PG-HMI:一种基于互信息的特征选择方法*

王 皓 孙宏斌 张伯明

(清华大学 电机工程与应用电子技术系 电力系统国家重点实验室 北京 100084)

摘 要 传统的基于样本的互信息估计方法不能直接处理离散、连续属性混合的情况.本文给出一种能够直接处理混合属性的互信息估计方法(PG 法).为了更好地考虑属性之间的关联,提出名为 HMI 的特征选择准则.结合 PG 互信息估计方法和 HMI 特征选择准则,给出一种新的特征选择方法(PG-HMI).实验结果验证 PG 互信息估计法的合理性及 PG-HMI 特征选择方法的有效性.

关键词 特征选择,互信息,混合互信息(HMI),分类器,数据挖掘

中图法分类号 TP181

PG-HMI: Mutual Information Based Feature Selection Method

WANG Hao, SUN Hong-Bin, ZHANG Bo-Ming

(*State Key Laboratory of Power Systems, Department of Electrical Engineering and Applied Electronic Techniques, Tsinghua University, Beijing* 100084)

ABSTRACT

Conventional sample-based mutual information estimation methods can't handle the mixed features directly that include both numeric attributes and nominal attributes. A Parzen window based general mutual information calculation method, PG method, is proposed in this paper, which could deal with the mixed attributes directly. A criterion named hybrid mutual information (HMI) is presented. Based on PG mutual information estimation method and HMI feature selection criterion, a feature selection algorithm (PG-HMI) is proposed. Experimental results show the correctness of PG and the effectiveness of PG-HMI.

Key Words Feature Selection, Mutual Information, Hybrid Mutual Information (HMI), Classifier, Data Mining

* 国家自然科学基金重大项目(No. 50595414)、国家重点基础研究发展规划项目(No. 2004CB217904)、国家自然科学基金项目(No. 50107005)和新世纪优秀人才支持计划项目资助

收稿日期:2006-01-18;修回日期:2006-07-20

作者简介 **王皓**,男,1981 年生,硕士研究生,主要研究方向为电力系统信息论和数据挖掘. E-mail: wang_hao00@tsinghua.org.cn. **孙宏斌**,男,1969 年生,教授,主要研究方向为全局电压优化控制、电力系统信息理论和数据挖掘. **张伯明**,男,1948 年生,教授,主要研究方向为电力系统运行、分析和控制.

1 引 言

在数据挖掘领域的许多应用中(如训练分类器、构建回归模型),寻找一组合适的特征作为输入对训练效率、模型复杂程度、泛化能力有至关重要的影响[1-7].特征选择是特征空间降维的重要手段,将特征选择所获得的特征属性作为数据挖掘的输入属性,可以有效加快分类器训练速度、降低分类模型复杂度、提高分类器泛化能力.

本文重点研究基于互信息的特征选择方法.在数据挖掘领域的应用中,互信息是通过统计样本数据得到的,本文称这类互信息估计方法为基于样本统计的互信息估计方法.基于样本统计的互信息估计方法按所能够处理的样本属性取值类型可分为3类:离散型、连续型、统一型.离散型方法只能处理离散属性,对于连续属性,则需要进行离散化处理,使之成为离散属性.传统离散型方法的计算时间复杂度和空间复杂度均为属性个数的指数函数[4-6],难以计算高维互信息.连续型方法(如PW[5]、OFS-MI[7])致力于直接处理连续属性.但是文献[5]和[7]中提出的计算方法没有考虑离散、连续属性同时存在的情况,也没有考虑目标属性 C 为连续属性的情况,使其应用受到局限.本文提出一种统一型计算方法——PG法,该方法对属性类型没有任何限制,是一种更加普遍的互信息估计方法.

特征选择准则在特征选择方法中起重要作用,极大地影响特征选择的效果.信息增益(Information Gain[2,8], IG)准则由于不考虑已选属性间的相关性,致使已选属性间可能有较大的信息重叠,难以获得精简、高效的结果.MIFS法[3]、MIFSU[4]用多个二维互信息来间接反映高维数据样本之间的关联,但是这些近似方法都是较为粗糙的.文献[5]和[7]直接采用高维互信息作为准则,更全面地考虑了属性之间的关联,但这类方法在高维时容易引入噪声属性.在已有特征选择准则的基础上,本文提出一种新的特征选择准则——混合互信息(Hybrid Mutual Information, HMI).HMI同时考虑属性 f 能够提供的新信息量 $I(C;f\mid S)$ 以及属性 f 与类别标号属性 C 的相关性 $I(C;f)$.以HMI作为属性度量准则,既能有效获取新信息,又能有效控制噪声的引入.

结合PG互信息计算方法和HMI特征选择准则,本文给出PG-HMI特征选择算法.

2 背景知识

2.1 熵、互信息、条件互信息

根据shannon的信息理论[9-10],熵(entropy)是随机变量不确定性的度量.一个离散随机变量 X,其可能取值集合记为 S_x,对应于 $x\in S_x$,其概率为 $p(x)$,则 X 的熵定义为

$$H(X)=-\sum_{x\in S_x}p(x)\log p(x). \tag{1}$$

当变量 Y 已知,变量 X 中剩余的不确定性用条件熵(conditional entropy)来度量:

$$H(X\mid Y)=-\sum_{x\in S_x}\sum_{y\in S_y}p(x,y)\log p(x\mid y). \tag{2}$$

条件熵与熵有如下关系:

$$H(X\mid Y)=H(XY)-H(Y). \tag{3}$$

两个随机变量 X 和 Y 的统计依存关系用互信息(mutual information)来度量:

$$I(X;Y)=\sum_{x\in S_x}\sum_{y\in S_y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

如果两个随机变量的互信息较大,则这两个随机变量相关性较大.互信息和熵有如下关系:

$$\begin{aligned}I(X;Y)&=H(X)-H(X\mid Y)\\&=H(Y)-H(Y\mid X)\\&=H(X)+H(Y)-H(XY).\end{aligned} \tag{5}$$

