第16卷 第5期
2006年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 5
May 2006

一种扩展的朴素贝叶斯分类器改进算法

张丽霞,赵大宇
(沈阳师范大学 数学与系统科学学院,辽宁 沈阳 110034)

摘 要:文中研究贝叶斯分类器家族中的一种扩展朴素贝叶斯分类器。此种扩展朴素贝叶斯分类器满足两个条件:一是类结点是所有属性的父结点;二是每个属性最多有一个属性父结点。其中有代表性的两种算法是贪婪爬山算法(Hill Climbing Search,即HCS算法)和超父结点算法(Superparent,即SP算法)。对两种算法进行了分析和比较,并在此基础上提出了一种改进算法。通过实验验证所改进的分类器是正确的、有效的。

关键词:朴素贝叶斯分类器;贪婪爬山算法;超父结点算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2006)05-0028-03

An Improved Algorithm for Learning Augmented Naive Bayes Classifier

ZHANG Li-xia, ZHAO Da-yu
(School of Mathematics and Systematic Science, Shenyang Normal University, Shenyang 110034, China)

Abstract: An augmented naive Bayes classifier of Bayes classifier family is studied in this paper. This classifier is defined by the following two conditions: one is that each attribute has the class attribute as parent; the other is that each attribute may have one other attribute as parent. Both representative algorithms are Hill Climbing Search and Superparent . Analysis and comparison are done to the above two algorithms, proposing an improved algorithm at the same time. It is sure that the modified algorithm is effective and correct during the demonstration.

Key words: NBC; HCS; SP

0 引 言

贝叶斯分类器[1~4]指的是基于贝叶斯学习方法的分类器。朴素贝叶斯分类器(NBC)[1~3]和扩展朴素贝叶斯分类器[1~4]是两种简单的贝叶斯分类器。NBC分类器在理论上满足其限定条件下是最优的,针对其较强的限定条件(即在类结点条件下属性结点之间是条件独立的),可以尝试着减弱它以扩大最优范围。于是扩展朴素贝叶斯分类器产生了:其中最简单的一种为树型扩展朴素贝叶斯分类器(TAN,Tree Augmented Naive Bayes)[1~3]。该分类器的类变量作为所有属性的父结点,而属性之间构成一个树形结构;当属性结点间关系不仅局限于树形,而是任意的贝叶斯网,这样的分类器称为BAN(Bayesian Network Augmented Naive Bayes)[3]。另外一种简单的扩展朴素贝叶斯分类器是属性之间构成一个有向森林[4],即类结点是所有属性的父结点,每个属性最多有一个非类的属性父结点(TAN是此种分类器的一种特殊情况)。以下提到的扩展朴素贝叶斯分类器都指此种分类器。构造该分类器有两种算法即HCS算法和SP算法。属性之间的条件依赖关系不是很复杂的情况下,HCS和SP算法的分类准确性非常高,通常好于TAN,在数据集规模不太大的情况下,优于BAN。文中分析和比较了HCS和SP算法,并在此基础上提出一个改进算法,提高了分类器的性能。

1 贝叶斯分类模型

贝叶斯分类模型是一种典型的基于统计方法的分类模型。贝叶斯定理是贝叶斯理论中最重要的一个公式,是贝叶斯学习方法的理论基础,它将事件的先验概率与后验概率巧妙地联系起来,利用先验信息和样本数据信息确定事件的后验概率。

令 $U=\{A_1,A_2,\cdots,A_n,C\}$ 是离散随机变量的有限集,其中 $A_1,A_2,\cdots,A_n$ 是属性变量,类变量 C 的取值范围为 $\{c_1,c_2,\cdots,c_l\}$,a_i 是属性 A_i 的取值。实例 $x_i=(a_1,a_2,\cdots,a_n)$(x 表示矢量)属于类 c_j 的概率,可由贝叶斯定理表示为:

$$P(c_j\mid a_1,a_2,\cdots,a_n)=\frac{P(a_1,a_2,\cdots,a_n\mid c_j)P(c_j)}{P(a_1,a_2,\cdots,a_n)}$$
$$=\alpha\cdot P(c_j)\cdot P(a_1,a_2,\cdots,a_n\mid c_j)$$

