约束高斯分类网研究

王双成 [1, 2] 高 瑞 [1, 3] 杜瑞杰 [1]

摘 要 针对基于一元高斯函数估计属性边缘密度的朴素贝叶斯分类器不能有效利用属性之间的依赖信息和使用多元高斯函数估计属性联合密度的完全贝叶斯分类器易于导致对数据的过度拟合而且高阶协方差矩阵的计算也非常困难等情况, 在建立属性联合密度分解与组合定理和属性条件密度计算定理的基础上, 将朴素贝叶斯分类器的属性选择、分类准确性标准和属性父结点的贪婪选择相结合, 进行约束高斯分类网学习与优化, 并依据贝叶斯网络理论, 对贝叶斯衍生分类器中属性为类提供的信息构成进行分析. 使用 UCI 数据库中连续属性分类数据进行实验, 结果显示, 经过优化的约束高斯分类网具有良好的分类准确性.

关键词 连续属性, 高斯网络, 贝叶斯网络, 贝叶斯衍生分类器, 密度估计

引用格式 王双成, 高瑞, 杜瑞杰. 约束高斯分类网研究. 自动化学报, 2015, **41**(12): 2164−2176

DOI 10.16383/j.aas.2015.c150106

Restricted Gaussian Classification Network

WANG Shuang-Cheng[1, 2] GAO Rui[1, 3] DU Rui-Jie[1]

Abstract Naive Bayes classifier using unitary Gaussian function to estimate attribute marginal density can not effectively use dependency information between attributes, and the full Bayes classifier using multivariate Gaussian function to estimate attribute joint density often leads to over fitting and the difficulty to calculate high-order covariance matrix. In this paper, based on Gaussian network theory, a decomposition and combination theorem for attribute joint density and a calculation theorem for attribute conditional density are established. A restricted Gaussian classification network is presented by combining the attribute selection of naive Bayes classifier, the evaluation criteria of classification accuracy and the greedy search of attribute parent nodes. The information composition of attributes provided for class is analyzed in Bayesian derivative classifiers. Experiment and analysis are done by using data sets with continuous attributes in UCI. The results show that restricted Gaussian classification networks have very good classification accuracy.

Key words Continuous attributes, Gaussian networks, Bayesian networks, Bayesian derivative classifiers, density estimation

Citation Wang Shuang-Cheng, Gao Rui, Du Rui-Jie. Restricted Gaussian classification network. *Acta Automatica Sinica*, 2015, **41**(12): 2164−2176

由贝叶斯分类器衍生的分类器 (简称为贝叶斯衍生分类器)[1−2] 已成为分类器家族的重要组成部分, 这些分类器各具特色, 可分别满足不同的实际需求, 有着广阔的应用前景. 对贝叶斯衍生分类器已有许多研究, 但主要是针对离散属性的情况, 如 Friedman 等[3] 的 TAN (Tree augmented naive Bayes) 分类器, Jing 等[4] 对 TAN 分类器的属性选择和参数集成, 王中锋等[5] 基于条件对数似然函数导数标准的贝叶斯网络分类器优化, Webb 等[6] 对朴素贝叶斯分类器的各阶依赖扩展研究, Wang 等[2] 建立的离散属性约束贝叶斯分类网, 冯月进等[7] 给出的限定的贝叶斯网络分类器, López-Cruz 等[8] 将经过结构优化的贝叶斯网络分类器用于神经元分类器等. 对于贝叶斯衍生分类器中的连续属性, 可采用两种处理方法: 一种是将其离散化, 最终转化为离散属性的分类器问题; 另一种是不离散化连续属性, 但需要估计属性密度. 两种方法各有优势和不足[1], 本文研究第二种方法, 其研究成果也可扩展到混合属性的情况. 关于贝叶斯衍生分类器的属性密度估计, 目前主要采用高斯函数和高斯核函数方法. John 等[9] 分别使用高斯函数和高斯核函数估计属性边缘密度建立 Gaussian naive Bayes classifier 和

收稿日期 2015-03-11 录用日期 2015-09-06
Manuscript received March 11, 2015; accepted September 6, 2015
国家自然科学基金 (61272209), 上海市自然科学基金 (15ZR1429700), 上海市教委科研创新项目 (15ZZ099) 资助
Supported by National Natural Science Foundation of China (61272209), Natural Science Foundation of Shanghai (15ZR1429700), and Innovation Program of Shanghai Municipal Education Commission (15ZZ099)
本文责任编委 杨健
Recommended by Associate Editor YANG Jian
1. 上海立信会计学院数学与信息学院 上海 201620 2. 上海立信会计学院立信会计研究院 上海 201620 3. 上海财经大学统计与管理学院 上海 200433
1. School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 201620 2. Lixin Accounting Research Institute, Shanghai Lixin University of Commerce, Shanghai 201620 3. School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433

Flexible Bayes classifier 两种朴素贝叶斯分类器, 虽然分类效果并不理想 (不能利用属性之间的依赖信息), 但奠定了基于属性密度估计研究贝叶斯衍生分类器的基础. Pérez 等[10−11] 在 John 和 Langley 工作的基础上, 采用高斯函数、高斯核函数和 Filter 方法, 通过连续属性之间的条件互信息计算、比较和选择, 对朴素贝叶斯分类器进行依赖扩展, 使扩展后分类器的分类准确性得到改进. 王双成等[1] 使用多元高斯核函数建立了连续属性完全贝叶斯分类器, 并通过平滑参数的调整来优化分类器, 使经过优化的完全贝叶斯分类器具有良好的分类准确性. Bounhas 等[12]、He 等[13]、Dong 等[14] 和 Pavani 等[15] 基于高斯函数估计属性联合密度建立贝叶斯分类器, 夏战国等[16] 和 Liu 等[17] 利用高斯过程进行分类, 这些分类器均需要进行复杂的多元高斯函数计算, 因此会影响它们的效率、可靠性和适用范围.

使用高斯核函数和高斯函数估计属性密度各有特点. 一元高斯核函数能够估计复杂的属性边缘密度, 但抗噪声能力相对较弱, 而且学习与分类的效率低; 采用边缘高斯函数的乘积进行叠加的多元高斯核函数 (一般的多元高斯核函数具有过多的平滑参数, 不适合于贝叶斯衍生分类器中的属性密度估计), 能够直接估计属性联合密度, 不需要计算协方差矩阵, 但用于叠加的边缘高斯函数的乘积蕴含着条件独立的假设, 因此通过叠加也只能部分地利用属性之间的依赖信息, 而且也同样存在效率问题. 基于一元高斯函数估计属性边缘密度的朴素贝叶斯分类器具有高效率和良好的抗噪声性, 但可能导致对数据的欠拟合, 从而降低分类器的分类准确性; 直接使用多元高斯函数估计属性联合密度的完全贝叶斯分类器能够充分利用属性之间的依赖信息, 但当属性较多时, 协方差矩阵的计算非常困难. 我们通过属性联合密度的分解计算和分解方式的优化来解决这些问题, 实现高斯朴素贝叶斯分类器与高斯完全贝叶斯分类器的优势互补.

本文的主要贡献如下:

1) 连续属性贝叶斯衍生分类器研究的核心问题是属性联合密度计算. 在高斯网络理论的基础上, 给出了联合高斯密度的分解与组合定理, 以及条件高斯密度计算定理, 为属性联合密度的分解计算提供理论依据和实现方法, 可将高阶属性联合密度计算转化为低阶 (一般不超过 4 阶) 属性条件密度计算, 从而使属性联合密度估计具有较高的效率, 计算结果也更加可靠.

2) 结合朴素贝叶斯分类器的属性子集选择、分类准确性标准和贪婪属性父结点搜索方法, 建立了一种新的高斯网络分类器–约束高斯分类网, 通过结构优化可使分类器以局部最优的方式利用属性为类提供的信息, 从而能够提高分类器与数据的拟合程度, 并使用 UCI 数据库[18] 中连续属性分类数据对分类器学习方法的有效性和分类可靠性进行了验证.

3) 依据贝叶斯网络和高斯网络理论, 分离出属性为类提供的三种依赖信息, 它们分别是传递依赖信息、直接导出依赖信息和间接导出依赖信息, 以此为依据论述了朴素贝叶斯分类器依赖扩展的必要性, 并给出了根据实际需求可供选择的一些依赖扩展方案.

本文分为五个部分: 第 1 节论述连续属性贝叶斯分类器和它的衍生分类器; 第 2 节在高斯网络理论的基础上, 给出高斯联合密度的分解与组合定理, 以及属性条件密度计算定理, 以为建立约束高斯分类网提供理论依据; 第 3 节给出约束高斯分类网的学习方法和算法, 并对属性为类提供的信息构成进行分析, 为朴素贝叶斯分类器依赖扩展的必要性和扩展方法优化选择提供支持; 第 4 节是实验与分析; 最后是结论和进一步的工作.