在 Z 已知的条件下,X 和 Y 的统计依存度可以用条件互信息来表示:

$$I(X;Y\mid Z)=\sum_{x\in S_x}\sum_{y\in S_y}\sum_{z\in S_z}p(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)}. \tag{6}$$

条件互信息与熵有如下关系:

$$I(X;Y\mid Z)=H(XZ)+H(YZ)-H(Z)-H(XYZ). \tag{7}$$

对于连续随机变量,微分熵、互信息、条件互信息分别定义如下:

$$H(X)=-\int_x p(x)\log p(x)\mathrm{d}x, \tag{8}$$

$$I(X;Y)=\iint_{x,y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y, \tag{9}$$

$$I(X;Y\mid Z)=\iiint_{x,y,z}p(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)}\mathrm{d}x\mathrm{d}y\mathrm{d}z, \tag{10}$$

式(1)~式(10)给出随机变量X、Y、Z的熵、互信息、条件互信息的计算公式及关系,对于随机矢量X、Y、Z,公式形式保持一致.

2.2 用Parzen窗估计概率密度函数

Parzen窗估计(亦称核函数估计)[11-12]是一种

非参数估计的概率密度函数估计方法. 对于给定 n 个样本的 D 维连续随机矢量 $\boldsymbol{X}$,其概率密度函数 $p(\boldsymbol{x})$ 可以按下式进行估计:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{x}-\boldsymbol{x}_i,h), \tag{11}$$

其中,$\boldsymbol{x}_i$ 是第 i 个 D 维样本,n 为样本总数,h 为窗宽,$\phi(\cdot)$ 为 Parzen 窗函数.

高斯窗函数如下式所示:

$$\phi(\boldsymbol{x},h) = \frac{1}{(2\pi)^{D/2}h^D \mid \boldsymbol{\Sigma} \mid^{1/2}}\exp\left(-\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}{2h^2}\right), \tag{12}$$

其中 $\boldsymbol{\Sigma}$ 为协方差矩阵.

Parzen 证明[11]:当 $\phi(\cdot)$ 和 h 选择适当时,$\hat{p}(\boldsymbol{x})$ 收敛到真实概率密度函数 $p(\boldsymbol{x})$.

3　统一型互信息计算方法——PG 法

互信息可以由熵表示,因此首先讨论统一型熵的计算. 计算出相应的熵后,便可用式(5)得到互信息.

在讨论统一型熵的计算方法时,为了表达简便,假设矢量 $\boldsymbol{Z}$ 的前 K 个属性 $U_1,\cdots,U_K$ 为离散属性,后 $D-K$ 个属性 $X_1,\cdots,X_{D-K}$ 为连续属性,即

$$\boldsymbol{Z} = UX = U_1,\cdots,U_K,X_1,\cdots,X_{D-K}.$$

$H(\boldsymbol{Z})$ 可以分成 $H(U)$ 和 $H(X \mid U)$ 两项,参见下式:

$$H(\boldsymbol{Z}) = H(UX) = H(U) + H(X \mid U). \tag{13}$$

下面分别讨论这两项的计算方法.

3.1　$H(U)$ 的计算

传统的离散熵算法需要开辟额外的内存空间保存联合概率,导致算法的空间复杂度为属性个数的指数函数[4-6]. 通过研究发现,额外开辟内存空间保存联合概率是不必要的. 本文提出一种"排序 - 遍历"(Sort - Traverse,ST) 式的离散熵算法. 该法不需额外内存,时间复杂度为 $O(Dn\log n)$,其中,n 为样本数,D 为属性数.

下面说明 K 维矢量 $\boldsymbol{U} = U_{i_1},\cdots,U_{i_K}$(其中 $i_1,\cdots,i_K \in \{1,\cdots,D\}$)的熵 $H(\boldsymbol{U})$ 的计算方法. K 维矢量 $\boldsymbol{U}$ 某一特定取值 $\boldsymbol{u}$ 的概率 $p(\boldsymbol{u})$,可以用 n_u/n 估计,其中 n_u 为样本空间中对应于 $\boldsymbol{u}$ 的样本数,n 为样本总数. 由下式可知,只要逐个确定 K 维矢量所有出现的取值[1)] 的样本数 n_u,便可算出 $H(\boldsymbol{U})$.

$$\begin{aligned} H(\boldsymbol{U}) &= -\sum_{u\in S_u} p(\boldsymbol{u})\log p(\boldsymbol{u}) \\ &= -\sum_{u\in S_u}\frac{n_u}{n}\log\frac{n_u}{n} \\ &= \log n - \frac{1}{n}\sum_{u\in S_u} n_u\log n_u. \end{aligned} \tag{14}$$

为此,ST 法计算熵 $H(\boldsymbol{U})$ 的总体思想为:首先通过排序,使具有相同取值的样本聚集在一起;其次,通过遍历,获得所有出现的取值的样本数;最后,按照式(14)得到 $H(\boldsymbol{U})$.

用 ST 法计算 K 维矢量 $\boldsymbol{U}$ 的熵 $H(\boldsymbol{U})$ 的总体计算步骤如下:

输入　数据编码表,K 维矢量 $\boldsymbol{U}$

输出　$\hat{H}(\boldsymbol{U})$

step1　$\hat{H}(\boldsymbol{U}) = 0$;

step2　对数据编码表"按 $\boldsymbol{U}$ 排序";

step3　顺序遍历有序编码表,对于每一个具有相同 $\boldsymbol{U}$ 码值的数据段

　　step3.1　统计其段内样本数 n_u;

　　step3.2　$\hat{H}(\boldsymbol{U}) = \hat{H}(\boldsymbol{U}) + n_u\log n_u$;

　　step3.3　$n = n + n_u$;

step4　$\hat{H}(\boldsymbol{U}) = \log(n) - \hat{H}(\boldsymbol{U})/n$;

step5　输出 $\hat{H}(\boldsymbol{U})$.

