doi:10.11835/j.issn.1000-582X.2014.06.011

最大相关最小冗余限定性贝叶斯网络分类器学习算法

冯月进,张凤斌
(哈尔滨理工大学 计算机科学与技术学院,哈尔滨 150080)

摘 要:朴素贝叶斯分类器(naïve bayes)是一种简单而有效的基于贝叶斯思想的分类方法,但它的属性条件独立性假设并不符合实际,影响了它的分类性能。BAN(bayesian network augmented naïve bayes)分类器扩展了朴素贝叶斯分类器,使其表示属性之间依赖关系的能力增强,但是其学习算法需要大量的高维计算,在小采样数据集上,影响BAN分类器的分类性能。基于改进的最大相关-最小冗余特征选择技术,提出限定性贝叶斯网络分类器学习算法(k-BAN)。本算法使用改进的最大相关-最小冗余特征选择技术,通过选择属性结点的连接关系集合建立属性之间的依赖性关系。将该分类方法与NB,TAN和BAN分类器进行实验比较。实验结果表明,在小采样数据集上,本算法获得的限定性贝叶斯网络分类器具有更高的分类准确性。

关键词:朴素贝叶斯;贝叶斯网络分类器;最大相关性;最小冗余性;依赖性
中图分类号:TP183 **文献标志码**:A **文章编号**:1000-582X(2014)06-071-07

Max-relevance min-redundancy restrictive BAN classifier learning algorithm

FENG Yuejin, ZHANG Fengbin
(Computer Science and Technology Institute, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: NB (Naïve Bayes) classifier is a simple and effective classification method, which is based on Bayes theorem. However, its attribute conditional independence assumption usually doesn't correspond to reality, which affects its classification performance. BAN (Bayesian network Augmented Naïve Bayes) classifier extends the ability to represent the dependence among attributes. However, BAN learning algorithms need a large amount of high dimensional computations, which impairs the classification accuracy of BAN, especially on small sample datasets. Based on the variant of max-relevance min-redundancy feature selection technology, a new restrictive BAN classifier learning algorithm (k-BAN), which builds the dependence by selecting the set of edges for each attribute node, is proposed. Compared with NB, TAN and BAN classifiers by an experiment, the restrictive BAN classifier of our algorithm has better classification accuracy, especially on small sample datasets.

Key words: Naïve Bayes; Bayesian network Augmented Naïve Bayes; max-relevance; min-redundancy; dependence

分类在生物信息学、图像识别、医疗诊断、自然语言处理等领域有着广泛的应用[1-2]。分类的目标是构造一个分类器,对由属性集描述的实例指定最合适的类标签。贝叶斯网络[3]提供了一种表示因果关系的方法,它结合图模型理论和统计学来表达随机变量之间的不确定性知识,并高效地执行推理任务。由于具有坚实的理论基础以及综合先验信息和数据样本信息的能力,利用贝叶斯网络处理分类问题由来已久。朴素贝叶

收稿日期:2013-12-26
基金项目:国家自然科学基金资助项目(61172168)
作者简介:冯月进(1970-),哈尔滨理工大学博士,主要从事自动分类,聚类及人工智能的信息处理研究,(Tel)13701255205;(E-mail)yifeng@hotmail.com。

斯分类器[4]是一种简单而有效的贝叶斯网络,它假设各个属性变量在给定分类变量后,是相互条件独立的。朴素贝叶斯分类器对很多问题有着很好的分类准确性,但属性变量间的条件独立性在大多数实际问题中明显不成立。因此,通过对朴素贝叶斯分类模型的改进,使得分类器的分类准确性得到进一步提升是机器学习和数据挖掘的研究热点之一。

通常存在2类方法改进朴素贝叶斯分类器:第一类方法是选择特征子集[5-6],寻找一个最佳的自变量子集来构造模型;第二类方法是放松独立性假设[7-9],改进贝叶斯网络分类器模型,使得模型对自变量的相关性结构有更为准确和灵活的建模,从而提高模型的分类准确性。另外,还有把这两种方法结合起来的混合式学习方法[10]。当前,对于第2类方法的研究比较多。Friedman[7]研究了具有树结构的TAN (tree augmented naïve bayes)分类器,它放松了朴素贝叶斯分类模型中的独立性假设条件,扩展了朴素贝叶斯网络的结构。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。BAN[8]进一步扩展了TAN结构,允许属性之间可以形成任意有向无循环图,使其表示属性之间依赖关系的能力增强,可以进一步提高分类准确性。但是,BAN结构的学习需要高维计算,在小采样数据集上,学到的结果模型可能存在较大误差,影响分类准确性。

