一种限定性的双层贝叶斯分类模型*

石洪波[1,2+], 王志海[1], 黄厚宽[1], 励晓健[1]

[1](北京交通大学 计算机与信息技术学院,北京 100044)
[2](山西财经大学 信息与管理学院,山西 太原 030006)

A Restricted Double-Level Bayesian Classification Model

SHI Hong-Bo[1,2+], WANG Zhi-Hai[1], HUANG Hou-Kuan[1], LI Xiao-Jian[1]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)
[2](School of Information and Management, Shanxi University of Finance and Economics, Taiyuan 030006, China)
+ Corresponding author: Phn: +86-10-51683602, E-mail: shb710@163.com
Received 2003-01-22; Accepted 2003-07-25

Shi HB, Wang ZH, Huang HK, Li XJ. A restricted double-level Bayesian classification model. *Journal of Software*, 2004,15(2):193~199.
http://www.jos.org.cn/1000-9825/15/193.htm

Abstract: Naive Bayes classifier is a simple and effective classification method, but its attribute independence assumption makes it unable to express the dependence among attributes, and affects its classification performance. On the basis of analyzing the classification principle of Bayesian classification model and a variant of Bayes theorem, a new classification model based on Bayes theorem, DLBAN (double-level Bayesian network augmented naive Bayes), which adds the dependence among attributes by selecting the key attributes, is proposed. DLBAN classifier is compared with Naive Bayes classifier and TAN (tree augmented naive Bayes) classifier by an experiment. Experimental results show this model has higher classification accuracy in most data sets.
Key words: naive Bayes; TAN (tree augmented naive Bayes); Bayes theorem; dependence

摘 要: 朴素贝叶斯分类模型是一种简单而有效的分类方法,但它的属性独立性假设使其无法表达属性变量间存在的依赖关系,影响了它的分类性能.通过分析贝叶斯分类模型的分类原则以及贝叶斯定理的变异形式,提出了一种基于贝叶斯定理的新的分类模型 DLBAN(double-level Bayesian network augmented naive Bayes).该模型通过选择关键属性建立属性之间的依赖关系.将该分类方法与朴素贝叶斯分类器和 TAN(tree augmented naive Bayes)分类器进行实验比较.实验结果表明,在大多数数据集上,DLBAN 分类方法具有较高的分类正确率.
关键词: 朴素贝叶斯;TAN(tree augmented naive Bayes);贝叶斯定理;依赖关系
中图法分类号: TP181 **文献标识码**: A

* Supported by the Key Science-Technology Project of the National 'Tenth Five-Year-Plan' of China under Grant No.2002BA407B (国家"十五"重点科技攻关项目)

作者简介: **石洪波**(1965－),女,山西太原人,博士生,副教授,主要研究领域为机器学习,数据挖掘;**王志海**(1963－),男,博士,副教授,主要研究领域为数据仓库,数据挖掘,机器学习;**黄厚宽**(1940－),男,教授,博士生导师,主要研究领域为数据挖掘,分布式人工智能;**励晓健**(1976－),男,硕士,主要研究领域为数据挖掘,人工智能.

分类是数据挖掘和机器学习中的一个重要研究课题.它的目标是构造一个分类器,对由属性集描述的实例指定最适合的类标签.许多分类方法和技术用于构造分类模型,例如决策树、决策表、神经网络、*k*-最近邻、贝叶斯方法以及支持向量机等.而贝叶斯方法由于具有坚实的数学理论基础以及综合先验信息和数据样本信息的能力,使其正在成为当前机器学习和数据挖掘的研究热点之一.

朴素贝叶斯分类器是目前公认的一种简单而有效的概率分类方法,其性能可与决策树、神经网络等算法相竞争,在某些领域中甚至表现出更优的性能[1,2].然而,朴素贝叶斯分类方法中的"独立性假设"在大多数现实世界中明显不成立,于是人们设想能否构造一种模型,可以放松朴素贝叶斯中不现实的独立性假设,从而提高分类器的性能.