"按 $\boldsymbol{U}$ 排序"的核心是排序中"按 $\boldsymbol{U}$ 比较"两个样本的大小. 本文定义:对于两个样本(每个样本为一个数组),逐个比较属性值的大小,第一个不相等属性值大的为大. 在计算过程中,离散属性的取值由一个以零起始的编号代替,从而形成数据编码表. 因此,属性值的比较实际上是整数的比较.

ST 法的时间复杂度为 $O(Dn\log(n))$,空间复杂度为 $O(1)$,即不需额外内存.

3.2　$H(X \mid U)$ 的估计

由下式知,计算 $H(\boldsymbol{X} \mid \boldsymbol{U})$ 需要解决两个问题,即如何估计概率密度函数 $p(\boldsymbol{x} \mid \boldsymbol{u})$ 以及如何计算积分.

$$\begin{aligned} H(\boldsymbol{X} \mid \boldsymbol{U}) &= -\sum_{u\in S_u}\int_x p(\boldsymbol{u},\boldsymbol{x})\log p(\boldsymbol{x} \mid \boldsymbol{u})\mathrm{d}\boldsymbol{x} \\ &= -\sum_{u\in S_u} p(\boldsymbol{u})\int_x p(\boldsymbol{x} \mid \boldsymbol{u})\log p(\boldsymbol{x} \mid \boldsymbol{u})\mathrm{d}\boldsymbol{x} \\ &= \sum_{u\in S_u} p(\boldsymbol{u})H(\boldsymbol{X} \mid \boldsymbol{u}), \end{aligned} \tag{15}$$

其中

$$H(\boldsymbol{X} \mid \boldsymbol{u}) = -\int_x p(\boldsymbol{x} \mid \boldsymbol{u})\log p(\boldsymbol{x} \mid \boldsymbol{u})\mathrm{d}\boldsymbol{x}, \tag{16}$$

S_u 为 $\boldsymbol{u}$ 所有可能取值的集合.

1) 若某一取值 $\boldsymbol{x}$ 在样本中不出现,则 $n_x=0$,$n_x\log n_x=0$,对熵 $H(\boldsymbol{X})$ 的值没有影响.

$p(\boldsymbol{x}\mid\boldsymbol{u})$ 可以用 Parzen 窗函数估计. 如用高斯窗函数,则 $p(\boldsymbol{x}\mid\boldsymbol{u})$ 的估计式 $\hat{p}(\boldsymbol{x}\mid\boldsymbol{u})$ 可以写成如下的形式:

$$\hat{p}(\boldsymbol{x}\mid\boldsymbol{u})=\frac{1}{n_{\boldsymbol{u}}}\sum_{i\in I_{\boldsymbol{u}}}\phi(\boldsymbol{x}-\boldsymbol{x}_i,h)$$
$$=\frac{1}{n_{\boldsymbol{u}}(2\pi)^{D/2}h^{D}\mid\boldsymbol{\Sigma}\mid}\sum_{i\in I_{\boldsymbol{u}}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)}{2h^2}\right),\tag{17}$$

其中,$I_{\boldsymbol{u}}$ 是 $\boldsymbol{U}$ 值取 $\boldsymbol{u}$ 的样本的编号集合;$\boldsymbol{\Sigma}$ 为协方差矩阵,由于概率密度估计时通常只考虑对角项(非对角项误差较大),而连续属性已经进行标准化处理(方差为 1),所以本法中的 $\boldsymbol{\Sigma}$ 为单位矩阵.

积分运算是很困难的,本法将积分运算简化为代数和. $\hat{p}(\boldsymbol{x}\mid\boldsymbol{u})$ 值大的区域表示该区域样本出现概率高、分布较为密集,因而在积分转化为代数和的时候每个样本所代表的区域相应地较小. 在积分转化为代数和的转换中,本文假设每个样本所能代表的区域与该样本所在点的概率密度成反比,即

$$\Delta\boldsymbol{x}_j\propto\frac{1}{\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})},$$

从而得到

$$\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})\Delta\boldsymbol{x}_j=\text{const.}$$

因为 $\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})$ 是概率密度函数的估计式,积分转化为代数和的过程中,理论上应满足概率密度函数的归一性,即

$$1=\int_{\boldsymbol{x}}p(\boldsymbol{x}\mid\boldsymbol{u})\mathrm{d}\boldsymbol{x}=\sum_{j\in I_{\boldsymbol{u}}}\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})\Delta\boldsymbol{x}_j,$$

又因为 $j\in I_{\boldsymbol{u}}$ 的样本共有 $n_{\boldsymbol{u}}$ 个,从而得到

$$\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})\Delta\boldsymbol{x}_j=1/n_{\boldsymbol{u}}.$$

按以上方式简化后,$H(\boldsymbol{X}\mid\boldsymbol{U})$ 的估计值 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{u})$ 计算如下:

$$\hat{H}(\boldsymbol{X}\mid\boldsymbol{u})=-\int_{\boldsymbol{x}}\hat{p}(\boldsymbol{x}\mid\boldsymbol{u})\log\hat{p}(\boldsymbol{x}\mid\boldsymbol{u})\mathrm{d}\boldsymbol{x}$$
$$=-\sum_{j\in I_{\boldsymbol{u}}}\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})\log\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})\Delta\boldsymbol{x}_j$$
$$=-\frac{1}{n_{\boldsymbol{u}}}\sum_{j\in I_{\boldsymbol{u}}}\log\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u}).\tag{18}$$