提出改进的最大相关最小冗余技术,用于确定每个属性结点的候选连接,在这个过程中,改进的最大相关最小冗余技术使用三维计算代替高维计算;同时,为了减小BAN分类器学习的计算维度,本文提出了一种受限的BAN分类器 k-BAN,即允许每个属性结点最多可以依赖 k($k<<N$,N 为属性结点个数)个非类结点,提高了计算的可靠性和健壮性,使BAN分类器适用于小采样数据集。并且,本文提出的受限BAN分类器学习算法(k-BAN)对结点与其候选父结点集之间的所有连接实现了穷尽查找,选择使分类准确性改进最大的连接关系集,由于每次加入多条有向边($k\geqslant$edges$\geqslant 0$),与通用启发式学习算法(每次最多加入1条边)相比较,进一步减小了产生局部最优化结果的可能性。

1 相关概念和定理

1.1 有限样本数据集合

假设,存在关于一组变量 $X=\{x_1,x_2,\cdots,x_n\}$ 的采样数据集合 D,$P(x_i,x_k,\cdots,x_j)$,$(1\leqslant i,k,j\leqslant n)$表示变量组$(\{x_i,x_k,\cdots,x_j\})(1\leqslant i,k,j\leqslant n)$之间的联合概率分布。采样数据集合 D 称为有限样本数据集合,当且仅当,从采样数据集合 D,只能准确计算低维(维度≤3)变量组之间的联合概率值,并对于较高维(维度>3)变量组之间的联合概率不能准确计算,只能获得近似值。

1.2 贝叶斯网络分类器模型(BAN)

BAN进一步扩展了NB和TAN,它允许属性结点之间可以形成任意的有向无循环图,而不只是树,使其表示依赖关系的能力增强,从而进一步增强分类的正确性。Friedman提出conditional log likelihood(CLL)计分方法[7],学习BAN分类器。Cheng和Greiner提出基于条件独立性测试(conditional mutual information,CMI)[8]的BAN学习算法。BAN学习算法都需要高维计算,因此,在大多数测试数据集,特别是在高维小采样数据集上,其分类效果较NB和TAN差。

1.3 最大相关-最小冗余和最大互信息关系定理

从信息论上看,特征选择中,最大互信息方法[11]就是找到一个含有 m 个特征的特征集合 S_m,使其与目标变量 t 的互信息最大,定义为:$\underset{S_m}{\operatorname{argmax}} I(S_m;t)$。

最大相关-最小冗余是特征选择中涉及的概念[12],定义为

公式1 假设从特征集合 X 中已经选择了 $m-1$ 个特征,用 S_{m-1} 表示。目标变量为 t,目标是从特征集合 $X-S_{m-1}$ 中,选择一个特征 $x_m=\underset{x_j\in X-S_{m-1}}{\operatorname{argmax}}\left[I(x_j;t)-\frac{1}{m-1}\sum_{x_i\in S_{m-1}} I(x_j;x_i)\right]$。

公式2 在每次特征选择,只选择一个特征的条件下,最大相关-最小冗余方法最佳近似于

$$x_m=\underset{x_j\in X-S_{m-1}}{\operatorname{argmax}}\left[I(x_j;t)-\frac{1}{m-1}\sum_{x_i\in S_{m-1}} I(x_j;x_i)\right]\approx\underset{x_j\in X-S_{m-1}}{\operatorname{argmax}}[I(\{S_{m-1},x_j\};t)]。$$

公式2相似于最大相关-最小冗余特征选择公式1,公式(2)的第二项要求特征集合中的各个特征最大相互独立(即最小冗余),同时,第一项要求特征集合中的每个特征最大依赖于目标变量 C。Peng和Ding通过

实验已经验证[12],在小采样数据集上,如果一次只加入一个特征(即贪婪算法),那么 MRMR 准则是最大互信息方法的近似最佳实现方案。

2　改进最大相关最小冗余特征选择技术和受限 BAN 模型

2.1　改进的最大相关最小冗余特征选择技术

假设分类属性为 C,目标属性为 t,从属性集合 X 中已经选择了 $m-1$ 个特征,用 S_{m-1} 表示;从属性集合 $X-S_{m-1}$ 中,选择一个属性

$$x_m = \underset{x_j \in X-S_{m-1}}{\operatorname{argmax}} \left[I(x_j;t \mid C) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;x_i \mid C) \right], \tag{1}$$