收稿日期:2005-09-01
基金项目:辽宁省高等学校科学研究项目(202112020)
作者简介:张丽霞(1979-),女,辽宁铁岭人,硕士研究生,研究方向为神经网络、数据挖掘;赵大宇,教授,硕士研究生导师,研究方向为数据挖掘、物流。

(1)

其中 α 是正则化因子。依据概率的链规则,式(1)可以表示为:

$$P(c_j \mid a_1, a_2, \cdots, a_n) = \alpha \cdot P(c_j) \cdot \prod_{i=1}^{n} P(a_i \mid a_1, a_2, \cdots, a_{i-1}, c_j) \quad (2)$$

根据贝叶斯最大后验准则,给定某一实例 $x_i = (a_1, a_2, \cdots, a_n)$,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 C^* 作为该实例的类标签。因此,贝叶斯分类模型的关键是如何计算 $P(a_i \mid a_1, a_2, \cdots, a_{i-1}, c_j)$。

目前,不同贝叶斯分类模型的区别就在于以不同的方式来求 $P(a_i \mid a_1, a_2, \cdots, a_{i-1}, c_j)$。在 NBC 分类器中,每个结点只与类结点相关联,因此,式(2)中的 $P(a_i \mid a_1, a_2, \cdots, a_{i-1}, c_j)$ 简化为 $P(a_i \mid c_j)$。

文中所讨论的简单扩展朴素贝叶斯分类器所构成的有向森林的结构中(包括 TAN 分类器),类变量是根,没有父结点,是每个属性的父结点。属性 A_i 除了类变量作为其父结点以外,最多有一个其他属性结点作为其父结点。因此,式(2)中的 $P(a_i \mid a_1, a_2, \cdots, a_{i-1}, c_j)$,或者简化为 $P(a_i \mid c_j)$,或者简化为 $P(a_i \mid a_p, c_j)$,其中 $a_p \in \{a_1, a_2, \cdots, a_{i-1}\}$。Friedman[2]等人提出了利用条件互信息构造 TAN 分类器的算法;Pazzani 和 Keogh[4,5]采用不同的思路构造扩展朴素贝叶斯分类器,选择使分类精度改进最大的弧作为扩展朴素贝叶斯分类器的增强弧。这两种方法的不同之处在于,采用不同的评价准则来选择增强弧。

2　扩展朴素贝叶斯分类器算法的讨论

扩展朴素贝叶斯分类器是在 NBC 基础上放松属性之间条件独立的假设,使属性之间存在一定的简单依赖关系,即在属性之间增添连接弧,这些弧称为扩展弧。从结点 A_i 到 A_j 的扩展弧表示属性 A_j 对分类的影响也取决于 A_i 的值。可能有这种情况:待分类事例的属性值 a_j 和 a_i 对分类的影响都不大,即 $P(a_j \mid c)$ 和 $P(a_i \mid c)$ 的值低,但 $P(a_j \mid c, a_i)$ 的值却高,这时 NBC 降低了事例属于类 c 的概率,而扩展的朴素贝叶斯分类器可以避免这一点。当属性之间的依赖关系不太复杂的时候,可以进行有效的简单的参数学习来建立网络。

2.1　HCS 算法

由于每个属性结点最多有一个父结点(非类结点),所以最多可增加 $N-1$ 条相关联的弧,这里 N 是属性结点个数;最少可增加 0 条弧(即 NBC)。扩展朴素贝叶斯分类器是在 NBC 分类器基础上的一种扩展的贝叶斯网络,为了更好地理解这个网络结构,先给出一个定义。

定义 1　在扩展贝叶斯网络中,除了类结点以外没有父结点的属性结点称为孤点。

爬山搜索算法(HCS):