将式(17)和式(18)带入式(15),则得到 $H(\boldsymbol{X}\mid\boldsymbol{U})$ 的估计式:

$$\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})$$
$$=\sum_{\boldsymbol{u}\in S_{\boldsymbol{u}}}p(\boldsymbol{u})\hat{H}(\boldsymbol{X}\mid\boldsymbol{u})$$
$$=-\frac{1}{n}\sum_{\boldsymbol{u}\in S_{\boldsymbol{u}}}\sum_{j\in I_{\boldsymbol{u}}}\log\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})$$
$$=-\frac{1}{n}\sum_{\boldsymbol{u}\in S_{\boldsymbol{u}}}\sum_{j\in I_{\boldsymbol{u}}}\log\frac{1}{n_{\boldsymbol{u}}(2\pi)^{D/2}h^{D}\mid\boldsymbol{\Sigma}\mid}\cdot$$
$$\sum_{i\in I_{\boldsymbol{u}}}\exp\left(-\frac{(\boldsymbol{x}_j-\boldsymbol{x}_i)^{T}\Sigma^{-1}(\boldsymbol{x}_j-\boldsymbol{x}_i)}{2h^2}\right).\tag{19}$$

$H(\boldsymbol{X}\mid\boldsymbol{U})$ 的计算步骤如下:

输入 部分编码表,D 维矢量 $\boldsymbol{Z}$,Parzen 窗宽 h

输出 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})$

step1 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})=0$;

step2 对部分编码表"按 $\boldsymbol{U}$ 排序";

step3 遍历部分编码表,统计每一个具有相同 $\boldsymbol{U}$ 码值的数据段的样本数 $n_{\boldsymbol{u}}$ 和样本总数 n;

step4 对于每一个具有相同 $\boldsymbol{U}$ 码值的数据段

step4.1 按式(18)计算 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{u})$,其中 $\hat{p}(\boldsymbol{x}_j\mid\boldsymbol{u})$ 按式(17)计算;

step4.2 $p(\boldsymbol{u})=n_{\boldsymbol{u}}/n$;

step4.3 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})=\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})+p(\boldsymbol{u})*\hat{H}(\boldsymbol{X}\mid\boldsymbol{u})$;

step5 输出 $\hat{H}(\boldsymbol{X}\mid\boldsymbol{U})$.

因为计算 $H(\boldsymbol{X}\mid\boldsymbol{U})$ 的时间复杂度为

$$O(D\sum\nolimits_{\boldsymbol{u}\in S_{\boldsymbol{u}}}n_{\boldsymbol{u}}^2=O(Dn^2),$$

计算 $H(\boldsymbol{U})$ 的时间复杂度为 $O(Dn\log n)$,所以 PG 法计算 $H(\boldsymbol{Z})$ 的时间复杂度为 $O(Dn^2)$. PG 法的空间复杂度为 $O(1)$,即不需额外内存.

3.3 通过熵计算互信息、条件互信息等量

在计算出 $H(\boldsymbol{U})$ 和 $H(\boldsymbol{X}\mid\boldsymbol{U})$ 后,便可得到 $H(\boldsymbol{Z})$. 计算出相应的熵后,便可根据式(3)、式(5)及式(7)计算出条件熵、互信息、条件互信息等量.

在用 PG 法的离散特例——ST 法计算离散熵、互信息、条件互信息时,不同的计算顺序,计算量会有差异. 例如,当按式(3)计算条件熵 $H(\boldsymbol{X}\mid\boldsymbol{Y})$ 时,应先按 $\boldsymbol{YX}$ 排序,计算出 $H(\boldsymbol{XY})$. 由于按 $\boldsymbol{YX}$ 排序的结果自然满足按 $\boldsymbol{Y}$ 排序的要求,故在计算 $H(\boldsymbol{Y})$ 时,不必重新排序,直接遍历即可. 用类似的方法也可以对离散互信息和离散条件互信息的计算进行进一步的简化,避免不必要的排序.

4 特征选择准则——HMI

IG 法将属性 f 与目标属性的互信息 $I(C;f)$ 作为特征选择准则,只考虑属性与目标属性的相关性,而没有考虑属性间的关联.

MIFS、MIFSU、PWFS、mRMR、OFS-MI 等采用了最大化 $I(C;f,S)$ 的特征选择准则(PWFS 和 OFS-MI 直接估计 $I(C;f_i,S)$; MIFS、MIFSU 和 mRMR 则通过二维互信息间接估计 $I(C;f,S)$). 由下式可知,最大化 $I(C;f,S)$ 等价于最大化 $I(C;f\mid S)$:

$$\max I(C;f,S)=\max\{I(C;S)+I(C;f\mid S)\}$$
$$=I(C;S)+\max I(C;f\mid S).\tag{20}$$

而 $I(C;f \mid S)$ 是在 S 已知条件下,属性 f 所能提供的新信息量.因此最大化 $I(C;f \mid S)$ 的特征选择过程便是逐个选择能够提供最多新信息量属性的过程.这种只考虑属性能够提供的新信息量的准则是存在局限的,即在已经选择较多属性的情况下,任一未选属性 f 所能提供的新信息量都是较小的,若仅从新信息量角度选择属性,则所选属性很可能与类别标号属性相关性较弱,即 $I(C;f)$ 较小,这样的属性很可能引入噪声,给特征选择的结果带来不利影响.

本文认为特征选择准则既要考虑属性能够提供的新信息量 $I(C;f \mid S)$,还要兼顾属性与类别标号属性的相关度 $I(C;f)$.因此,本文提出混合互信息(Hybrid Mutual Information,HMI)特征选择准则.混合互信息的定义如下:

$$I_{\text{hybrid}}(C;f \mid S) = (1-W) \cdot I(C;f) + W \cdot I(C;f \mid S), \quad (21)$$

其中,互信息 $I(C;f)$ 表示属性 f 与分类目标属性 C 的相关程度;$I(C;f \mid S)$ 表示在已经向特征属性集合 S 中选取一些属性的条件下,候选属性 f 所能提供的关于 C 的新信息量;式中 $W \in [0,1]$ 为混合系数.