为了改进朴素贝叶斯分类器的性能,人们提出了许多方法和技术[1,3~7].在现有的改进方法中,一个关键思路是,当放弃独立性假设以后,如何表示属性变量之间可能存在的依赖关系.Kononenko 的 semi-naive[3]贝叶斯分类器将属性集分割成若干个不相交的属性组,假设在不同组中的属性之间是相互独立的,而同一属性组内的各属性相互关联.Friedman 和 Goldszmidt[1]研究了具有树结构的 TAN(tree augmented naive Bayes)分类器,它放松了朴素贝叶斯中的独立性假设条件,扩展了朴素贝叶斯的结构,允许每个属性结点最多可以依赖于 1 个非类结点.TAN 具有较好的综合性能,体现了学习效率与分类精度之间的一种适当的折衷[1,8].BAN(Bayesian network augmented naive Bayes)[1,9]进一步扩展了 TAN 的结构,允许属性之间可以形成任意的有向图,使其表示依赖关系的能力增强,然而,由于其结构的任意性,与一般贝叶斯网络一样,BAN 结构的学习是不容易的(文献[10]已证明贝叶斯网的学习是一个 NP-Complete 问题).

通过放松朴素贝叶斯的独立性假设来改进分类器性能的许多方法,往往都是在完全图中对整个弧空间进行搜索,选择最佳的弧集,如 TAN 和 BAN,而任意的弧都反映了两个属性之间的依赖关系,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会依赖于它.因此,我们希望通过对由属性结点组成的属性空间的搜索,找出一些对其他属性具有较强影响的属性,属性集中的所有其他属性仅通过与这些属性的关联就可以将属性集中重要的依赖关系表达出来.

本文从一个新的思路对贝叶斯网络的分类原则进行了讨论,认真分析了朴素贝叶斯、TAN 和 BAN 的模型结构特点以及构造这些分类器的方法,提出了基于贝叶斯定理的一种新的分类模型 DLBAN(double-level Bayesian network augmented naive Bayes),给出了构造 DLBAN 模型的算法,并实验比较了 DLBAN,TAN 和朴素贝叶斯分类器,最后总结了本文的工作以及下一步的研究方向.

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

$$\begin{aligned}P(c_j|a_1,a_2,...,a_n)&=\frac{P(a_1,a_2,...,a_n|c_j)\cdot P(c_j)}{P(a_1,a_2,...,a_n)}\\&=\alpha\cdot P(c_j)\cdot P(a_1,a_2,a_3,...,a_n|c_j)\end{aligned}\tag{1}$$

其中 α 是正则化因子, $P(c_j)$ 是类 c_j 的先验概率, $P(c_j|a_1,a_2,...,a_n)$ 是类 c_j 的后验概率,先验概率独立于训练样本数据,而后验概率反映了样本数据对类 c_j 的影响.

依据概率的链规则,式(1)可以表示为

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

给定训练数据集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_N\}$,分类任务的目标是对数据集 D 进行分析,确定一个映射函数 $f:(A_1,A_2,...,A_n)\to C$,使得对任意的未知类别的实例 $\boldsymbol{x}_i=(a_1,a_2,...,a_n)$ 可标以适当的类标签 C^*.

根据贝叶斯最大后验准则,给定某一实例 $\boldsymbol{x}_i=(a_1,a_2,...,a_n)$,贝叶斯分类器选择使后验概率 $P(c_j\,|\,a_1,a_2,...,a_n)$ 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 $P(a_i\,|\,a_1,a_2,...,a_{i-1},c_j)$.

目前,不同贝叶斯分类模型的区别就在于,它们以不同的方式来求 $P(a_i|a_1,a_2,\ldots,a_{i-1},c_j)$.在朴素贝叶斯分类器中,假定所有的属性变量都是相互类条件独立的,每个结点 A_i 只与类结点 C 相关联,因此,式(2)中的 $P(a_i|a_1,a_2,\ldots,a_{i-1},c_j)$简化为 $P(a_i|c_j)$.相对于其他分类方法,朴素贝叶斯分类器的最大特点是不需要搜索[11],只需简单地计算训练例中各个属性值发生的频率数,就可以估计出每个属性的概率估计值,因而朴素贝叶斯分类器的效率特别高.