1 贝叶斯衍生分类器

用 $X_1,\cdots,X_n,C$ 表示连续属性和类, $x_1,\cdots,x_n$, c 为其值, $\boldsymbol{X}=(X_1,\cdots,X_n)$, $\boldsymbol{x}=(x_1,\cdots,x_n)$, D 是具有 N 个记录的数据集, 数据随机产生于混合分布 P. 使用满条件概率 $p(c|x_1,\cdots,x_n)$ 进行分类的分类器 $\arg\max_{c(x_1,\cdots,x_n)}\{p(c|x_1,\cdots,x_n)\}$ 一般称为贝叶斯分类器. 贝叶斯分类器是一个依据概率理论进行分类的基础分类器, 在理论上我们能够证明它是最优分类器[1−2], 但直接计算满条件概率 $p(c|x_1,\cdots,x_n)$ 非常困难, 根据概率公式, 可得:

$$p(c|x_1,\cdots,x_n)=\frac{p(c,x_1,\cdots,x_n)}{f(x_1,\cdots,x_n)}=\frac{p(c)f(x_1,\cdots,x_n|c)}{f(x_1,\cdots,x_n)}=\alpha p(c)f(x_1,\cdots,x_n|c) \tag{1}$$

其中, α 是与 C 无关的量. 式 (1) 将满条件概率计算转化为类先验概率 $p(c)$ 与属性联合密度 $f(x_1,\cdots,x_n|c)$ 的计算问题, 由对属性联合密度计算方式的不同便可产生一系列分类器, 我们将这些分类器统称为贝叶斯衍生分类器, 如图 1 所示.

贝叶斯分类器简记为 BC, 结构见图 2 (a). 我们可以在属性之间条件独立性假设下来估计属性联合密度, 得到朴素贝叶斯分类器 (最简单的贝叶斯衍生分类器 (Naive Bayes classifier, NBC), 结构见图 2 (b)). 条件独立性使得 $f(x_1,\cdots,x_n|c)=\Pi_{i=1}^{n}f(x_i|c)$, 因此, 只需估计属性边缘密度即可, 但是,这样会导致估计得到的属性联合密度与实际联合

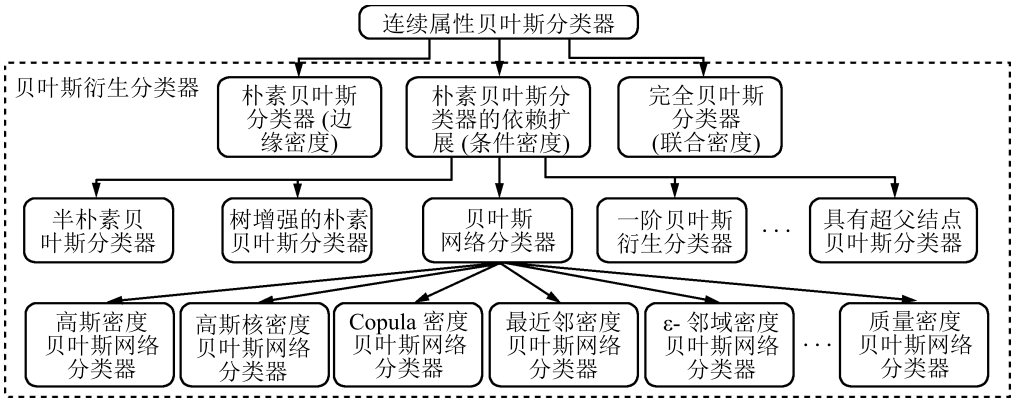


图 1 连续属性贝叶斯衍生分类器

Fig. 1 Bayesian derivative classifiers with continuous attributes

密度有较大的差距, 从而影响分类器的分类准确性 (丢失一些重要的信息). 我们也可以使用多元高斯函数直接估计属性联合密度, 获得完全贝叶斯分类器 (最复杂的贝叶斯衍生分类器 (Full Bayes classifier, FBC), 结构见图 2 (c)). 但直接估计属性联合密度需要计算协方差矩阵, 当属性数量较多时这种计算非常困难, 而且可靠性也往往很难得到保障. 就像基于贝叶斯网络能够进行联合概率的分解一样, 依据高斯网络可实现属性联合密度的分解计算, 并能够避免 NBC 和 FBC 存在的一些弊端.

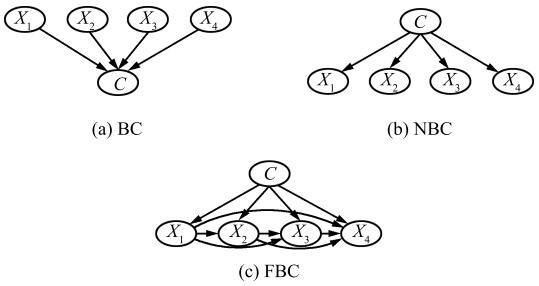


图 2 4 个属性的 BC、NBC 和 FBC 结构

Fig. 2 Structure of BC, NBC and FBC with four attributes

2 高斯网络

高斯网络 (Gaussian network)[19] 是高斯分布与贝叶斯网络 (Bayesian network)[20] 的结合, 是传统基于数据的统计方法和强调知识的人工智能方法的统一, 是研究属性联合密度分解计算的有力工具.

对于连续变量集 $\boldsymbol{X}$, 如果 $\boldsymbol{X}$ 服从联合高斯分布, 那么

$$f(x)=\mathrm{N}(\boldsymbol{\mu},\Sigma^{-1})= (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2}$$

其中, $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)$ 是 $\boldsymbol{X}$ 的无条件均值向量, $\Sigma=(\sigma_{ij})$ 是一个 $n\times n$ 协方差矩阵, $|\Sigma|$ 是 Σ 的行列式, Σ^{-1} 是 Σ 的逆矩阵, 也称为决策矩阵 (Precision matrix), 用 W 表示 Σ^{-1}.

按照链式规则, 可得

$$f(\boldsymbol{x})=\prod_{i=1}^{n}f(x_i|x_1,\cdots,x_{i-1}) \tag{3}$$

根据条件高斯分布理论, 可知

$$f(x_i|x_1,\cdots,x_{i-1})=\mathrm{N}\left(\mu_i+\sum_{j=1}^{i-1}b_{ij}(x_j-\mu_j),\frac{1}{\nu_i}\right) \tag{4}$$

其中, ν_i 是给定 $x_1,\cdots,x_{i-1}$ 时 X_i 的条件方差, b_{ij} 是反映 X_i 和 X_j 之间依赖强度的线性相关系数 (也称为线性回归系数).

Shachter 等[21] 给出了决策矩阵 W、线性相关系数 b_{ij} 和条件方差 ν_i 之间的关系:

$$W(i+1)=\begin{pmatrix} W(i)+\dfrac{\boldsymbol{b}_{i+1}\boldsymbol{b}_{i+1}^{\mathrm{T}}}{\nu_{i+1}} & -\dfrac{\boldsymbol{b}_{i+1}}{\nu_{i+1}} \\ -\dfrac{\boldsymbol{b}_{i+1}^{\mathrm{T}}}{\nu_{i+1}} & \dfrac{1}{\nu_{i+1}} \end{pmatrix} \tag{5}$$

其中, $W(i)$ 是 W 的左上角 $i\times i$ 阶主子式矩阵, $\boldsymbol{b}_i$ 是列向量, $\boldsymbol{b}_i^{\mathrm{T}}=(b_{1,i},\cdots,b_{i-1,i})$ 是 $\boldsymbol{b}_i$ 的转置向量, $W(1)=1/\nu_1$.

在 Shachter 等研究的基础上, Geiger 等[19] 给出了如下的高斯网络定义:

定义 1. 高斯网络是一个二元组 (G,P), 其中 G 是高斯网络结构, 是一个有向无环图, X_j 到 X_i 之间有一条弧当且仅当 $b_{ij}\neq 0$, $j<i$, P 是高斯网络参数, 包括 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)$, $\boldsymbol{\nu}=(\nu_1,\cdots,\nu_n)$ 和 $\boldsymbol{B}=(b_{ij},i<j)$.

根据式 (2)、式 (4) 和定义 1, 可以得到:

$$\prod_{i=1}^{n} f(x_i|x_1,\cdots,x_{i-1}) = \prod_{i=1}^{n} f(x_i|\pi_i, G) \tag{6}$$

其中, π_i 是 G 中 X_i 的父结点集 Π_i 的配置.

定理 1 (分解与组合定理). 如果 $\boldsymbol{x}$ 具有多元高斯密度, 我们能够获得由 $X_1,\cdots,X_n$ 构成的高斯网络, 反之, 在 W 是可逆矩阵的假设下, 由 X_1, $\cdots$, X_n 的高斯网络也可得到关于 $\boldsymbol{x}$ 的多元高斯密度.

证明.

1) 多元高斯密度 $\Rightarrow$ 高斯网络.