5　特征选择方法——PG-HMI

结合PG互信息估计方法、HMI特征选择准则,并采用顺序前进搜索[12](Sequence Forward Search,SFS)搜索方式,形成一种新的特征选择方法——PG-HMI,其基本步骤如下:

step1　初始化:$F \leftarrow$ 所有的 N 个候选属性,$S \leftarrow \varnothing$.

step2　$\forall f \in F$,计算 $I(C;f)$.

step3　选择最大化 $I(C;f)$ 的属性 f,设置 $F \leftarrow F \backslash \{f\}$,$S \leftarrow S \cup \{f\}$.

step4　重复 step4.1、step4.2 直至满足终止条件:

step4.1　$\forall f \in F$,计算 $I_{\text{hybrid}}(C;f \mid S)$;

step4.2　对于最大化 $I_{\text{hybrid}}(C;f \mid S)$ 的属性 $f \in F$,设置 $F \leftarrow F \backslash \{f\}$,$S \leftarrow S \cup \{f\}$.

step5　输出特征属性集 S.

其中,终止条件为如下3个条件中任意一个满足:

1) $|S| \geq k$ 或 $|F| = 0$. $|\cdot|$ 为属性集合中属性个数.

2) $\dfrac{\Delta I(C;S)}{H(C)} < \alpha$,

其中 $\Delta I(C;S)$ 为添加 f 引起的 $I(C;S)$ 的增长量.

3) $\dfrac{I(C;S)}{H(C)} \geq \beta$.

k、α、β 的具体设置可根据实际情况确定.如可取 $\alpha = 0.005$,$\beta = 0.99$,$k = N$(N 为候选属性总数).

应用HMI法时,混合系数的确定是很重要的.混合系数应使得在算法初期,使条件互信息 $I(C;f \mid S)$ 占主导地位,使算法具有良好地获取新信息的能力;而在算法后期,使互信息 $I(C;f)$ 占较大比重,一定程度上控制噪声的引入.为了达到这个目的,可令混合系数 W 为随着属性个数递减的函数.通过分析可以发现,如果将 W 取一个较大的常数值(例如0.9)也可以满足上述要求.这是因为在特征选择前期,条件互信息数值上与互信息相差不大,所以较大的 W 便可以使得条件互信息起主导作用;在属性选取后期,条件互信息数值较小,这时互信息将起较大作用.本文实验中混合系数 $W = 0.9$.

6　实验研究

6.1　PG法计算互信息的精度

对于文献[5]提出的抑或问题(eXclusive OR,XOR),数据如表1所示.类别标号属性 C 的熵 $H(C) = 1$,且 C 可以由 X_1、X_2 完全确定,则理论上应有 $I(C;X_1X_2) = H(C) = 1$.

表1　XOR问题数据表

Table 1　XOR problem

C	X_1	X_2
−1	0	0
1	0	1
1	1	0
−1	1	1

若将 X_1、X_2 看成离散属性,按照CD/ST法(属性全为离散属性时,PG法退化为ST法)计算,则 $I(C;X_1X_2) = 1$.

若将 X_1、X_2 看成连续属性,按照PW/QMI-P/PG法计算,则 $I(C;X_1X_2)$ 的值与窗宽有关.当窗宽较小时 $I(C;X_1X_2) \approx 1$;当窗宽增大时样本间相互干扰变大,互信息逐渐减小至0,具体如图1.对于XOR问题,PG法在计算精度上优于PW法[5]、QMI-P法[7].值得指出的是QMI-P法即使在 h 接近于0时,计算出的二次型互信息并不接近于1,而是收敛到0.6931,这也印证二次型互信息(QMI)与shannon互信息是不同的.本文阐述的是shannon互信息的计算方法,因而下文主要与PW法进行比较.

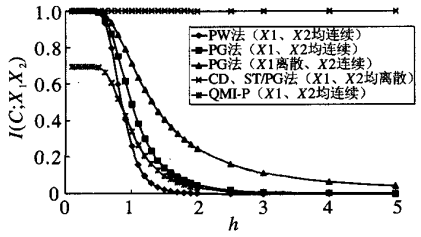

图 1 $I(C;X1X2)$ 随窗宽的变化图

Fig. 1 Estimated $I(C;X1X2)$ for different h

若将 X_1、X_2 分别看作离散、连续属性，用 PG 法计算互信息（PW 法、QMI - P 法不能处理混合属性），则其计算精度比将 X_1、X_2 均看作连续属性时高. 实际上，对于处理混合属性，精度高已然不那么重要，能够同时处理离散、连续属性这一能力本身才是更重要的.

本实验验证 PG 法的正确性，展示 PG 法同时处理离散、连续属性的能力. 同时也说明合理设定窗宽 h 的重要性. 窗宽可按文献[13]取如下形式：

$$h = \left(\frac{4}{2d+1}\right)^{1/(d+4)} n^{-1/(d+4)}, \tag{22}$$

其中，d 为连续属性个数，n 为样本个数.