TAN[1,8]是一种树状结构的贝叶斯模型.在 TAN 结构中,类变量是根,没有父结点,即$\Pi_C=\varnothing$(Π_C 表示 C 的父结点集),类变量是每个属性变量的父结点,即 $C\in\Pi_{A_i}$(Π_{A_i} 表示 A_i 的父结点集,i=1,2,…,n);属性变量 A_i 除了类变量 C 作为其父结点以外,最多有 1 个其他属性变量作为其父结点,即$|\Pi_{A_i}|\le 2$.因此,式(2)中的 $P(a_i|a_1,a_2,\ldots,a_{i-1},c_j)$或者简化为 $P(a_i|c_j)$,或者简化为 $P(a_i|a_p,c_j)$,其中 $a_p\in\{a_1,a_2,\ldots,a_{i-1}\}$.Friedman 等人提出了利用条件互信息构造 TAN 分类器的算法[1];Keogh 和 Pazzani 采用不同的思路构造 TAN 分类器,选择使分类精度改进最大的弧作为 TAN 的增强弧[8].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.

BAN 模型[9]原则上对每个结点的父结点个数没有限定,只需按照事先选定的评价准则,在与属性结点 A_i 相关联的结点 $A_1,A_2,\ldots,A_{i-1},C$ 中寻找 A_i 的父结点,每个结点 A_i 可以找出多个父结点,每个结点 A_i 的父结点可能不同.与一般贝叶斯网的学习方法类似,BAN 的学习有两种方式[1,12]:启发式搜索方法和相关性分析方法.

2 DLBAN 模型

2.1 贝叶斯定理变形公式

令 G_1 和 G_2 是属性集 $\{A_1,A_2,...,A_n\}$ 的一个划分,$\boldsymbol{g}_1$ 和 $\boldsymbol{g}_2$ 分别是属性集 G_1 和 G_2 的取值,实例 $(a_1,a_2,...,a_n)$(或记为 $(\boldsymbol{g}_1,\boldsymbol{g}_2)$)属于类 c_j 的概率,可由贝叶斯定理的变形公式[13,14]表示为

$$\begin{aligned}P(c_j\big|\boldsymbol{g}_1,\boldsymbol{g}_2)&=\frac{P(\boldsymbol{g}_2\,|\,c_j,\boldsymbol{g}_1)}{P(\boldsymbol{g}_2\,|\,\boldsymbol{g}_1)}\cdot P(c_j\,|\,\boldsymbol{g}_1)\\&=\beta\cdot P(\boldsymbol{g}_2\big|c_j,\boldsymbol{g}_1)\cdot P(c_j\,|\,\boldsymbol{g}_1)\end{aligned}\tag{3}$$

其中 β 是一个正则化因子.假设 $\boldsymbol{g}_1=\{a_{k_1},a_{k_2},...,a_{k_m}\}$,$\boldsymbol{g}_2=\{a_{l_1},a_{l_2},...,a_{l_{n-m}}\}$,并且在给定 c_j 和 $\boldsymbol{g}_1$ 时,$\boldsymbol{g}_2$ 中的各属性是条件独立的,那么式(3)等号右侧可以表示为

$$\beta\cdot P(c_j\,|\,a_{k_1},a_{k_2},...,a_{k_m})\cdot P(\boldsymbol{g}_2\,|\,c_j,a_{k_1},a_{k_2},...,a_{k_m})=\beta\cdot P(c_j\,|\,a_{k_1},a_{k_2},...,a_{k_m})\cdot\prod_{s=1}^{n-m}P(a_{l_s}\,|\,c_j,a_{k_1},a_{k_2},...,a_{k_m})\tag{4}$$

式(4)等号右侧假设在较少的变量(即 $A_{l_1},A_{l_2},...,A_{l_{n-m}}$)之间存在条件独立性,这个假设是比朴素贝叶斯独立性假设较弱的独立性假设,该假设的强弱取决于 $G_1=\{A_{k_1},A_{k_2},...,A_{k_m}\}$ 中属性的个数.G_1 中属性个数越多,该独立性假设越弱.由于

$$P(c_j\,|\,a_{k_1},a_{k_2},...,a_{k_m})=\gamma\cdot P(c_j)\cdot\prod_{t=1}^{m}P(a_{k_t}\,|\,c_j,a_{k_1},a_{k_2},...,a_{k_{j-1}})\tag{5}$$