使得$\{S_{m-1},x_m\}$属性集合与目标属性 t 的条件互信息近似最大,即 $x_m \approx \underset{x_j \in X-S_{m-1}}{\operatorname{argmax}} [I(\{S_{m-1},x_j\};t|C)]$。

2.2　受限 BAN 分类模型(k-BAN)

贝叶斯分类模型是一种典型的基于统计方法的分类模型,它将事件的先验概率与后验概率联系起来,利用先验信息和样本数据信息确定事件的后验概率。

假设 $U=\{A_1,A_2,\cdots,A_n,C\}$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,Π_{A_i} 表示 A_i 的父结点集合(不考虑 C)。实例 $x_i=\{a_1,a_2,\cdots,a_n\}$属于类 c_j 的概率,可表示为

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

$$\alpha P(c_j)P(a_1,a_2,\cdots,a_n \mid c_j) = \alpha P(c_j)\prod_{i=1}^{n} P(a_i \mid \Pi_{A_i},c_j)。$$

因此,给定某一实例$\{a_1,a_2,\cdots,a_n\}$,选择使 $P(c_j)\prod_{i=1}^{n} P(a_i \mid \Pi_{A_i},c_j)$ 最大的类 c_j 作为该实例的类标签。

受限 BAN 分类模型(k-BAN)假设 $|\Pi_{A_i}| \leqslant k$,其中$(i=1,\cdots,n)$。当 $k=0$ 时,受限 BAN 模型等价为 NB 分类模型;当 $k=1$ 时,受限 BAN 模型等价为 TAN 分类模型;当 $k=n$ 时,受限 BAN 模型等价为不受限 BAN 分类模型。其中,A_i 候选父结点集合 S_m 为$\{A_{l1},A_{l2},\cdots,A_{lk}\}$,$A_{lj} \neq A_i, j=1,\cdots,k$,使得 $I(S_m,A_i \mid C)$ 条件互信息最大,即,$\underset{S_m}{\operatorname{argmax}} I(S_m;A_i \mid C)$。

3　受限 BAN 分类器学习算法

受限 BAN 分类器学习算法 k-BAN 实现了 4 个方面的改进。

1)使用改进的最大相关最小冗余特征选择技术,即公式(1),替换基于最大条件互信息的 CLL 或 CMI 公式,获得给定变量的候选父结点集合,从而将确定候选父结点集合的计算维度降到三维,在小采样数据集上,提高了计算的可靠性和健壮性。

2)提出受限 BAN 模型,限制每个结点的父结点集合中变量的个数不超过 k,将条件概率 $P(a_i|\Pi_i,C)$的计算维度降到 $k+1$,当 k 较小时,在小采样数据集上,进一步提高了计算的可靠性和健壮性。

3)本算法的每次贪婪查找学习到多条有向连接边$(0 \leqslant \text{edges} \leqslant k)$,与通用 BAN 学习算法(每次贪婪查找最多学习到 1 条有向连接边)相比较,能够更容易地跳出局部最优值。

4)每次贪婪查找,通过对结点和其候选父结点之间的所有有向连接边集进行穷尽搜索,进一步减小了所学分类器的局部最优化。

3.1　算法描述

输入:分类结点为 C,n 个结点变量集合 $X=\{A_1,A_2,\cdots,A_n\}$,X_{handled}表示已经处理过的结点集合,A_{best}表示在一次贪婪查找后,要加入到 X_{handled}的结点;$c\pi_i$ 表示 A_i 的候选父结点集合,π_i 表示 A_i 的当前父结点集合;E_{im}表示 A_i 与 $c\pi_i$ 之间任意一个有向连接边集,E_i 表示 A_i 与 $c\pi_i$ 之间的所有的有向边集的集合 $E_i=\{E_{im}\}$,$(0 \leqslant m < 3^k)$,S_{edges}表示一次贪婪查找中,可能加入到当前分类器的所有有向连接边集的集合 $S_{\text{edges}}=\{E_i\}$,$(1 \leqslant i \leqslant n)$;$E_{\text{cur}}$表示当前分类器 $\text{Class}_{\text{cur}}$的有向边集,$E_{\text{new}}$表示新分类器 $\text{Class}_{\text{new}}$的有向边集;$\text{Accuracy}_{\text{cur}}$表示当前分类器 $\text{Class}_{\text{cur}}$的分类性能,$\text{Accuracy}_{\text{new}}$表示新分类器 $\text{Class}_{\text{new}}$的分类性能;$\text{Class}_{\text{best}}$表示当前获得的