设 $\boldsymbol{x}$ 具有多元高斯密度 $f(\boldsymbol{x}) = \mathrm{N}(\bar{\boldsymbol{x}}, \Sigma^{-1})$, 我们需要依据 $\bar{\boldsymbol{x}}$ 和 $W = \Sigma^{-1}$ 来获得 G 和 P, 从而得到高斯网. 根据迭代式 (5) 能够获得 $\boldsymbol{\nu} = (\nu_1,\cdots,\nu_n)$ 和 $\boldsymbol{B} = (b_{ij}, i<j)$. 实际上, 对于 ν_j, b_{ij}, $i<j$, 由关于 $W(j)$ 的公式中矩阵的最后一列 (或最后一行) 可以得到 $\nu_j = 1/w_{j,j}$, $b_{1,j} = \nu_j w_{1,j},\cdots, b_{i,j} = \nu_j w_{i,j},\cdots, b_{j-1,j} = \nu_j w_{j-1,j}$, 从而获得 $\boldsymbol{\nu} = (\nu_1,\cdots,\nu_n)$、$\boldsymbol{B} = (b_{ij}, i<j)$ 和 $\bar{\boldsymbol{x}}$ (网络参数 P), 再根据式 (3) 和 X_j 与 X_i 之间有一条弧当且仅当 $b_{ij} \neq 0$, $j<i$, 又可建立有向无环图 (网络结构 G), 最终得到关于 $X_1,\cdots,X_n$ 的高斯网络 (G,P).

2) 高斯网络 $\Rightarrow$ 多元高斯密度.

设 (G,P) 是一个关于 $X_1,\cdots,X_n$ 的高斯网络, 也就是我们需要由一个网络结构 (有向无环图) G 和网络参数 $\boldsymbol{\mu} = (\mu_1,\cdots,\mu_n)$, $\boldsymbol{\nu} = (\nu_1,\cdots,\nu_n)$ 与 $\boldsymbol{B} = (b_{ij}, i<j)$ 来得到 $\boldsymbol{x}$ 所具有的多元高斯密度 $f(\boldsymbol{x}) = \mathrm{N}(\bar{\boldsymbol{x}}, \Sigma^{-1})$. □

确定 $\boldsymbol{x}$ 所具有的多元高斯密度, 只需求得 $\boldsymbol{\mu}$ 和 Σ 即可. $\boldsymbol{\mu}$ 是已知, 下面来获得 Σ.

根据迭代关系式 (5), 通过迭代我们能够由 $\boldsymbol{\nu}$ 和 $\boldsymbol{B}$ 确定 W. 由 W 的可逆性假设, 对 W 进行求逆运算便得到 Σ, 由 $\boldsymbol{\mu}$ 和 Σ 所确定的多元高斯密度便是 $\boldsymbol{x}$ 所具有的多元高斯密度.

式 (5)、定义 1 和定理 1 将为建立约束高斯分类网提供理论依据.

3　约束高斯分类网

定义 2. 假设 $f(x_1,\cdots,x_n|c)$ 具有多元高斯密度, 那么在类 C 约束下, 由 $X_1,\cdots,X_n$ 构成用于分类的高斯网络称为约束高斯分类网 (Restricted Gaussian classification network, RGCN).

RGCN 是一种特殊的高斯网络, 其结构同样是一个有向无环图, 属性除类之外还可以有属性父结点, 仍用 G 表示 RGCN 的结构, 如图 3 所示.

根据高斯网络理论、式 (6) 和图 3 中所蕴含的的条件独立性关系, 可得

$$p(c|x_1,\cdots,x_n) = \alpha p(c) f(x_1,\cdots,x_n|c) = \alpha p(c)\prod_{i=1}^{n} f(x_i|\pi_i, c, G) \tag{7}$$

其中, $f(x_i|\pi_i, c, G)$ 是属性条件密度, π_i 是 G 中 X_i 的属性父结点集 Π_i 的配置. 那么, RGCN 能够被表示为

$$\arg\max_{c(x_1,\cdots,x_n)} \{p(c|x_1,\cdots,x_n)\} = \arg\max_{c(x_1,\cdots,x_n)} \left\{ p(c)\prod_{i=1}^{n} f(x_i|\pi_i, c, G) \right\} \tag{8}$$

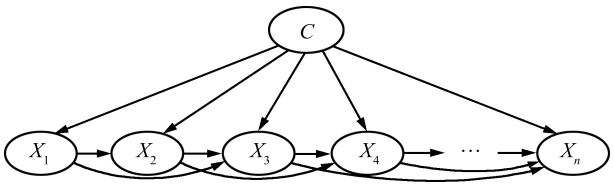


图 3　RGCN 结构

Fig. 3　Structure of RGCN

RGCN 的学习与分类过程如图 4 所示.

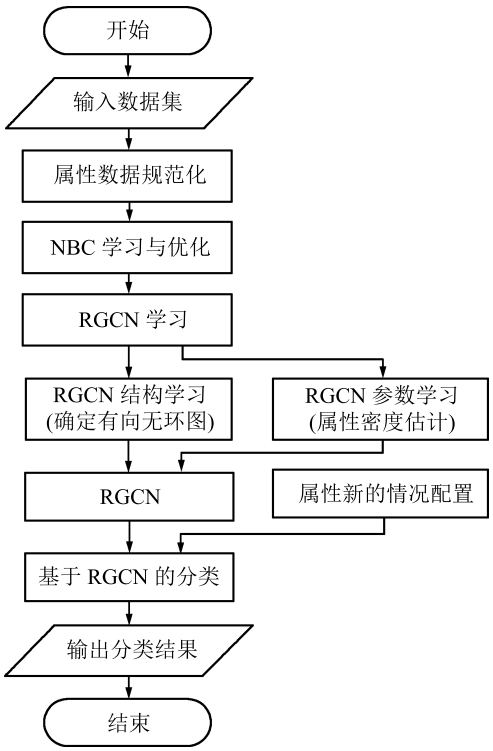


图 4　RGCN 学习与分类流程图

Fig. 4　Flowchart for RGCN learning and classification

从式 (8) 和图 4 中能够发现, RGCN 是依据后验概率的顺序进行分类, 而顺序性比较稳定, 因此 RGCN 具有良好的抗噪声性. RGCN 基于高斯网

络理论分解属性联合密度, 这也使分类计算具有更高的效率和可靠性. 建立 RGCN 需要解决两个问题, 分别是属性条件密度估计和分类器结构学习.

3.1 属性条件密度估计

引理 1. 设随机向量

$$\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}$$

服从均值向量, 协方差矩阵分别为

$$\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\quad \Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

的高斯分布, 记

$$\Sigma^{-1}=\begin{pmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{pmatrix}$$

其中, $\boldsymbol{X}_1$ 为 k_1 维向量, $\boldsymbol{X}_2$ 为 k_2 维向量, 且 $k_1+k_2=k$, 则有

$$\Sigma_{22}^{-1}=T_{22}-T_{12}T_{11}^{-1}T_{21} \tag{9}$$

$$T_{11}^{-1}T_{12}=-\Sigma_{12}\Sigma_{22}^{-1} \tag{10}$$

$$T_{11}^{-1}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{11}$$

证明. 由于 $\Sigma\Sigma^{-1}=I_{k\times k}$, 即

$$\Sigma_{11}T_{11}+\Sigma_{12}T_{21}=I_{k_1\times k_1} \tag{12}$$

$$\Sigma_{11}T_{12}+\Sigma_{12}T_{22}=O_{k_1\times k_2} \tag{13}$$

$$\Sigma_{21}T_{11}+\Sigma_{22}T_{21}=O_{k_2\times k_1} \tag{14}$$

$$\Sigma_{21}T_{12}+\Sigma_{22}T_{22}=I_{k_2\times k_2} \tag{15}$$

由式 (14) 可得 $\Sigma_{21}^{-1}=-\Sigma_{22}T_{21}T_{11}^{-1}$, 结合式 (15) 得到 $\Sigma_{22}(T_{22}-T_{12}T_{11}^{-1}T_{21})=I$, 从而式 (9) 得证.

由式 (14) 可得 $T_{21}=-\Sigma_{22}^{-1}T_{21}T_{11}$, 结合式 (12) 得到 $T_{11}(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})=I$, 从而式 (11) 得证.

对式 (14) 两边求转置, 由于 $\Sigma_{21}^{\mathrm{T}}=\Sigma_{12}$, $T_{21}^{\mathrm{T}}=T_{12}$, 可得 $T_{11}^{\mathrm{T}}\Sigma_{12}+T_{12}\Sigma_{22}=O_{k_1\times k_2}$. 该式两边同时乘以 $(T_{11}^{\mathrm{T}})^{-1}$ 得到 $T_{11}^{-1}T_{12}\Sigma_{22}+\Sigma_{12}=O_{k_1\times k_2}$, 从而式 (10) 得证. □

引理 2. 引理 1 的条件, 用 $f(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 表示 $\boldsymbol{X}$ 的联合密度函数, $f(\boldsymbol{x}_1|\boldsymbol{x}_2)$ 表示给定 $\boldsymbol{X}_2=\boldsymbol{x}_2$ 时的条件下 $\boldsymbol{X}_1$ 的条件密度函数, 则有

$$f(\boldsymbol{x}_1|\boldsymbol{x}_2)=(2\pi)^{-\frac{k_1}{2}}\left(\frac{|\Sigma|}{|\Sigma_{22}|}\right)^{-\frac{1}{2}}\times$$

$$\exp[-\frac{1}{2}(\boldsymbol{x}_1-\boldsymbol{v}_1)^{\mathrm{T}}(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})(\boldsymbol{x}_1-\boldsymbol{v}_1)]$$

其中, $\boldsymbol{v}_1=\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)$.