其中 γ 是一个正则化因子,将式(5)代入式(4)等号右侧,得到式(6):

$$\begin{aligned}\beta\cdot\gamma\cdot P(c_j)\cdot\prod_{t=1}^{m}P(a_{k_t}\mid c_j,a_{k_1},a_{k_2},...,a_{k_{t-1}})\cdot\prod_{s=1}^{n-m}P(a_{l_s}\mid c_j,a_{k_1},a_{k_2},...,a_{k_m})\\ =\beta\cdot\gamma\cdot P(c_j)\cdot\prod_{t=1}^{m}P(a_{k_t}\mid c_j,K(a_{k_t}))\cdot\prod_{s=1}^{n-m}P(a_{l_i}\mid c_j,K(a_{l_s}))\\ =\beta\cdot\gamma\cdot P(c_j)\cdot\prod_{i=1}^{n}P(a_i\mid c_j,K(a_i))\\ \propto P(c_j)\cdot\prod_{i=1}^{n}P(a_i\mid c_j,K(a_i))\end{aligned}\tag{6}$$

其中 $K(a_i)$ 是 A_i 的非类父结点集 $K(A_i)$ 的取值.如果 $A_i\in G_1$,则 $K(A_i)\subseteq\{A_{k_1},A_{k_2},...,A_{k_m}\}$;如果 $A_i\in G_2$,则 $K(A_i)=\{A_{k_1},A_{k_2},...,A_{k_m}\}$.

因此,给定某一实例 $(a_1,a_2,...,a_n)$,应确定关键属性集 $\{A_{k_1},A_{k_2},...,A_{k_m}\}$,选择使 $P(c_j)\cdot\prod_{i=1}^{n}P(a_i\mid c_j,K(a_i))$ 最大的类 C^* 作为该实例的类标签.

2.2 DLBAN模型

定义(DLBAN 模型). 令属性集 $A=\{A_1,A_2,...,A_n\}$,类变量为 C, $G_1=\{A_{k_1},A_{k_2},...,A_{k_m}\}$ 和 $G_2=\{A_{l_1},A_{l_2},...,A_{l_{n-m}}\}$ 是属性集 A 的划分, G_1 中的任意两个属性之间都可以有依赖关系,给定 G_1 和 C, G_2 中的任意两个属性都是条件独立的,类变量 C 是 A 中每个属性的父结点, G_1 中的属性可以是 G_2 中每个属性的父结点,满足这些条件的贝叶斯分类模型称为 DLBAN 模型.

图 1 是一个简单的 DLBAN 模型.带箭头的连线表示依赖关系.

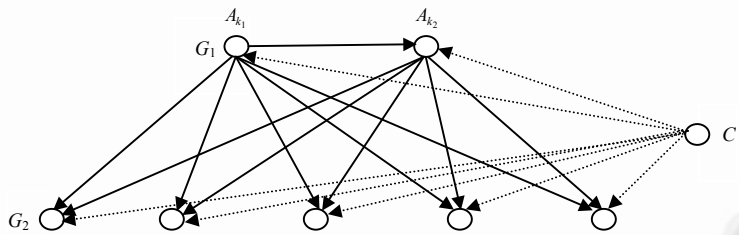

Fig.1 An example of the DLBAN model

图 1 DLBAN 模型实例

2.3 DLBAN模型结构的学习

构造 DLBAN 模型的关键是如何选择 G_1,一旦确定了属性集 G_1,在属性之间添加可以缓解朴素贝叶斯的强独立性假设的增强弧,就可以很容易地构造出 DLBAN 模型.

属性集 $\{A_1,A_2,...,A_n\}$ 中任意两个属性之间可能存在一定的依赖关系.在不同的类中,两个属性之间的依赖程度不同,而互信息可衡量两个属性相互提供信息的程度.本文用属性 A_i 与 A_j 的条件互信息表示两个属性之间的依赖关系,记为

$$I\ (A_i,A_j\mid C)=\sum_{a_i,a_j,c}P(a_i,a_j\mid c)\log\frac{P(a_i,a_j\mid c)}{P(a_i\mid c),P(a_j\mid c)}\tag{7}$$

属性 A_i 与其他属性的互信息均值为 $EI\ (A_i\mid C)=\frac{1}{n}\sum_{j=1}^{n}I(A_i,A_j\mid C)$.为了讨论方便,我们将属性集分为两个子集:强属性集和弱属性集.如果属性 A_i 的互信息均值较大,则称其为强属性;如果属性 A_i 的互信息均值较小,则称其为弱属性.强属性和弱属性是一个相对的概念,需要根据具体的数据集来判断具有某一互信息均值的属性是强属性还是弱属性.