为了进一步说明 PG 法的合理性，考虑如下问题：设 X/Y 是满足均值为 μ_x/μ_y，方差为 σ_x^2/σ_y^2 的正态分布，即 $X \sim N(\mu_x,\sigma_x^2)$，$Y \sim N(\mu_y,\sigma_y^2)$. 令 $U = X+Y$；$V = X-Y$，则可推导出：

$$I(U;V) = \log_2 \frac{\sigma_x^2+\sigma_y^2}{2\sigma_x\sigma_y}. \tag{23}$$

表 2 I(U;V) 估计值

Table 2 Estimated $I(U;V)$ value

X	Y	样本数 n	CD/ST	PG	$I(U;V)$ 理论值
$N(0,1)$	$N(0,1)$	10	2.0464	0.2202	0
$N(0,1)$	$N(0,1)$	50	1.1902	0.2071	0
$N(0,1)$	$N(0,1)$	100	0.5278	0.1290	0
$N(0,1)$	$N(0,1)$	500	0.1123	0.0482	0
$N(0,1)$	$N(0,1)$	1000	0.0531	0.0584	0
$N(0,1)$	$N(0,1)$	5000	0.0121	0.0213	0
$N(1,5)$	$N(-3,10)$	10	1.9610	0.5532	0.3219
$N(1,5)$	$N(-3,10)$	50	1.1854	0.3854	0.3219
$N(1,5)$	$N(-3,10)$	100	0.8249	0.3810	0.3219
$N(1,5)$	$N(-3,10)$	500	0.4332	0.3769	0.3219
$N(1,5)$	$N(-3,10)$	1000	0.3814	0.3648	0.3219
$N(1,5)$	$N(-3,10)$	5000	0.3002	0.3216	0.3219

下面分别随机产生若干组 U、V，用 CD/ST 法（这两种方法需要将连续属性 U、V 离散为 10 段）及 PG 法估计 $I(U;V)$，得到结果如表 2 所示.

PW 法不能处理两个属性都是连续属性的 I(U;V) 问题. PG 法可以直接计算这种情况下的互信息. 从表 2 可以看出，在相同样本条件下，PG 法较 CD/ST 等离散算法有更高的精度，在样本数目较少时优势更为明显.

6.2 互信息估计方法的算法复杂度比较

Led Display 数据产生器可以从 UCI 数据库[14]得到，其类别标号属性 $C = \{0,\cdots,9\}$ 由前 7 个 0/1 属性决定，另外 17 个 0/1 属性为噪声属性. 用 Led Display 数据发生器可以产生任意规模的样本量.

维数扩展性比较. 取 100 个样本，当 X 取前 1 至 20 个属性时，记录各方法计算互信息 $I(C;X)$ 所需时间（参见图 2）. 观察图 2 发现：高维时 CD 法计算时间随维数呈指数增长；ST、PG、PW 法大致呈线性增长，且 ST 法斜率显著小于 PG、PW 法. 这是因为 CD 法计算时间复杂度为 $O(nK_1 + K_2 P_{\max}^{M+N})$，在维数较高时 $P_{\max}^{M+N} \gg n$，因而高维时计算时间随维数指数增长. PG、PW 法与维数相关的操作为计算 $(\boldsymbol{x}_j - \boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{x}_i)$ 的步骤，它的计算量和维数成正比，导致连续方法计算量基本和维数成正比. ST 法与维数相关的操作主要为排序过程中“按 $\boldsymbol{X}$ 比较”的计算，其计算量不与维数成严格正比（通常只比较前几个属性就能判断大小），因而 ST 法的维数可扩展性很强.

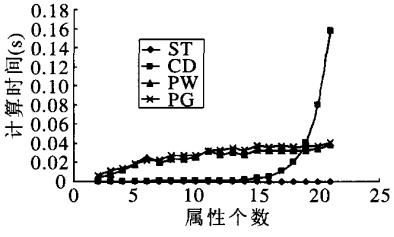

图 2 维数可扩展性比较

Fig. 2 Dimension flexibility of 4 methods

样本可扩展性比较. $\boldsymbol{X}$ 取前 7 个属性，用数据发生器分别产生 200，400，…，3 000 个样本，记录各样本量下 4 种算法计算 $I(C;\boldsymbol{X})$ 所需时间（参见图 3、图 4）. 观察图 3、图 4 发现，CD、ST 法样本可扩展性显著好于 PG、PW 法. 这是因为 CD 法计算时间复杂度为 $O(nk_1 + k_2 P_{\max}^{M+N})$，在维数较高时 $P_{\max}^{M+N} \gg n$，即主要由维数决定而与样本量关系不大；ST 法时间复杂

度为 $O(Dn\log n)$；而 PG、PW 法时间复杂度为 $O(Dn^2)$.

空间扩展性比较. ST、PG、PW 法均不需额外内存,而 CD 法内存需求量随维数指数增长(如图 5). 本例中属性不同值个数很少($P=2$)、计算维数较低(20 维),实际应用中属性不同值可能很多、计算维数可能达到几百、几千甚至更高,这时用 CD 法计算互信息是不现实的.

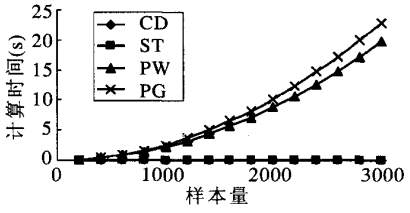

图 3　样本可扩展性比较:4 种算法计算时间

Fig. 3　Sample flexibility of 4 methods

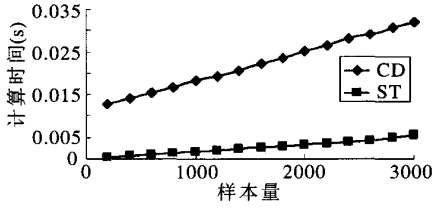

图 4　样本可扩展性比较:CD、ST 法计算时间

Fig. 4　Sample flexibility of CD and ST

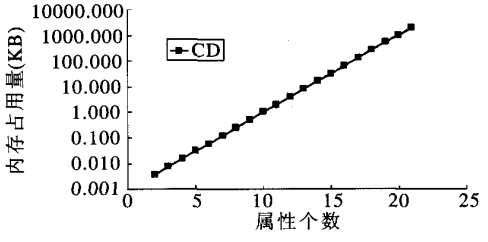

图 5　CD 法内存需求量

Fig. 5　RAM requirement of CD

实验结果表明:ST 法有很好的维数、样本、空间可扩展性,适于计算高维、海量样本的离散型互信息;PG 法有较好的维数、空间可扩展性,样本可扩展性较差,但 PG 法可以同时处理离散、连续属性,适合计算少量样本的连续 / 统一型互信息. ST、PG 法的特性汇总于表 3.