证明. $f(\boldsymbol{x}_1,\boldsymbol{x}_2)$ 可以表示为

$$f(\boldsymbol{x}_1,\boldsymbol{x}_2)=(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\times$$

$$\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

由于 $\boldsymbol{X}_2$ 也服从高斯分布, 用 $f(\boldsymbol{x}_2)$ 表示 $\boldsymbol{X}_2$ 的边缘密度函数, 则有

$$f(\boldsymbol{x}_2)=(2\pi)^{-\frac{k_2}{2}}|\Sigma_{22}|^{-\frac{1}{2}}\times$$

$$\exp\left\{-\frac{1}{2}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\right\}$$

又因为

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=\\&\quad(\boldsymbol{x}_1^{\mathrm{T}}-\boldsymbol{\mu}_1^{\mathrm{T}},\boldsymbol{x}_2^{\mathrm{T}}-\boldsymbol{\mu}_2^{\mathrm{T}})\begin{pmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{pmatrix}\times\\&\quad(\boldsymbol{x}_1-\boldsymbol{\mu}_1,\boldsymbol{x}_2-\boldsymbol{\mu}_2)=\\&\quad(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}T_{11}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+\\&\quad(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)+\\&\quad(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}T_{21}(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+\\&\quad(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}T_{22}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

重新组织上式, 可以得到

$$\begin{aligned}&(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=\\&\quad[(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+T_{11}^{-1}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)]^{\mathrm{T}}T_{11}\times\\&\quad[(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+T_{11}^{-1}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)]+\\&\quad(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}(T_{22}-T_{12}T_{11}^{-1}T_{21})(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

记

$$\begin{aligned}Q_1&=[(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+T_{11}^{-1}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)]^{\mathrm{T}}T_{11}\times\\&\quad[(\boldsymbol{x}_1-\boldsymbol{\mu}_1)+T_{11}^{-1}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)]\\Q_2&=(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}(T_{22}-T_{12}T_{11}^{-1}T_{21})(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

因此

$$\begin{aligned}Q_1&=(\boldsymbol{x}_1-\boldsymbol{v}_1)^{\mathrm{T}}T_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{v}_1)=\\&\quad(\boldsymbol{x}_1-\boldsymbol{v}_1)^{\mathrm{T}}(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})(\boldsymbol{x}_1-\boldsymbol{v}_1)\end{aligned}$$

其中

$$\begin{aligned}\boldsymbol{v}_1&=\boldsymbol{\mu}_1-T_{11}^{-1}T_{12}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)=\\&\quad\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

结合引理 1, 可得

$$f(\boldsymbol{x}_1,\boldsymbol{x}_2)=(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{Q_1+Q_2}{2}\right)$$

$$f(\boldsymbol{x}_2)=(2\pi)^{-\frac{k_2}{2}}|\Sigma_{22}|^{-\frac{1}{2}}\exp\left(-\frac{Q_2}{2}\right)$$

则

$$f(\boldsymbol{x}_1|\boldsymbol{x}_2)=\frac{f(\boldsymbol{x}_1,\boldsymbol{x}_2)}{f(\boldsymbol{x}_2)}=$$
$$(2\pi)^{-\frac{k_1}{2}}(\frac{|\Sigma|}{|\Sigma_{22}|})^{-\frac{1}{2}}\exp\left(-\frac{Q_1}{2}\right)$$

□

定理 2 (条件密度计算定理). 假设对给定的 $C=c$, $\boldsymbol{X}$ 服从均值向量为 $\boldsymbol{\mu}$, 协方差矩阵为 Σ 的高斯分布, 用 $\Pi_i=\{X_i^1,\cdots,X_i^{\omega(i)}\}$ 表示属性变量 X_i 的属性父结点集, $\omega(i)$ 表示 X_i 的属性父结点的个数, 那么

$$p(x_i|\pi_i,c,G)=\mathrm{N}(m_i(c),v_i(c))$$

$$m_i(c)=\mu_i(c)+\sum_{j=1}^{\omega(i)}\frac{\sigma_{ij}(c)}{\sigma_j^2(c)}\left(x_i^j-\mu_i^j(c)\right)\tag{16}$$

$$v_i(c)=\frac{|\Sigma_{X_i,\Pi_i}(c)|}{|\Sigma_{\Pi_i}(c)|}\tag{17}$$

其中, π_i 是 Π_i 的配置, $\mu_i(c)$, $\sigma_i^2(c)$, $\sigma_{ij}(c)$ 和 $\Sigma_R(c)$ 分别表示给定类值 $C=c$ 条件下, X_i 的均值、方差、与其属性父结点 X_i^j $(j=1,\cdots,\omega(i))$ 之间的协方差和属性变量集 R 的协方差矩阵.

证明. 令

$$\boldsymbol{X}=\begin{pmatrix}X_i\\ \Pi_i(c)\end{pmatrix},\quad \boldsymbol{\mu}=\begin{pmatrix}\mu_i(c)\\ \mu_{\Pi_i}(c)\end{pmatrix},$$
$$\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix}$$

那么

$$\Sigma=\begin{pmatrix}\sigma_i^2(c)&(\sigma_{ij}(c))_{1\times\omega(i)}\\ (\sigma_{ji}(c))_{\omega(i)\times1}&(\sigma_j^2(c))_{\omega(i)\times\omega(i)}\end{pmatrix}$$

且 $\Sigma_{22}^{-1}=(1/\sigma_j^2(c))_{\omega(i)\times\omega(i)}$, 结合引理 1 和引理 2 得到

$$m_i(c)=\mu_i(c)+\sum_{j=1}^{\omega(i)}\frac{\sigma_{ij}(c)}{\sigma_j^2(c)}\left(x_i^j-\mu_i^j(c)\right)$$

$$v_i(c)=\frac{|\Sigma|}{|\Sigma_{22}|}=\frac{|\Sigma_{X_i,\Pi_i}(c)|}{|\Sigma_{\Pi_i}(c)|}$$

□

定理 2 将为属性条件密度估计提供实现方法. 根据定理 1 和定理 2, 约束高斯分类网可表示为

$$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}f(x_i|\pi_i,c,G)\right\}=$$
$$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}\mathrm{N}\left(m_i(c),\frac{1}{\nu_i(c)}\right)\right\}=$$
$$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}\mathrm{N}\Bigg(\mu_i(c)+$$
$$\sum_{j=1}^{\omega(i)}b_{ij}(c)(x_j-\mu_i(c)),\frac{|\Sigma_{(\Pi_i,X_i)(c)}|}{|\Sigma_{\Pi_i(c)}|}\Bigg)\right\}\tag{18}$$

属性具有过多的父结点, 一方面易于导致对数据的过度拟合, 另一方面会降低学习与分类的效率, 一般可限定属性除类之外的父结点不超过 3 个 (或 2 个). 这样通过属性联合密度的高斯网络分解, RGCN 的分类时间复杂度是 $\mathrm{O}(n)$ (以最多 3 个父结点的情况为例, 需要进行不超过 $2(n-1)$ 次最多 4 阶的协方差矩阵运算), 而不进行属性联合密度分解计算的时间复杂度是 $\mathrm{O}(n^3)$ (需要进行 n 阶协方差矩阵运算).

3.2 RGCN 结构学习

RGCN 结构学习是在 NBC 的基础上发现属性父结点的过程. Geiger 等给出的高斯网络结构打分函数非常复杂, 而且所突出的是变量之间的相关性, 而不是属性和属性之间依赖关系对分类的贡献, 因此不适合于建立 RGCN 的结构. 我们结合分类准确性标准 (采用 10 折交叉有效性验证方法[22] 估计分类器的分类准确率) 和贪婪搜索方法分别进行 NBC 的属性子集选择和在此基础上的属性新父结点发现.

1) 属性父结点选择

分别用 Π_i 和 Ω_i 表示 X_i 的属性父结点集和子孙结点集, 我们能够证明 X_i 新的属性父结点候选集为 $\{X_1,\cdots,X_n\}-\Omega_i\cup\Pi_i\cup\{X_i\}$. 根据 Quinlan[23] 的信息增益率为所选择的属性排序, 其中的属性条件密度计算采用高斯函数. 按照属性的顺序依次发现属性的新父结点, 建立 RGCN 结构. 在属性父结点选择过程中, 首先进行一阶父结点选择 (为属性增加的第一个父结点), 在一阶父结点选择的基础上进行二阶父结点选择 (为属性增加的第二个父结点), 在二阶父结点选择过程中也可能新增一阶父结点, 在二阶父结点选择的基础上进行三阶父结点选择 (为属性增加的第三个父结点), 在三阶父结点选择过程中也可能新增一阶和二阶父结点. 我们通过实验发现具有三阶以上父结点的属性极其少见, 而

且对分类的贡献也非常小, 因此只进行到三阶父结点选择.