DLBAN 模型构造包括以下几个步骤:

(1) 令强属性集 G_1 为空;弱属性集 $G_2=\{A_1,A_2,...,A_n\}$,最大强属性个数为 m,阈值 ε 取小的实数;

(2) 由训练数据集 D 统计出任意两个属性的各个属性值在每个类中出现的次数 $counts[c][a_i][a_j]$,其中 c 是类值,a_i 是属性 A_i 的属性值,a_j 是属性 A_j 的属性值;

(3) 按照公式(7)以及 $counts[c][a_i][a_j]$ 计算每一对属性之间的条件互信息 $I(A_i,A_j|C)$;

(4) 计算每个属性的互信息均值 $EI\ (A_i|C)$;

(5) 估计当前分类器的分类性能,并将估计结果保存在 OldAccuracy 中;

(6) 从属性集 G_2 中选择 $EI\ (A_i|C)$ 最大的属性 A_{k_i},将其加入到强属性集 G_1 中,并从 G_2 中去掉该属性;

(7) 对于 G_2 中的每个属性 A_j,如果 $I(A_{k_i},A_j|C)>\varepsilon$,则将 A_{k_i} 作为 A_j 的父结点;对于 G_1 中的两个属性 A_{k_i},A_{k_j},如果 $EI(A_{k_i})>EI(A_{k_j})$,则 A_{k_i} 是 A_{k_j} 的父结点,否则,A_{k_j} 是 A_{k_i} 的父结点;

(8) 估计当前分类器的分类性能,并将结果存入 *NewAccuracy*,如果 *NewAccuracy*>*OldAccuracy*,并且 G_1 中的强属性个数小于最大强属性个数 m,则 *NewAccuracy*→*OldAccuracy*,返回步骤(6);否则,结束.

2.4 算法性能分析

算法第(2)步的时间复杂度为 $O(N\cdot n^2)$,N 是实例个数,n 是属性个数,第(3)步的时间复杂度为 $O(c\cdot n^2)$,c 是类数.一般情况下,$c<<N$,所以在 G_1 中加入一个结点的时间复杂度为 $O(N\cdot n^2)$.由于 G_1 中的属性个数不能很多(G_1 中属性个数的增加将使得用于估计条件概率的训练个数减少),所以需要估计正确率的分类器通常比较少,我们的实验中设置为 3,因此,该算法的时间复杂度为 $O(N\cdot n^2)$,与 TAN 的时间复杂度一样[1,8],但常数因子比 TAN 的常数因子要低.

3 实验结果

所有实验都是在 Weka 系统[15]上完成的,实验数据选自 UCI 资源库.表 1 列出了每个数据集的实例个数、类个数、属性个数以及是否有丢失值等数据信息.由于我们的算法不能处理连续型数值数据,因此,使用 Weka 中的"weak.filiters.Discretizefilter"对连续型数值离散化,将所有包含非序数型数据的数据集离散化,使得所有的数值属性值都转换为序数型数值.在有丢失值的数据集中,将所有的丢失值作为一个单独的值来处理.

Table 1 Description of data sets used in the experiments

表 1 实验数据集的构成描述

	Domain	Size#	Classes#	Attributes#	Missing value
1	Anneal	898	6	38	Yes
2	Bupa	345	2	6	No
3	Car	1 728	4	6	No
4	Cleveland	303	2	13	No
5	Contact-Lenses	24	3	4	No
6	Flare_C	1 389	2	13	No
7	German	1 000	2	20	No
8	Horse-Colic	368	2	22	No
9	House-Votes-84	435	2	16	Yes
10	Hypothyroid	3 163	2	25	Yes
11	Iris Classification	150	3	4	No
12	King-Rook-vs-King-Pawn	3 169	2	36	No
13	LED	1 000	10	7	No
14	Mushroom	8 124	2	22	Yes
15	Nursery	12 960	5	8	No
16	Post-Operative	90	3	8	Yes
17	Promoter Gene Sequences	106	2	57	No
18	Segment	2 310	7	19	No
19	Tic-Tac-Toe End Game	958	2	9	No
20	Vehicle	846	4	18	Yes
21	Zoology	101	7	16	No

实验的主要目的是对 DLBAN 与 Naive Bayes 和 TAN 分类器在每个数据集上的分类正确率进行比较,其中 TAN 分类器采用基于条件互信息的方法[1].每个分类器的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,采用 10 重交叉验证估计分类器的正确率.