①初始化网络为一个 NBC 网络。

②评估当前的分类器。

③在当前分类器中添加每条合法的弧。

④假如存在能够提高分类器性能的弧,那么就选择使分类器性能改进最大的弧,添加到当前分类器中,转向②;否则返回当前分类器。

初始化网络为一个 NBC 网,评估当前的分类器,孤点集 O 最初是属性点集 $A_1, A_2, \cdots, A_n$。评估从 A_i 到 A_j 的每条合法弧($A_i \neq A_j, A_j \in O$),利用留一交叉验证 LOO(Leaving One Out Cross Validation)。如果没有弧能提高分类器性能时,当前的分类器是最优的,否则保留对分类性能提高最大的那条弧,这条弧指向的那个属性结点从 O 中移出。当 O 只包含一个结点或是没有提高分类器性能的弧时,循环终止。

2.2　SP 算法

再介绍一种更有效率的,时间复杂性不太高,同样能提高分类器性能的算法——SP 算法。同 HCS 类似,SP 也是寻找最好的弧来有效地提高分类器的精度,把这个过程分为两个阶段:第一阶段先找到一个最优的超父结点;第二阶段找到这个超父结点对应的最优子结点。超父结点和最优子结点的定义见文献[4]。

超父结点算法(SP):

①初始化网络为一个 NBC 网络。

②评估当前的分类器。

③考虑每个结点为超父结点,设能提高分类器性能最好的那个结点为超父结点 A_{SP}。

④考虑从 A_{SP} 到每个孤点的弧,假如有提高分类器性能的弧,则把提高最大的那个弧加入到当前分类器中,并转向③;否则,返回当前的分类器。

SP 第三步是依次设每个属性结点为超父结点,评估对分类器性能的影响程度,把提高分类器性能最大的那个结点称为超父结点 A_{SP},其次是寻找 A_{SP} 的最优子结点,在保证属性节点中没有环生成的条件下,连接 A_{SP} 到每个孤点的弧,评估哪个弧的加入使分类器性能提高最大,提高最大的那个弧所对应的孤点即是最优子结点 A_{ch},然后把 A_{ch} 从 O 中移出,A_{SP} 到 A_{ch} 的弧加到当前分类器,循环依次下去,当没有提高分类器精度的弧或是孤点个数等于一时,循环终止。

3　算法的改进

文献[4]中的实验共有 14 个数据集,比较分类器 NBC,TAN,HCS 和 SP 的实验结果中,HCS 算法在 8 个数据库中的分类精度最好,SP 算法在 6 个数据库中的分类精度最好。于是从文献[3]得到启示,可以把 HCS 算法和 SP 算法拟和起来,构成一个每次输出的分类器都是性能最好的。

改进的算法:

①按 HCS 算法得到分类器 1,评估它的分类精度。

②按 SP 算法得到分类器 2,评估它的分类精度。

③利用评估的结果选择两个分类器中最优的。

④保留最优分类器的结构,利用整个的数据集进行参数学习(条件概率表)。

⑤输出这个新的分类器。

数据集规模不大的情况下,选择用 LOO 验证法。LOO 充分利用训练样本,是最严格、最精确的一种评估方法之一,但是它适用于小规模数据集,否则增加机器运行时间。数据集规模很大的情况下,选择保留(Holdout)方法来测试分类器的性能。保留方法中一般采用数据集的 2/3 作为训练集,数据集的 1/3 作为测试集。

利用保留方法测试小数据集的分类器性能会降低分类器的分类精度;而利用 LOO 方法测试大数据集的分类器,会增大机器的运行时间。因此不同的数据规模下选择合理的检验方法既保证分类器分类精度,又可以减少机器运行时间。

4 实验结果

标准数据集是从 UCI[5] 上下载的,选取了 8 个数据集。有关数据集的概况见表 1。所选择的数据集都是小规模的,所以本实验只针对评估方法是 LOO 情况进行。在实验中不同的数据集是在相同的环境下运行的。