用 S 表示 NBC 的结构, A 是经过选择的属性子集, 为表示方便, 令 $A=\{X_1,\cdots,X_m\}$, $m\le n$, $accuracy\_nbc(A|D,S)$ 表示在属性子集 A 上 NBC 的分类准确率, $accuracy\_bnc\_f(j|D,G_j)$, $j=1,2,3$ 和 $accuracy\_bnc\_b(j|D,G_j)$ 表示进行第 j 阶依赖扩展过程中父结点变化前和变化后的分类准确率, 其中, G_j 是对应的 RGCN 结构, $\Pi_i^{(1)}$、$\Pi_i^{(2)}$ 和 $\Pi_i^{(3)}$ 分别表示 X_i 的一阶、二阶和三阶属性父结点集, Γ_i 是属性 X_i 的父结点候选集 (可以包含 $\boldsymbol{X}-A$ 中的属性结点). 具体的属性父结点选择如算法 1 所示.

算法 1. 属性父结点选择算法

输入. 数据集 D, 属性子集 A 和 $accuracy\_nbc(A|D,S)$

设置 $\Pi_i^{(1)}=\Pi_i^{(2)}=\Pi_i^{(3)}=\phi$, $\Gamma_i=\{X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_n\}$

for $j=1,2,3$

if $j=1$

$accuracy\_bnc\_f(j|D,G_1)=accuracy\_nbc(A|D,G_1)$

用 S 初始化 G_1

else

$accuracy\_bnc\_f(j|D,G_1)=accuracy\_bnc\_b(j-1|D,G_j)$

用 G_{j-1} 初始化 G_j

for $i=1,\cdots,m$

计算 Γ_i, 让 $s=0$

for $k=1,\cdots,n$

if $X_k\in\Gamma_i$

计算 $accuracy\_bnc\_b(j|D,G_j)$

if $accuracy\_bnc\_b(j|D,G_j)>accuracy\_bnc\_f(j|D,G_j)$

$accuracy\_bnc\_f(j|D,G_j)=accuracy\_bnc\_b(j|D,G_j)$

$s=k$

if $s\neq 0$

if $\Pi_i^{(1)}=\phi$

$\Pi_i^{(1)}=\{X_s\}$, 更新 Γ_i 和 G_1

if $\Pi_i^{(1)}\neq\phi$ and $\Pi_i^{(2)}=\phi$

$\Pi_i^{(2)}=\{X_s\}$, 更新 Γ_i 和 G_2

if $\Pi_i^{(1)}\neq\phi$ and $\Pi_i^{(2)}\neq\phi$ and $\Pi_i^{(3)}=\phi$

$\Pi_i^{(3)}=\{X_s\}$, 更新 Γ_i 和 G_3

if $j>1$ and $accuracy\_bnc\_b(j|D,G_j)=accuracy\_bnc\_b(j-1|D,G_j)$

退出 j 循环

输出. $accuracy\_bnc\_b(j|D,G_j)$ $(j=1,2,3)$, G_j 和 G.

2) 时间复杂性分析

RGCN 学习主要包括属性子集选择和属性父结点选择. 在属性子集选择中, 最多需要进行 $(n-1)(n+2)/2$ 次分类准确率计算. 而在属性父结点选择中, 一阶、二阶和三阶依赖扩展的属性父结点选择最多需要进行 $3n^2$ 次分类准确率计算, 可知, RGCN 学习算法关于分类准确率运算的时间复杂度是 $\mathrm{O}(n^2)$. 计算分类准确率的主要运算是属性条件密度估计, 计算一次分类准确率最多需要进行 $10n$ 次属性条件密度估计, 因此, RGCN 学习算法关于属性条件密度估计运算的时间复杂度是 $\mathrm{O}(n^3)$.

3.3 属性为类提供的信息构成分析

根据贝叶斯网络和高斯网络理论, 在 RGCN 中, 属性可为类提供三种信息, 分别是传递依赖信息、直接导出依赖信息和间接导出依赖信息. 传递依赖信息是属性直接为类提供的信息, 是分类的主要信息; 直接导出依赖信息是由属性和类之间形成 V 结构[24] 直接诱导出的信息, 对分类也有较大的影响, 有时这种影响甚至是至关重要的; 间接导出依赖信息是由属性和属性之间形成的 V 结构间接诱导出的信息, 这种信息相对较弱, 在分类时起到辅助和补充的作用, 但有时这种信息也不容忽视, 具体情况如图 5 所示.

我们再从分类器的表示形式来分析属性为类提供的三种信息. 考察 RGCN 的表示形式

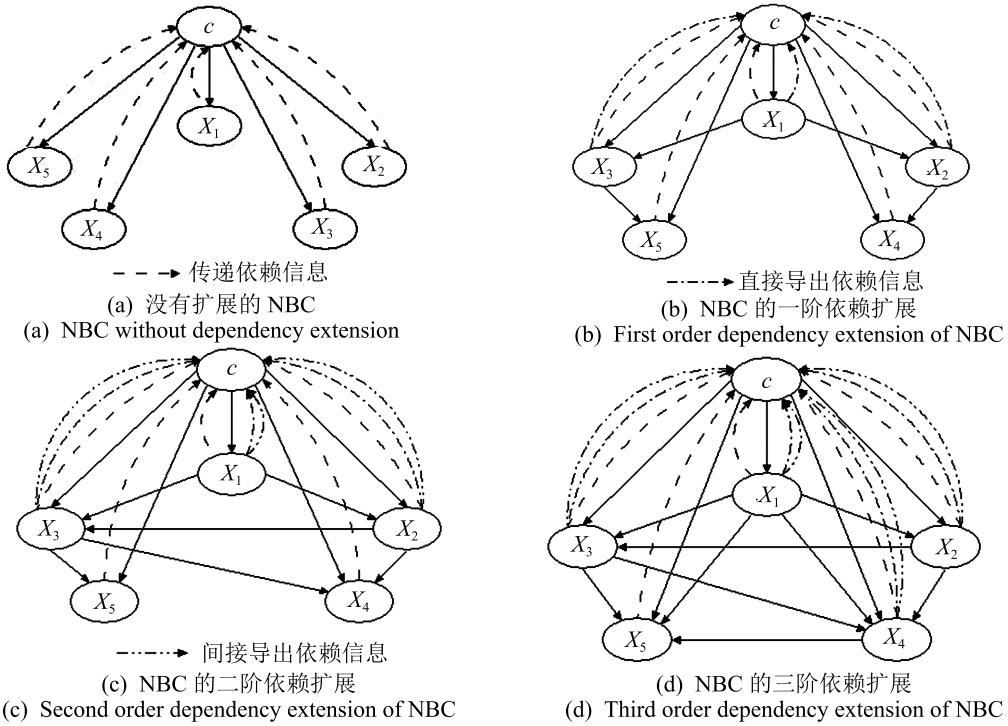
$$\arg\max_{c(x_1,\cdots,x_n)}\left\{p(c)\prod_{i=1}^{n}p(x_i|x_i^1,\cdots,x_i^{\omega(i)},c,G)\right\}$$

其中, X_i 为 C 提供传递依赖信息, $X_i^1,\cdots,X_i^{\omega(i)}$ 为 C 提供直接导出依赖信息, $X_i^1,\cdots,X_i^{\omega(i)}$ 之间的相互作用产生间接导出依赖信息 (或称为结合信息).

NBC (图 5 (a)) 虽然具有高效率, 但它的属性只能为类提供传递依赖信息, 从而会影响分类器的分类准确性. 对 NBC 的一阶依赖扩展 (图 5 (b)), 将使直接导出依赖信息得到有效的利用, 因此能够提高分类器的分类准确性. NBC 的二阶或更高阶 (包括完全依赖扩展) 的依赖扩展 (图 5 (c) 和图 5 (d)), 则可提取三种依赖信息用于分类, 分类器会更加可靠, 但会降低效率. 我们可以根据具体实际情况和需求确定依赖扩展策略, 建立相应的贝叶斯衍生分类器.

4 实验与分析

在 UCI 数据库中选择 30 个连续属性的分类数据集, 如表 1 所示, 删除具有丢失数据的记录, 数据集中记录的位置也进行随机初始化. 分别从分类器之间的分类错误率 (分类错误率 = 1 − 分类准确率/100) 比较、朴素贝叶斯分类器依赖扩展情况比较和属性所提供的三种信息对分类的贡献比例三个方面进行实验与分析.