3 个分类器在每个数据集上分别测试了 20 次,每次实验采用不同的 10 重划分.表 2 列出了 20 次测试的平

均正确率及标准离差,并在最后一行列出了 3 个分类器在 21 个数据集上分类正确率的平均值,DLBAN 分类器比 TAN 分类器提高约 1%,比朴素贝叶斯分类器提高 2%.为了比较 3 种分类方案在每个数据集上的优劣,采用双尾配对 t 检验,对 DLBAN 与 Naive Bayes 和 TAN 分别进行比较,表 3 显示了在显著性水平 0.05 的情况下,DLBAN 分类器明显优于、相当于以及明显劣于 Naive Bayes 和 TAN 分类器的数据集.从表 3 中可以看出,在 21 个数据集中,DLBAN 分类器在 14 个数据集上的分类正确率明显优于 Naive Bayes 分类器,在 15 个数据集上的分类正确率明显优于 TAN 分类器.

Table 2 Experimental results by comparing three classifiers
表 2 3 种分类器的实验结果

	Domain	Naive Bayes	TAN	DLBAN
1	Anneal	96.2475±0.27	96.2363±0.27	96.4031±0.35
2	Bupa	57.0580±0.87	57.0580±0.87	57.0580±0.87
3	Car	85.5757±0.32	91.6001±0.22	94.3302±0.38
4	Cleveland	83.1571±0.69	81.4483±0.94	83.2236±0.71
5	Contact-Lenses	72.7667±3.33	65.8333±4.68	72.7667±3.33
6	Flare_C	79.0137±0.23	83.1209±0.31	83.8408±0.19
7	German	73.3053±0.52	73.0105±0.65	74.4403±0.57
8	Horse-Colic	80.2581±0.52	80.9375±0.57	82.0788±0.53
9	House-Votes-84	90.0690±0.14	93.1954±0.32	94.1724±0.34
10	Hypothyroid	98.7449±0.23	98.7241±0.18	98.8686±0.16
11	Iris Classification	93.1667±0.73	91.7500±1.47	93.4000±1.07
12	King-Rook-vs-King-Pawn	87.8989±0.12	93.4473±0.12	87.8989±0.12
13	LED	73.8874±0.34	73.9600±0.24	73.8874±0.34
14	Mushroom	95.7680±0.03	99.4090±0.03	99.6147±0.05
15	Nursery	90.2847±0.05	92.5319±0.23	95.5128±0.16
16	Post-Operative	68.8889±0.86	66.0000±1.63	68.8890±0.86
17	Promoter Gene Sequences	91.2735±1.76	82.9717±3.26	91.2735±1.76
18	Segment	91.0044±0.20	94.8068±0.28	93.3334±0.35
19	Tic-Tac-Toe End Game	69.7390±0.32	74.4394±1.17	72.8497±0.80
20	Vehicle	61.4442±0.71	70.5004±0.94	68.9230±1.46
21	Zoology	94.0325±1.04	95.5270±0.84	96.0230±0.29
	Average accuracy	82.6056	83.7397	84.7417

Table 3 Experimental results of comparing three classifiers using a two-tailed pairwise t-test
表 3 DLBAN 与 Naive Bayes 和 TAN 在双尾配对 t 检验下的实验结果

	Naive Bayes	TAN
Higher than	1,3,6,7,8,9,10,11,14,15,18,19,20,21	1,3,4,5,6,7,8,9,10,11,14,15,16,17,21
Equal to	2,4,5,12,13,16,17	2,13
Lower than	–	12,18,19,20