分类性能最好的分类器，$Accuracy_{best}$表示 $Class_{best}$的分类性能；布尔变量 isModified=true 表示当前的贪婪查找获得了分类性能更好的分类器，否则，表示贪婪查找过程结束。MI(n,n)是一个 $n \cdot n$ 的二维数组。

约束条件：$|\pi_i| \leqslant k, 1 \leqslant i \leqslant n$，

输出：受限的 BAN 分类器。

算法：

/ * 第一部分：初始化 * /

/ * 求出结点间的两两条件互信息 * /

1) initialize $MI(n, n)$ to 0；
2) for i=1 to n do
3) for j= i+1 to n do
4) begin
5) $MI[i,j]=MI[j,i]= I(x_i,x_j|C)$；
6) end for
7) end for

/ * 初始化当前分类器为朴素贝叶斯分类器，并计算 NB 的分类性能 * /

8) $Class_{cur}=NB$；$Accuracy_{cur}=Accuracy_{NB}$；$\pi_i=\Phi,(1 \leqslant i \leqslant n)$；$X_{handled}=\Phi$；；

/ * 第二部分：进行贪婪查找，每次加入一个结点的多条有向连接边，使获得的分类器的分类性能最好，直到不能找到分类性能更好的分类器为止 * /

9) do { // 一次贪婪查找过程
10) for each A_i,$(1 \leqslant i \leqslant n)$
11) $c\pi_i=\Phi$；
12) end for
13) $S_{edges}=\Phi$；

/ * 求出每个未处理结点 A_i 的 $c\pi_i$ * /

14) for each A_i,$A_i \notin X_{handled}$,$1 \leqslant i \leqslant n$
15) while $|c\pi_i|+|\pi_i|<k$
16) $$y=\underset{A_j \in X-A_i-c\pi_i-\pi_i}{\operatorname{argmax}}\left[MI[A_i,A_j|C]-\frac{1}{|\pi_i|+|c\pi_i|}\sum_{A \in \pi_i+c\pi_i} MI[A,A_j|C]\right]$$
17) $c\pi_i=c\pi_i \bigcup \{y\}$；
18) end while
19) end for

/ * 对每个 $c\pi_i$，确定 A_i 与 $c\pi_i$ 结点之间的所有有向连接边集的集合 E_i，加入 S_{edges} * /

20) for each $c\pi_i$,$c\pi_i \neq \Phi$,$1 \leqslant i \leqslant n$
21) acquire all the sets E_i of directed edges between A_i and $c\pi_i$；
22) $S_{edges}=S_{edges} \bigcup E_i$；
23) end for

/ * 对于每个有向连接边集，将其加入当前分类器，获得一个新的分类器，并进行比较处理 * /

24) $Class_{best}=Class_{cur}$；$Accuracy_{best}=Accuracy_{cur}$；$A_{best}=\Phi$；stModified=false；
25) for each E_{im},$E_{im} \in S_{edges}$,$1 \leqslant i \leqslant n$。
26) acquire a new classifier $Class_{new}$,$E_{new}=E_{cur} \bigcup E_{im}$；

/ * 判断分类器 $Class_{new}$是否满足受限 BAN 的约束条件 * /

27) for each nodeA_j(1≤j≤n)
28) Compute the number $ParentNum_j$ of parent nodes for A_j in the $Class_{new}$；
29) end for
30) if all $ParentNum_j \leqslant k$,$(1 \leqslant j \leqslant n)$ then

```
        /* 如果满足约束条件,进行如下比较处理: */
31)       compute Accuracy_new for Class_new;
32)     if Accuracy_new > Accuracy_best then
33)         Class_best = Class_new; Accuracy_best = Accuracy_new; A_best = A_i;
34)         stModified = true;
35)       end if
36)     end if
37)   end for // corresponding to line 25)
/* 如果在一次贪婪查找中,获得了一个分类性能更好的分类器,则如下处理: */
38)   ifstModified == true then
39)     Class_cur = Class_best, Accuracy_cur = Accuracy_best, X_handled = X_handled ∪ {A_best};
40)     for each node A_j (1≤j≤n)
41)       Update π_j according to the Class_cur;
42)     end for
43)   endif
44) while,                                                    (stModified == true);
45) return Class_cur;
```