(a) 没有扩展的 NBC
(a) NBC without dependency extension

(b) NBC 的一阶依赖扩展
(b) First order dependency extension of NBC

(c) NBC 的二阶依赖扩展
(c) Second order dependency extension of NBC

(d) NBC 的三阶依赖扩展
(d) Third order dependency extension of NBC

图 5　贝叶斯衍生分类器中属性为类提供的信息

Fig. 5　Information of attribute providing for class in Bayesian derivative classifiers

表 1　用于实验的数据集情况

Table 1　Data sets for experiments

序号	数据集	例子	属性	类值	序号	数据集	例子	属性	类值
1	Ae_train	774	12	9	16	Horse_colic	300	22	2
2	Allbp	747	20	4	17	Image_Segmentation	209	16	7
3	Annealing	798	22	4	18	Ionosphere	349	33	2
4	Arabic_digit	736	13	2	18	Iris	150	4	3
5	Automobile	205	22	5	20	Magic_Gamma_telescope	718	10	2
6	Banques	102	12	4	21	New_thyroid	215	5	3
7	Breast_cancer	699	10	2	22	Pima	768	8	2
8	Breast_tissue	106	9	6	23	Spambase	601	30	2
9	Cmc	1 376	9	2	24	Spectf_incorrect	269	44	2
10	Column_3C	310	6	3	25	Steel_plates_faults	441	33	2
11	Credit (crx)	690	15	2	25	Thyroid0387	787	27	4
12	Echocardiogram	132	10	3	27	Transfusion	748	4	2
13	Ecoli	292	5	4	28	Wdpc	569	31	2
14	Flags	214	9	6	29	Wine	178	13	3
15	Heart_disease	270	13	2	30	Wpbc	198	34	2

4.1　分类错误率比较

选择 12 个分类器, 其中前 5 个是离散属性分类器 (对连续属性采用 Fayyad 等[25] 方法进行离散化), 后 7 个是连续属性分类器, 将它们与 RGCN 进行分类错误率比较实验, 分类器的具体情况是: 离散属性朴素贝叶斯分类器 (用 DNB 表示); 基于 Quinlan 的信息增益率为属性排序, 按照属性顺序对朴

素贝叶斯分类器进行链依赖扩展的离散属性分类器 (用 DCEB 表示); Friedman 等给出的树增强朴素贝叶斯分类器 (用 DTAN 表示); 采用属性排序和贪婪打分–搜索方法所建立的约束贝叶斯网络分类器 (用 DCBN 表示), 其中属性排序采用 Quinlan 的信息增益率, 打分函数使用 MDL (Minimal description length) 标准; 结合分类准确性标准和前向贪婪搜索而建立的离散属性完全贝叶斯分类器 (用 DFB 表示); 使用高斯函数估计属性条件边缘密度而建立的朴素贝叶斯分类器 (用 GNB 表示); 使用高斯函数来估计属性的条件密度, 并基于 Quinlan 的信息增益率为属性排序, 对 GNB 进行链依赖扩展而得到的分类器 (用 CCEN 表示); 基于高斯函数来估计属性的条件密度和属性之间的条件互信息计算, 以属性之间的条件互信息为权重, 采用 Friedman 等方法进行连续属性的最大权重跨度树学习, 对 GNB 进行树依赖扩展而得到的分类器 (用 CTAN 表示); 基于高斯函数估计属性联合密度, 结合分类准确性标准和前向贪婪搜索而建立的连续属性完全贝叶斯分类器 (用 GFB 表示); 使用多元高斯核函数估计属性联合密度, 并进行单平滑参数优化的连续属性完全贝叶斯分类器 (用 GKFB 表示); Quinlan 的决策树分类器 (C4.5); 支持向量机分类器 (Support vector machine, SVM); 约束高斯分类网 (RGCN).

采用 10 折交叉有效性验证方法进行分类器的分类错误率估计, 并使用 Wilcoxon Signed-ranks Test 和 Friedman Test with Post-hoc Bonferroni Test (Demšar 2006)[26] 进行两个分类器分类错误率之间差异的置信打分, 其中, "$*$" 表示 RGCN 和用于比较的分类器相对于给定的检验方法差别显著. 12 个分类器与 RGCN 的分类错误率实验结果如表 2 所示.

在表 2 中, 使用 Wilcoxon Signed-ranks Test 和 Friedman Test with Post-hoc Bonferroni Test 的检验结果显示, RGCN 与 12 个分类器在分类错误率方面差异显著. 再考察总体平均值, RGCN 优于其他分类器的程度依次是 8.65 %, 6.07 %, 10.11 %, 7.39 %, 4.34 %, 14.32 %, 12.51 %, 11.71 %, 8.86 %, 6.53 %, 8.20 % 和 3.05 %, 可见 RGCN 具有优势的程度也非常明显.

RGCN 与其他分类器在 30 个数据集的分类错误率具体比较情况如图 6 所示. 图中每一个点的坐标是用于比较的两个分类器的分类错误率, 在 45 度线上方和下方的点分别表示 RGCN 的分类错误率小于和大于用于比较的分类器.

从图 6 中的 12 个散点图来看, RGCN 的分类错误率明显优于其他分类器. 在 30 个数据集中, RGCN 相对于其他分类器, 分类准确率之差大于或等于 0.5 % (用区间 $[0.5, \infty)$ 表示)、大于 -0.5 % 且小于 0.5 % (用区间 $(-0.5, 0.5)$ 表示)、小于或等于 -0.5 % (用区间 $(-\infty, -0.5]$ 表示), 三方面的百分比情况如表 3 所示.

综合分类器之间的分类错误率差异的显著性检验、分类错误率平均值比较和分类准确率差异百分比计算三方面的结果, 显示了 RGCN 相对于其他 12 个分类器在分类错误率 (或分类准确率) 方面具有明显的优势.

高斯函数被广泛用于各领域的密度估计 (是一种最有效实用的密度估计函数), 但直接基于高斯函数估计属性密度存在两个主要问题: 一个是可能导致对数据的欠拟合; 另一个是多元高斯函数的计算非常困难. 本文通过分类器结构的调整来提高分类器与数据的拟合程度 (见属性父结点选择算法), 将属性联合高斯密度的贝叶斯网络分解和条件高斯函数计算相结合来解决运算复杂性问题, 这样便能够有效地提高 RGCN 的效率和可靠性, 也是 RGCN 相对于其他 12 个分类器在分类准确性方面具有优势的主要原因.

4.2 朴素贝叶斯分类器依赖扩展情况比较

分别用 GNB、CCEN、CTAN、GFB 和 RGCN 表示相应分类器的分类准确率, 对 GNB 进行依赖扩展后分类准确率的变化情况如图 7 所示.

图 7 显示, RGCN-GNB 曲线基本是在零线和其他三条曲线的上方, 而且另外三条曲线均在零线上下波动. 再考察对生成曲线的数据, 分别从最大值、最小值、平均值、大于零数量、小于零数量和等于零数量的统计结果为:

1) CCEN-GNB: 21.14, −22.50, 1.17, 18, 11, 1;

2) CTAN-GNB: 18.18, −9.00, 1.70, 22, 8, 0;

3) GFB-GNB: 32.01, −12.00, 3.64, 17, 13, 0;

4) RGCN-GNB: 34.00, 0.74, 10.40, 30, 0, 0.

从以上比较结果可以看出, 四种情况的平均值都大于零, 而且大于零数据集的数量明显多于小于或等于零的数量, 但其他三种情况小于零的数量也占有较大的比例, 而且最坏情况降低分类准确率的程度也比较大. 由此可以得出结论: 对 GNB 进行依赖扩展可以使属性之间的条件依赖信息得到有效的利用, 提高分类器的分类准确率, 但依赖扩展方法的选择非常重要, 不同的依赖扩展方法之间可能存在较大的差异, 而且不适宜的依赖扩展方法也会显著降低分类器的分类准确率. RGCN 使所有数据集的 GNB 分类准确率都有所改进, 但提高程度的差别较大, 最大的情况达到 33.26, 这主要是由属性之间条件依赖情况的差异所导致.

表 2 RGCN 和其他分类器的分类错误率实验结果

Table 2 Experimental results on classification error for RGCN and other classifiers