表 1 算法的实验结果

Dataset	Attributes	Classes	Instances	HCS	SP	HCS&SP
Vehicle	18	4	846	70.1	70.3	70.3
Post-op	9	3	90	72.7	72.1	72.7
Australia	14	2	690	84.7	85.3	85.3
Hepatitis	19	2	270	84.8	84.3	84.8
Vote	16	2	435	95.6	95.7	95.7
Heart	13	2	270	78.7	76.1	78.7
Soybean-Large	35	19	562	88.6	88.4	88.6
Pima	8	2	768	78.0	78.2	78.2

表 1 列出了实验结果,实验结果基本和文献[4]中的数据差不多,第七列是改进算法的实验结果,每次输出的分类器确实是 HCS 和 SP 中性能最好的,验证了所建立的分类器的有效性和正确性。同时也发现当属性变量之间的依赖关系不是很复杂的情况下,分类器的分类精度会很高。例如数据集 Vote,说明文中讨论的分类器在对 Vote 操作时几乎把影响分类准确率的弧都加上了,丢失有价值的弧的个数相应减少。当属性间的关系很复杂时,由于分类器属性之间构成的是有向森林,有可能丢失一些对分类有价值的弧,如 Vehicle。再一次表明了属性结点之间的条件依赖关系不是很复杂的情况下,HCS 和 SP 算法的分类准确性非常高,改进的算法的分类准确性更高。

5 结束语

通过对 HCS 算法和 SP 算法的分析,又提出了一个改进的算法,明显提高了分类器的分类精度。扩展朴素贝叶斯分类器既有朴素贝叶斯分类器的简单性,又有很好的分类能力,应努力把它应用到更多的现实领域中。当属性结点之间的相关性很复杂的情况,即有简单性又有很好的分类性能的分类器待于进一步的研究。

参考文献:

[1] 林士敏,田凤占,陆玉昌.用于数据采掘的贝叶斯分类器研究[J].计算机科学,2000,27(10):73-76.

[2] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning,1997(29):131-163.

[3] Cheng J, Greiner R. Comparing Bayesian network classifiers [A]. Proceedings of the fifteenth conference on uncertainty in artificial intelligence[C]. San Francisco: Morgan Kaufmann, 1999.101-108.

[4] Eamonn J, Pazzani J. Learning augmented bayesian classifiers: a comparison of distribution-based and classification-based approaches[A]. Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics[C]. Lauderdale:[s.n.],1999.225-230.

[5] Blake C, Keogh E, Merz C. UCI repository of machine learning database[EB/OL]. http://www.ics.uci.edu/mlearn/ML Repository.html,1998.

(上接第 27 页)

就可以确定输出层的神经元数目为 4。但是通过其后的实践发现,当采用(0,0,0,0)这样的目标输出向量时,BP 网络无法收敛,那是因为采用的激发函数永远不可能达到 0 或 1,而只能是接近。所以还要对其重新编码,最后确定其编码为:0(0.1,0.1,0.1,0.1),1 (0.1,0.1,0.1,0.9),…,9 (0.9,0.1,0.1,0.9)。

3 实验结果

本系统采用了 100 个训练样本用来测试算法的性能,这 100 幅图像包括 Arial, Batang, Gautami 字体,单个数字的平均识别率达到 94%,网络训练时间大约 3~6 秒,当然如果采用更多的训练样本,则识别率将会进一步提高。

参考文献:

[1] 张宏林,蔡 锐. Visual C++数字图像识别技术及工程实践[M].北京:人民邮电出版社,2003.422-442.

[2] 冈萨雷斯.数字图像处理(第 2 版)[M].北京:科学出版社,2003.60-127.

[3] Hagan M T. Neural Network Design[M].北京:机械工业出版社,2002.16-64.

[4] Bruck J, Sanz J. A Study on Neural Networks[J]. Int J Intelligent System,1988(3):4-15.

[5] 徐丽娜.神经网络控制[M].哈尔滨:哈尔滨工业大学出版社,1999.7-16.

[6] 袁曾任.人工神经网络及其应用[M].北京:清华大学出版社,1999.16-29.