有最大 $R(A_i)$值的属性被依次删除之后，数据集的 R 值逐渐变小直至为 0，这说明剩余属性之间的关联度在逐渐下降，但是使用剩余属性建立的 NBC 模型的预测准确率却很少能够超越使用全部属性建立的原始 NBC 模型。所以本实验证实，删除与其他属性有较强关联的属性后，数据集中的属性的独立性得到增强，但是 NBC 模型却没有被改进。本文从 7 个数据集的实验结果中任意选取了两个结果，分别由图 2 和图 3 来表示。

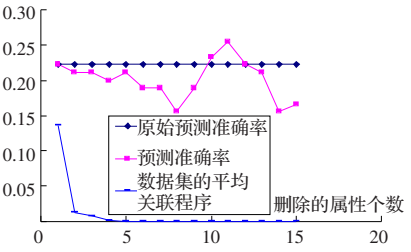

图 2　数据集 primary-tumor 的实验结果

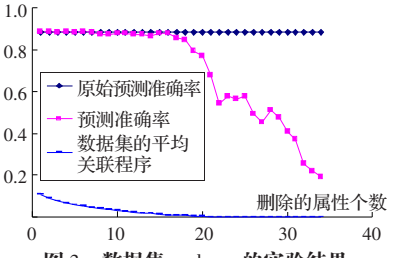

图 3　数据集 soybean 的实验结果

5　结束语

本文介绍了 NBC 模型的基本原理，并着重分析了该模型的独立性假设条件。在总结现有独立性假设研究的基础上，本文通过例子和实验分析得出结论：NBC 模型的表现和独立性假设是否满足没有必然联系，提高属性之间的独立性对改进 NBC 模型没有明显效果，这对今后从理论上探索如何改进 NBC 模型将起到积极作用。

参考文献：

[1] Rish I.An empirical study of the naive Bayes classifier[C]//Proceedings of IJCAI Workshop on Empirical Methods in Artificial Intelligence，2001.

[2] Hand D J，Yu K.Idiot's Bayes-not so stupid after all?[J].International Statistical Review，2001(69)：388-369.

[3] Domingos P，Pazzani M.On the optimality of the simple Bayesian classifier under zero-one loss[J].Machine Learning，1997(29)：103-130.

[4] Merz C J，Murphy P M.UCI repository of machine learning datasets[EB/OL].http://www.ics.uci.edu/mlearn/MLRepository.html.

[5] 张静，王建民，何华灿.基于属性相关性的属性约简新方法[J].计算机工程与应用，2005，41(28)：55-57.

（上接 68 页）

在 $p\geq 1$ 的实数范围内任意选取（通常取正整数），α,β 可在 $[0,1]$内选取（只要 $\alpha+\beta\leq 1$ 即可），所以这是数量庞大的一类 Vague 集间的相似度公式。而文献[4-5]提出的公式中有 3 个相似度量公式是本文公式的特例。把本文公式应用于模式识别的例子表明这类公式是实用的。

参考文献：

[1] Gau W L，Buehrer D J.Vague sets[J].IEEE Tran Systems Man and Cybemetics，1993，23(2)：610-614.

[2] 刘华文，王凤英.Vague 值的转化与相似度量[J].计算机工程与应用，2004，40(32)：79-81.

[3] 赵亚娟，王鸿绪.关于 Vague 集间相似度量的缺陷及修补[J].计算机工程与应用，2007，43(5)：49-51.

[4] 黄国顺.一类新 Vague 集相似度量[J].计算机应用与软件，2005，7：24-26.

[5] 黄国顺.Vague 集相似度量及其在模式识别中的应用[J].复旦学报：自然科学版，2004，43(5)：869-872.

[6] Liang Z Z，Shi P F.Similarity measures on intuitionisistic fuzzy sets[J].Pattern Recognition Letters，2003，24：2687-2693.