数据集	DNB	DCEB	DTAN	DCBN	DFB	GNB	CCEN	CTAN	GFB	GKFB	C4.5	SVM	RGCN
Ae_train	0.2650	0.2818	0.2624	0.2598	0.3325	0.1935	0.1962	0.1532	0.1325	0.0858	0.2076	0.0562	0.08804
Allbp	0.3286	0.4202	0.4352	0.4203	0.4244	0.5945	0.4595	0.6095	0.3621	0.4567	0.4418	0.3439	0.3311
Annealing	0.1102	0.0949	0.1279	0.0608	0.0760	0.3127	0.1773	0.2254	0.1773	0.1000	0.0853	0.0611	0.0589
Arabic_Digit	0.2740	0.2602	0.2795	0.2357	0.2905	0.3494	0.3494	0.2726	0.2616	0.2220	0.3233	0.2327	0.2215
Automobile	0.3750	0.3050	0.3350	0.3200	0.2750	0.4400	0.4450	0.4200	0.5600	0.3700	0.2633	0.2710	0.2700
Banques	0.2800	0.4500	0.4400	0.5400	0.3700	0.3900	0.3500	0.4800	0.4300	0.4100	0.5000	0.3900	0.2807
Breast_cancer	0.0358	0.0371	0.0486	0.0406	0.0348	0.0372	0.0272	0.0500	0.0343	0.0377	0.0601	0.0431	0.0286
Breast_tissue	0.2900	0.2600	0.3200	0.3400	0.2800	0.3700	0.5600	0.3100	0.4300	0.3000	0.3491	0.3225	0.2400
Cmc	0.3285	0.3124	0.3271	0.3358	0.3176	0.3504	0.3475	0.3270	0.3014	0.3212	0.3176	0.3080	0.3044
Column_3C	0.2259	0.1709	0.1871	0.1904	0.1517	0.1936	0.3033	0.1678	0.2194	0.2355	0.1936	0.1355	0.1484
Credit(crx)	0.1406	0.1347	0.1334	0.1363	0.14492	0.2493	0.2363	0.2334	0.1885	0.2463	0.1406	0.1783	0.1354
Echocardiogram	0.6385	0.3230	0.5154	0.3539	0.3384	0.4385	0.3539	0.4626	0.3461	0.3384	0.4470	0.3754	0.3000
Ecoli	0.1035	0.0931	0.1000	0.0794	0.0725	0.0794	0.0587	0.0689	0.0896	0.0621	0.0994	0.0515	0.0587
Flags	0.4158	0.4736	0.5790	0.4736	0.3895	0.5527	0.4422	0.5316	0.5264	0.4263	0.4805	0.4092	0.3264
Heart_disease	0.1334	0.1370	0.1445	0.1519	0.14077	0.1408	0.1667	0.1555	0.1630	0.1630	0.2260	0.1719	0.1334
Horse_colic	0.2834	0.2433	0.2834	0.3133	0.24337	0.3400	0.3433	0.3600	0.3300	0.2967	0.3133	0.3100	0.2700
Image_Segmentation	0.1450	0.1250	0.1350	0.2700	0.1100	0.2300	0.1650	0.2050	0.3000	0.1300	0.1145	0.1172	0.0900
Ionosphere	0.2572	0.2314	0.2858	0.2943	0.2171	0.5372	0.3258	0.5200	0.2171	0.3629	0.2349	0.2209	0.1972
Iris	0.0400	0.3333	0.0400	0.0467	0.0200	0.0400	0.0467	0.0267	0.0200	0.0333	0.0467	0.0400	0.0200
Magic_Gamma_telescope	0.2522	0.2522	0.2817	0.2930	0.2578	0.3676	0.2862	0.3198	0.2677	0.2508	0.1894	0.1907	0.2170
New_thyroid	0.0454	0.0318	0.0454	0.0381	0.0286	0.0500	0.0864	0.0545	0.0590	0.0410	0.0931	0.0562	0.0410
Pima	0.2390	0.2285	0.2356	0.2250	0.2211	0.2481	0.2403	0.2441	0.2572	0.2403	0.2657	0.2329	0.2221
Spambase	0.1684	0.1383	0.1417	0.1466	0.1267	0.1616	0.3866	0.1566	0.1750	0.2417	0.1610	0.1665	0.1134
Spectf_incorrect	0.2000	0.1423	0.2539	0.1308	0.2076	0.3270	0.2077	0.2770	0.2076	0.2076	0.2677	0.2364	0.2000
Steel_plates_faults	0.2387	0.2000	0.2728	0.2091	0.1432	0.0160	0.1955	0.0091	0.0590	0.0204	0.0069	0.0000	0.0069
Thyroid0387	0.4142	0.4128	0.4321	0.4257	0.4065	0.5243	0.4988	0.5167	0.4706	0.4935	0.4447	0.4180	0.3667
Transfusion	0.2414	0.1946	0.2040	0.2014	0.2095	0.2507	0.2347	0.3160	0.2520	0.2240	0.2300	0.2207	0.2010
Wdbc	0.1412	0.0357	0.1295	0.0463	0.0286	0.0661	0.0572	0.0589	0.0821	0.0412	0.0739	0.0250	0.0233
Wine	0.0447	0.0117	0.0375	0.0530	0.0412	0.0295	0.0412	0.0177	0.0117	0.0322	0.1012	0.0176	0.0059
Wpbc	0.4200	0.1850	0.3650	0.1737	0.2300	0.3350	0.2750	0.1532	0.1900	0.2300	0.3031	0.2287	0.1950
平均	0.2359	0.2173	0.2460	0.2269	0.2043	0.2738	0.2621	0.2568	0.2374	0.2207	0.2327	0.1944	0.1698
Wilcoxon SR Test	$*-4.72$	$*-3.71$	$*-4.76$	$*-4.06$	$*-4.14$	$*-4.78$	$*-4.74$	$*-4.53$	$*-4.67$	$*-4.72$	$*-4.55$	$*-3.58$	RGCN
Friedman/ Bonferroni Test	$*-4.50$	$*-2.37$	$*-5.17$	$*-4.13$	$*-2.28$	$*-5.63$	$*-5.57$	$*-5.42$	$*-4.75$	$*-4.45$	$*-5.35$	$*-2.66$	RGCN

表 3 RGCN 和其他分类器的分类准确率差异比较 (%)

Table 3 Difference comparison on classification accuracy for RGCN and other classifiers (%)

差异情况	DNB	DCEB	DTAN	DCBN	DFB	GNB	CCEN	CTAN	GFB	GKFB	C4.5	SVM
$[0.5, \infty)$	83.33	76.66	90.00	76.66	83.33	100.00	93.33	93.34	90.00	86.67	90.00	70.00
$(-0.5, 0.5)$	16.67	6.67	10.00	16.67	10.00	0.00	6.67	3.33	6.67	13.33	3.33	13.33
$(-\infty, -0.5]$	0.00	16.67	0.00	6.67	6.67	0.00	0.00	3.33	3.33	0.00	6.67	16.67