表 3　ST、PG 法特性汇总

Table 3　Comparison between ST and PG

		ST 法	PG 法
时间复杂度	维度可扩展性	$O(D)$	$O(D)$
	样本可扩展性	$O(n\log n)$	$O(n^2)$
	时间复杂度	$O(Dn\log n)$	$O(Dn^2)$
空间复杂度		$O(1)$	$O(1)$
预处理方法		离散属性编码、连续属性离散化	离散属性编码、连续属性标准化
优势		时间、空间复杂度低	能处理连续属性能处理混合属性
不足		需要对连续属性离散化	窗宽 h 确定困难计算复杂度较高
适用场合		海量、高维、离散型互信息	少量、高维、连续 / 统一型互信息

6.3　PG - HMI 特征选择的有效性

实验数据采用 UCI 数据库中 sonar 数据集[14],该数据集有 104 个训练样本,104 个测试样本,每个样本有一个类别标号属性 $C=\{R,M\}$,60 个连续属性.

表 4　实验中涉及的算法

Table 4　Feature selection methods used in experiments

算法简记	算法名称
IG	信息增益法[8]
MIFS	MIFS 法[3]
MIFSU	MIFSU 法[4]
mRMR	离散型最小冗余最大互信息法[6]
CmRMR	连续型最小冗余最大互信息法[6]
PWFS	PWFS 法[5]
OFS-MI	OFS-MI 法[7]
ST-HMI	离散型混合互信息法(将连续属性离散化,用 ST 法算互信息,其余同 PG-HMI)
PG-HMI	统一型混合互信息法

为了验证 PG - HMI 法的有效性,本文将 PGFB 法与 7 种(参见表 4)基于互信息的特征选择方法进行比较. 应用表 4 中算法时,首先要进行适当的数据预处理. 在应用 IG、MIFS、MIFSU、mRMR 和 ST - HMI 法时,将连续属性按文献[3]的方法离散为 5 段:如果属性的分布未知,则计算其均值 μ,标准差 σ,并将区间$[\mu-2\sigma,\ \mu+2\sigma]$等宽分成 5 段,超出最左(右)子区间的样本被分配到最左(右)子区间. 在应用 CmRMR、PWFS、OFS - MI 和 PG - HMI 法时,

将连续属性进行标准化，使属性的均值为0，方差为1. 例如：对连续属性 X 标准化时，首先应计算出 X 的均值 μ、标准差 s，然后对每一个属性值 x_i，令 $x_i = (x_i - \mu)/s$.

为了验证特征选择结果，本文选用 weka[15] 提供的5种分类器（参见表5），实验过程中每种分类器均使用默认参数.

表5 参与比较的分类器

Table 5 Classifiers used in experiments

分类器简记	分类器名称
NB	Naive Bayes，朴素贝叶斯
KNN	K-Nearest Neighbours，k近邻
NN	Neural Network，BP人工神经网络
SMO	Sequential Minimal Optimization algorithm for training a support vector，顺序最小化的支持向量机
C4.5	决策树

实验步骤：首先，对于每种特征选择方法，分别选出3、5、8、10、13、15、18、20个属性（在应用 CmRMR、PWFS、OFS-MI、PG-HMI 法时，Parzen 窗宽取 $h = 0.4611$）. 然后，用表5所列的5种分类器根据不同属性集合分别进行训练. 最后，用测试集验证分类准确率.

表6 PG-HMI 法对于 sonar 数据特征选择结果的分类准确率

Table 6 Accuracy of PG-HMI for sonar dataset

属性个数	NB	KNN	NN	SMO	C4.5
3	70.19	75.00	68.27	69.23	69.23
5	71.15	87.50	74.04	71.15	75.00
8	71.15	91.35	77.88	76.92	71.15
10	75.00	89.42	75.96	76.92	78.85
13	75.00	92.31	71.15	75.96	75.96
15	75.00	89.42	78.85	75.96	75.96
18	76.92	90.38	81.73	75.00	75.96
20	75.96	92.31	75.00	75.96	75.96
分类器平均准确率	73.80	88.46	75.36	74.64	74.76

以连续型 PG-HMI 法为例，分类准确率结果汇总于表6. 所有特征选择方法的分类器平均准确率（不同属性个数时特定分类器分类准确率的均值）和综合准确率（5个分类器平均准确率的均值）汇总于表7.

依据表7，ST-HMI 法和 PG-HMI 法综合准确率较高，说明这两种方法的有效性.

表7 各特征选择算法对于 sonar 数据分类器平均准确率和综合准确率

Table 7 Accuracy comparison of feature selection methods for sonar dataset

算法	NB	KNN	NN	SMO	C4.5	综合
IG	73.08	82.69	76.08	75.72	73.56	76.23
MIFS	65.14	81.49	77.88	75.84	**75.84**	75.24
MIFSU	70.31	83.89	**81.49**	**79.21**	69.11	76.80
mRMR	70.43	87.26	79.81	77.52	70.55	77.12
CmRMR	73.56	84.38	77.28	75.00	66.71	75.38
PWFS	69.35	84.98	76.92	72.96	73.68	75.58
OFS-MI	69.83	86.06	75.24	72.60	72.00	75.14
ST-HMI	73.08	84.86	78.61	75.96	73.56	**77.21**
PG-HMI	**73.80**	**88.46**	75.36	74.64	74.76	**77.40**

7 结束语

本文给出一种能够直接处理混合属性的互信息计算的方法——PG 法，并提出能够同时考虑新信息量和相关性的特征选择准则——HMI. 结合两者，本文给出特征选择的 PG-HMI 法，实验结果表明该方法能取得较好的特征选择结果.