从实验结果可以看出,DLBAN 在大部分实验数据集上取得了最好的分类性能.对 Car,Flare_C, House-Votes-84,Mushroom,Nursery 等数据集,DLBAN 的分类正确率均比朴素贝叶斯和 TAN 分类器的分类正确率要高.但是,King-Rook-vs-King-Pawn 的情况有些特殊,最好的分类器是 TAN,并且 TAN 的分类准确率比其他分类器的分类准确率要高出许多.在本文的 DLBAN 的实验中,强属性最多选为 3 个.跟踪 King-Rook-vs-King-Pawn 和 Vehicle 的实验过程,我们发现,强属性从 1 个增加至 3 个,King-Rook-vs-King-Pawn 的分类准确率从 84.6996%增加至 87.8989%,每增加一个强属性,分类准确率平均增加 1.5%.Segment 和 Vehicle 也出现类似的情况,继续增加强属性,分类准确率也许会进一步增加.

4 结 论

朴素贝叶斯分类器是一种简单而有效的分类算法,但它的属性独立性假设使其无法表达实际数据中属性间存在的依赖关系.目前有许多种方法和技术用于改进朴素贝叶斯的性能.本文提出了一个新的贝叶斯模型 DLBAN,它通过选择某些合适的属性建立起属性之间的依赖关系,一方面扩大了每个属性可依赖的属性个数,另一方面通过属性空间的搜索来建立属性之间的依赖关系.我们的实验表明,DLBAN 分类器具有较高的分类性能.

在 DLBAN 算法的实现中,强属性的选择方法是非常重要的.本文采用的方法是按照每个属性的互信息均值来选择的.是否还有其他更好的强属性选择方法,是我们下一步研究的一个内容.另外,在实验中我们采用的强属性个数最大选择为 3.实际上,对于不同的数据集,最佳的强属性个数是不同的,并不是强属性个数越多越好,强属性个数多少才是最佳的,这也是需要进一步研究的一个问题.

References:

[1] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997,29(2-3):131~163.

[2] Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers. In: Rosenbloom P, Szolovits P, eds. Proc. of the 10th National Conf. on Artificial Intelligence. Menlo Park: AAAI Press, 1992. 223~228.

[3] Kononenko I. Seminaive Bayesian classifier. In: Kodratoff Y, ed. Proc. of the 6th European Working Session on Learning. New York: Springer-Verlag, 1991. 206~219.

[4] Pazzani MJ. Searching for dependencies in Bayesian classifiers. In: Fisher D, Lenz HJ, eds. Learning from Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag. 1996. 239~248.

[5] Langley P, Sage S. Induction of selective Bayesian classifiers. In: Mantaras RL, Poole DL, eds. Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1994. 399~406.

[6] Webb GI, Pazzani MJ. Adjusted probability naive Bayesian induction. In: Antoniou G, Slaney JK, eds. Proc. of the 11th Australian Joint Conf. on Artificial Intelligence. Berlin: Springer-Verlag, 1998. 285~295.

[7] Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In: Simoudis E, Han J, Fayyad UM, eds. Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996. 202~207.

[8] Keogh EJ, Pazzani MJ. Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. In: Heckerman DE, Whittaker J, eds. Proc. of the Uncertainty'99: The 7th Int'l Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 1999. 225~230.

[9] Cheng J, Greiner R. Comparing Bayesian network classifiers. In: Laskey KB, Prade H, eds. Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999. 101~108.

[10] Chickering DM, Geiger D, Heckerman D. Learning Bayesian networks is NP-complete. In: Fisher DH, Lenz HJ, eds. Learning from Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag, 1996. 121~130.

[11] Mitchell TM. Machine Learning. New York: McGraw-Hill Companies, Inc., 1997. 154~200.

[12] Cheng J, Bell D, Liu W. Learning belief networks from data: An information theory based approach. Artificial Intelligence, 2002, 137(1-2):43~90.

[13] Lu RQ. Artificial Intelligence. Beijing: Science Press, 1989. 1134~1147 (in Chinese).

[14] Zheng Z, Webb GI. Lazy learning of Bayesian rules. Machine Learning, 2000,41(1):53~84.

[15] Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Seattle: Morgan Kaufmann Publishers, 2000. 265~314.

附中文参考文献:

[13] 陆汝钤.人工智能.北京:科学出版社,1989.1134~1147.