受限 BAN 分类器学习算法分为 2 个部分:

第一部分主要包括计算结点间的互信息,初始化分类器为朴素贝叶斯分类器,并计算其分类性能;

第二部分采用贪婪启发式方法,结合改进的最大相关最小冗余技术,确定结点的候选父结点集(结点个数不超过 k),穷尽查找每个结点和其候选父结点之间的有向连接边集,把分类性能近似最佳的有向连接边集加入当前分类器中,直到无法获得具有更好分类性能的分类器。

注:算法中第(21)行的解释:假设结点 A_i 的候选父结点集合 $c\pi_i=\{A_1,A_2\}$,那么,A_i 与 $c\pi_i$ 结点间的所有有向连接边集合 $E_i=\{E_{im},1\leqslant m\leqslant 3^{|c\pi_i|}\}$,其中,$E_{i1}=\Phi$;$E_{i2}=\{A_1\rightarrow A_i\}$;$E_{i3}=\{A_i\rightarrow A_1\}$;$E_{i4}=\{A_2\rightarrow A_i\}$;$E_{i5}=\{A_i\rightarrow A_2\}$;$E_{i6}=\{A_1\rightarrow A_i,A_2\rightarrow A_i\}$;$E_{i7}=\{A_1\rightarrow A_i,A_i\rightarrow A_2\}$;$E_{i8}=\{A_i\rightarrow A_1,A_2\rightarrow A_i\}$;$E_{i9}=\{A_i\rightarrow A_1,A_i\rightarrow A_2\}$。

4　实验结果和分析

实现了受限 BAN 分类器 k-BAN($k=3$)、NB 分类器、TAN 分类器和 BAN 分类器。在配置为 Pentium5 3GHz,2G RAM,Windows XP 的计算机上进行实验,并且在算法的分类正确性上给出比较结果。考虑到,在小采样数据集上,$P(a_i|\Pi_i,C)$的可靠性和健壮性,选取较小的 k 值,$k=3$。

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。由于算法不能处理连续型数值数据,因此,使用 MLC++中[13]的离散化工具对连续型数值离散化。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。

表 1　实验数据集的构成描述

Domain	Size#	Classes#	Attributes#	Missing vlaue
Anneal	898	6	38	Yes
Car	1728	4	6	No
Cleveland	303	2	13	No
Horse-Colic	368	2	22	No
House-Votes-84	435	2	16	Yes
Mushroom	8124	2	22	Yes
Nursery	12960	5	8	No
Promoter Gene Sequences	106	2	57	No
Flare-C	1389	2	13	No

实验的主要目的是对 k-BAN(k=3)与 NB,TAN 和 BAN 分类器在每个数据集上的分类正确率进行比较。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。4 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方差。

表 2 4 种分类器的实验结果

Domain	NB	TAN	BAN	k-BAN(k=3)
Anneal	96.247 5±0.27	96.236 3±0.27	92.655 3±0.41	96.742 9±0.29
Car	85.575 7±0.32	94.600 1±0.21	94.040 1±0.44	93.691 0±0.27
Cleveland	83.157 1±0.69	81.448 3±0.94	80.921 1±0.33	86.196 2±0.34
Horse-Colic	80.258 1±0.52	80.937 5±0.57	76.273 3±1.12	81.235 7±0.52
House-Votes-84	90.069 0±0.14	93.195 4±0.32	88.147 6±0.26	95.034 4±0.22
Mushroom	95.768 0±0.03	99.409 0±0.03	99.831 1±0.03	99.488 9±0.03
Flare-C	79.013 7±0.23	83.120 9±0.31	82.850 9±0.31	83.106 1±0.29
Nursery	90.284 7±0.05	92.531 9±0.23	93.750 8±0.39	93.033 3±0.19
Promoter Gene Sequences	91.273 8±1.76	82.971 7±3.26	80.462 3±2.94	92.276 7±1.19