PG-HMI 法计算复杂度较高，在处理海量样本时能力不足. 这时可以用“聚类-采样”的方法降低样本量；或者将连续属性离散化，然后采用较为快速的 ST-HMI 法. 但无论那种方法都会一定程度影响特征选择的效果. 如何改进 PG-HMI，使之适于海量样本还需进一步研究.

参 考 文 献

[1] Piramuthu S. Evaluating Feature Selection Methods for Learning in Data Mining Applications // Proc of the 31st Hawaii International Conference on System Sciences. Kohala Coast, USA, 1998, V: 294-301

[2] Han J, Kamber M. Data Mining: Concepts and Techniques. New York, USA: Morgan Kaufman, 2000
(Han J, Kamber M. 数据挖掘：概念与技术. 范 明，孟小峰，等，译. 北京：机械工业出版社，2001)

[3] Battiti R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Trans on Neural Networks, 1994, 5(4): 537-550

[4] Kwak N, Choi C H. Input Feature Selection for Classification Problems. IEEE Trans on Neural Networks, 2002, 13(1): 143-159

[5] Kwak N, Choi C H. Input Feature Selection by Mutual Information Based on Parzen Window. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1667-1671

[6] Peng Hanchuan, Long Fuhui, Ding C. Feature Selection Based

on Mutual Information Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[7] Chow T W S, Huang D. Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information. IEEE Trans on Neural Networks, 2005, 16(1): 213-224

[8] Quinlan J R. C4.5: Programs for Machine Learning. San Mateo, USA: Morgan Kaufmann, 1993

[9] Shannon C E, Weaver W. The Mathematical Theory of Communication. Urbana, USA: University of Illinois Press, 1949

[10] Zhu Xuelong. Fundamentals of Applied Information Theory. Beijing, China: Tsinghua University Press, 2000 (in Chinese)
(朱雪龙. 应用信息论基础. 北京:清华大学出版社, 2000)

[11] Parzen E. On Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics, 1962, 33(3): 1065-1076

[12] Bian Zhaoqi, Zhang Xuegong. Pattern Recognition. Beijing, China: Tsinghua University Press, 2000 (in Chinese)
(边肇祺,张学工. 模式识别. 北京:清华大学出版社, 2000)

[13] Silverman B W. Density Estimation for Statistics and Data Analysis. London, UK: Chapman & Hall, 1986

[14] Hettich S, Bay S D. The UCI KDD Archive[DB/OL]. [1999-09-09]. http://kdd.ics.uci.edu

[15] The University of Wakato. WEKA Software [CP/OL]. [2005-10-20]. http://www.cs.waikato.ac.nz/~ml/weka

PG-HMI:一种基于互信息的特征选择方法

作者： 王皓， 孙宏斌， 张伯明， WANG Hao， SUN Hong-Bin， ZHANG Bo-Ming
作者单位： 清华大学,电机工程与应用电子技术系,电力系统国家重点实验室,北京,100084
刊名： 模式识别与人工智能 ISTIC EI PKU
英文刊名： PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE
年，卷(期)： 2007,20(1)
被引用次数： 1次

参考文献(16条)

1. Piramuthu S Evaluating Feature Selection Methods for Learning in Data Mining Applications/ 1998

2. Han J;Kamber M;范明;孟小峰 数据挖掘:概念与技术 2001

3. Battiti R Using Mutual Information for Selecting Features in Supervised Neural Net Learning[外文期刊] 1994(04)

4. Kwak N;Choi C H input Feature Selection for Classification Problems 2002(01)

5. Kwak N;Choi C H Input Feature Selection by Mutual Infor mation Based on Parzen Window[外文期刊] 2002(12)

6. Peng Hanchuan;Long Fuhui;Ding C Feature Selection Based on Mutual Information Criteria of Max-Dependency,Max-Relevance and Min-Redundancy[外文期刊] 2005(08)

7. Chow T W S;Huang D Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual lnformation[外文期刊] 2005(01)

8. Quinlan J R CA.5:Programs for Machine Learning 1993

9. Shannon C E;Weaver W The Mathematical Theory of Communication 1949

10. 朱雪龙 应用信息论基础 2000

11. Parzen E On Estimation of a Probability Density Function and Mode[外文期刊] 1962(03)

12. 边肇祺;张学工 模式识别 2000

13. Silverman B W Density Estimation for Statistics and Data Analysis 1986

14. Hettich S;Bay S D The UCI KDD Archive 1999

15. The University of Wakato WEKA Software 2005

16. 若某一取值x在样本中不出现,则nx=0,nxlognx=0,对熵H(X)的值没有影响

本文读者也读过(6条)

1. 卢新国.林亚平.陈治平.LU Xin-guo.LIN Ya-ping.CHEN Zhi-ping 一种改进的互信息特征选取预处理算法[期刊论文]-湖南大学学报（自然科学版）2005,32(1)

2. 黄德根.马玉霞.杨元生 基于互信息的中文姓名识别方法[期刊论文]-大连理工大学学报2004,44(5)

3. 谭金波.黄峰.杨晓江.李艺.Tan Jinbo.Huang Feng.Yang Xiaojiang.Li Yi 一种改进的互信息特征选择算法[期刊论文]-情报学报2006,25(6)

4. 高伟.田铮.GAO Wei.TIAN Zheng 基于条件互信息的多维时间序列图模型[期刊论文]-控制理论与应用2008,25(2)

5. 谢文彪.樊绍胜.费洪晓.樊晓平.Xie Wen-biao.Fan Shao-sheng.Fei Hong-xiao.Fan Xiao-ping 基于互信息梯度优化计算的信息判别特征提取[期刊论文]-电子与信息学报2009,31(12)

6. 李顺山.杨明星.庄天戈 基于互信息方法的医学图像检索[期刊论文]-红外与毫米波学报2001,20(5)

引证文献(1条)

1. 令狐红英. 陈梅. 王翰虎. 娄皴 基于互信息可信度的贝叶斯网络入侵检测研究[期刊论文]-计算机工程与设计 2009(14)