从结果比较中可以看出,对于采样充分的大数据集(mushroom nursery),BAN 的分类正确性最好,k-BAN(k=3)次之,但是分类正确性优于 TAN 和 NB。当数据集合采样充分时,由于可以准确地计算变量组的联合概率值,因此,不受限的(即,无前提假设的)BAN 分类模型的分类正确性好于受限的(即,存在前提假设的)BAN 分类模型;限制条件弱的(例,3-BAN(k=3))BAN 分类模型的分类正确性好于限制条件强(例,TAN(k=1),NB(k=0))的 BAN 分类模型。

对于有限样本数据集,k-BAN(k=3)分类性能优于 BAN,TAN 和 NB。当数据集合样本数目相对变量维度较少时,只能准确计算低维变量组(维度 3)的联合概率值,3-BAN 分类正确性优于 BAN;但,3-BAN 的分类正确性优于限制条件强(例,TAN(k=1),NB(k=0))的 BAN 分类模型。

5 结 语

在贝叶斯分类器学习中,正确性是评估学习算法优劣的主要指标。通过采用改进的最大相关–最小冗余特征选择技术建立候选父节点集合和限制父结点的个数,减少了高维计算。同时,提出的算法在每次贪婪查找中,学习到多条有向连接边(最多不超过 k 条),能够更好地减小所学分类器的局部最优化;而且,每次贪婪查找,通过对结点和其候选父结点之间的所有有向连接边集进行穷尽搜索,进一步减小了所学分类器的局部最优化。因此,在维数较高和小采样数据集上,受限 BAN 学习算法在实际测试中,与 NB\TAN\BAN 分类器比较,体现出了更好的分类正确性。

参考文献:

[1] 周国强,崔荣一. 基于朴素贝叶斯分类器的朝鲜语文本分类的研究[J]. 中文信息学报,2011,25(4):16-19.
ZHOU Guoqiang,CUI Rongyi. Research on Korean text categorization based on native Bayesian classifier[J]. Journal of Chinese Information Processing,2011,25(4):16-19.

[2] 李冠广,王占杰. 贝叶斯分类器在入侵检测中的应用[J]. 信息安全与技术,2010,11:63-66.
LI Guanguang, WANG Zhanjie. The application of Bayesian classifier in intrusion de-tection[J]. Technology and

Application,2010,11:63-66.

[3] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference[M]. San Francisco: Morgan Kaufman,1998.

[4] Muralidharan V,Sugumaran V. A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis[J]. Applied Soft Computing,2012,12(8):2023-2029.

[5] David R,Telesca D,Johnson V E. High-dimensional Bayesian classifiers using non-local priors[J]. Statistical Models for Data Analysis,2013:305-313.

[6] 程玉虎,仝遥遥,王雪松.类相关性影响可变选择性贝叶斯分类器[J].电子学报,2011,39(7):1628-1633.
CHENG Yuhu, TONG Yaoyao, WANG Xuesong. A selective Bayesian classifer based on change of class relevance influence[J]. Acta Electronica Sinica,2011,39(7):1628-1633.

[7] Chen L F,Wang S R. Automated feature weighting in naive bayes for high-dimensional data classification[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management, October 29- November2, Maui, USA. New York:ACM,2012:1243-1252.

[8] Malovini H,Barbarini N,Bellazzi R,et al. Hierarchical naive Bayes for genetic association studies[J]. BMC Bioinformatics, 2012,13(Sup14):1-11.

[9] 石洪波,王志海,黄厚宽,等.一种限定性的双层贝叶斯分类模型[J].软件学报,2004,15(2):193-199.
SHI Hongbo,WANG Zhihai,HUANG Houkuang,et al. A restricted double-level Bayesian classification model[J]. Journal of Software,2004,15(2):193-199.

[10] Singh M,Provan G M. Efficient learning of selective Bayesian network classifiers[M]. Italy:Morgan Kaufmann,1996.

[11] Lin H C,Su C T. A selective Bayes classifier with meta-heuristics for incomplete data[J]. Neurocomputing,2013,106: 95-102.

[12] Peng H C,Long F M,Ding C. Feature selection based on mutual information:criteria of max-dependency,max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(8):1226-1233.

[13] Kohavi R, John G, Long R. MLC++: A machine learning library in C++. ICTAI, New Orleans, Louisiana: IEEE Computer Society,1994. 740-743.

(编辑　侯　湘)