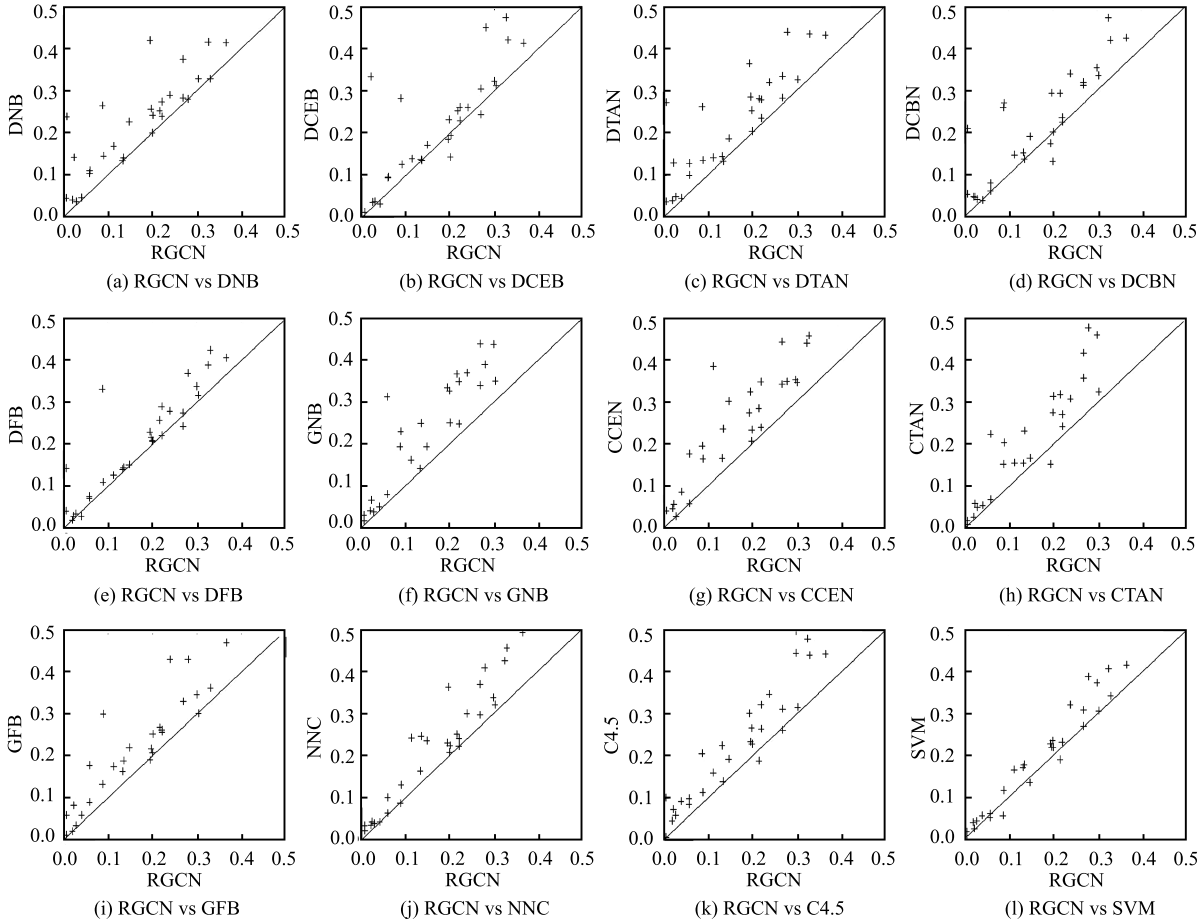


图 6 RGCN 和其他分类器的分类错误率散点图

Fig. 6 Scatter plot graphs of classification errors

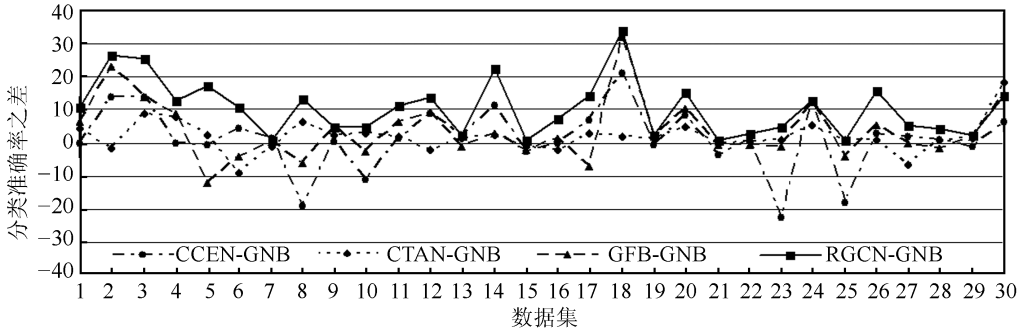


图 7 对 GNB 进行依赖扩展后分类准确率的变化情况图

Fig. 7 Change of classification accuracy when GNB is extended

4.3 属性所提供的三种信息对分类的贡献比例

分别用 I1、I2 和 I3 表示属性提供的传递依赖、直接导出依赖和间接导出依赖信息对分类的贡献, 它们的比例如图 8 所示.

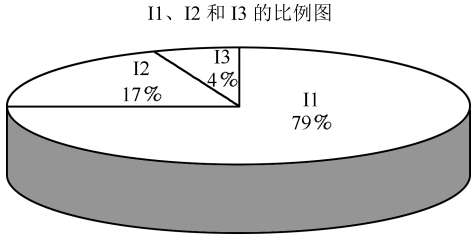


图 8 分类贡献比例图

Fig. 8 Proportion graph of classification contribution

朴素贝叶斯分类器中的属性只为类提供传递依赖信息 (I1, 大约占 79%), 不能提供导出依赖信息 (I2 和 I3, 大约占 21%), 因此需要对朴素贝叶斯分类器进行依赖扩展. 通过属性依赖扩展可使导出依赖信息得到有效的利用, 从而能够提高分类器的分类准确率. 对朴素贝叶斯分类器的一阶依赖扩展, 可使直接导出依赖信息 (I2) 得到利用, 这部分信息也占有较大的比例. 如果需要突出效率, 可只进行一阶扩展; 二阶和以上阶的依赖扩展能够利用三种依赖信息, 但高阶依赖扩展会显著降低学习效率, 综合考虑效率和分类准确率, 二阶依赖扩展应该是比较好的选择, 这些结论同样适合于离散属性贝叶斯衍生分类器.

5 结论

本文在高斯网络理论的基础上, 建立了用于分类的 RGCN. RGCN 中的属性联合密度分解计算, 即可避免复杂的高阶协方差矩阵计算问题, 又能够使属性之间的依赖信息得到充分的利用. 而结合分类准确性标准和属性父结点贪婪选择会以局部最优的方式构建分类器结构, 这将有助于提高分类器与数据的拟合程度. 对贝叶斯衍生分类器中属性为类提供的信息构成分析和使用 UCI 数据库中连续属性分类数据的实验, 分别理论论证和实验验证了对朴素贝叶斯分类器进行依赖扩展的必要性和扩展方法的有效性. 我们进一步的研究工作是采用不同的属性密度估计方法进行贝叶斯网络分类器学习、优化和比较.

References

1 Wang Shuang-Cheng, Du Rui-Jie, Liu Ying. The learning and optimization of full Bayes classifiers with continuous attributes. *Chinese Journal of Computers*, 2012, **35**(10): 2129−2138
(王双成, 杜瑞杰, 刘颖. 连续属性完全贝叶斯分类器的学习与优化. 计算机学报, 2012, **35**(10): 2129−2138)

2 Wang S C, Xu G L, Du R J. Restricted Bayesian classification networks. *Science China Information Sciences*, 2013, **56**(7): 1−15

3 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine Learning*, 1997, **29**(2−3): 131−163

4 Jing Y S, Pavlović V, Rehg J M. Boosted Bayesian network classifiers. *Machine Learning*, 2008, **73**(2): 155−184

5 Wang Zhong-Feng, Wang Zhi-Hai. An optimization algorithm of Bayesian network classifiers by derivatives of conditional log likelihood. *Chinese Journal of Computers*, 2012, **35**(2): 364−374
(王中锋, 王志海. 基于条件对数似然函数导数的贝叶斯网络分类器优化算法. 计算机学报, 2012, **35**(2): 364−374)

6 Webb G I, Boughton J R, Zheng F, Ting K M, Salem H. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. *Machine Learning*, 2012, **86**(2): 233−272

7 Feng Yue-Jin, Zhang Feng-Bin. Max-relevance min-redundancy restrictive BAN classifier learning algorithm. *Journal of Chongqing University*, 2014, **37**(6): 71−77
(冯月进, 张凤斌. 最大相关最小冗余限定性贝叶斯网络分类器学习算法. 重庆大学学报, 2014, **37**(6): 71−77)

8 López-Cruz P L, Larrañaga P, DeFelipe J, Bielza C. Bayesian network modeling of the consensus between experts: an application to neuron classification. *International Journal of Approximate Reasoning*, 2014, **55**(1): 3−22

9 John G H, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. 338−345

10 Pérez A, Larrañaga P, Inza I. Supervised classification with conditional Gaussian networks: increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, 2006, **43**(1): 1−25

11 Pérez A, Larrañaga P, Inza I. Bayesian classifiers based on kernel density estimation: flexible classifiers. *International Journal of Approximate Reasoning*, 2009, **50**(2): 341−362

12 Bounhas M, Mellouli K, Prade H, Serrurier M. Possibilistic classifiers for numerical data. *Soft Computing*, 2013, **17**(5): 733−751

13 He Y L, Wang R, Kwong S, Wang X Z. Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis. *Information Sciences*, 2014, **259**(2): 252−268

14 Dong W Y, Zhou M C. Gaussian classifier-based evolutionary strategy for multimodal optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(6): 1200−1216

15 Pavani S K, Gomez D D, Frangi A F. Gaussian weak classifiers based on co-occurring Haar-like features for face detection. *Pattern Analysis and Applications*, 2014, **17**(2): 431−439

16 Xia Zhan-Guo, Xia Shi-Xiong, Cai Shi-Yu, Wan Ling. Semi-supervised Gaussian process classification algorithm addressing the class imbalance. *Journal on Communications*, 2013, **34**(5): 42−51
(夏战国, 夏士雄, 蔡世玉, 万玲. 类不均衡的半监督高斯过程分类算法. 通信学报, 2013, **34**(5): 42−51)

17 Liu G Q, Wu J X, Zhou S P. Probabilistic classifiers with a generalized Gaussian scale mixture prior. *Pattern Recognition*, 2013, **46**(1): 332−345

18 Murphy S L, Aha D W. UCI repository of machine learning databases [Online], available: http://archive.ics.uci.edu/ml/datasets.html, January 1, 2014

19 Geiger D, Heckerman D. Learning Gaussian Networks. Technical Report MSR-TR-94-10, Microsoft Research, Advanced Technology Division, 1994.

20 Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California, USA: Morgan Kaufmann Publishers Inc., 1988. 383−408

21 Shachter R D, Kenley C R. Gaussian influence diagrams. *Management Science*, 1989, **35**(5): 527−550

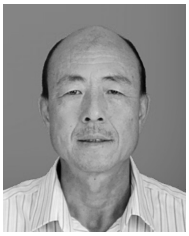
22 Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. 1137−1143

23 Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, **1**(1): 81−106

24 Wang Shuang-Cheng, Yuan Sen-Miao. Research on learning Bayesian networks structure with missing data. *Journal of Software*, 2004, **15**(7): 1042−1048
(王双成, 苑森淼. 具有丢失数据的贝叶斯网络结构学习研究. 软件学报, 2004, **15**(7): 1042−1048)

25 Fayyad U M, Irani K B. Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambery, France, 1993. 1022−1027

26 Demšar J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 2006, **7**(1): 1−30



王双成 上海立信会计学院数学与信息学院教授. 2004 年获吉林大学通信工程学院博士学位. 主要研究方向为人工智能, 机器学习和数据挖掘. 本文通信作者. E-mail: wangsc@lixin.edu.cn
(**WANG Shuang-Cheng** Professor at the School of Mathematics and Information, Shanghai Lixin University of Commerce. He received his Ph. D. degree from the College of Communication Engineering, Jilin University in 2004. His research interest covers artificial intelligence, machine learning, and data mining. Corresponding author of this paper.)

高 瑞 上海财经大学数学与统计学院博士研究生. 上海立信会计学院数学与信息学院讲师. 2006 年获上海理工大学理学院硕士学位. 主要研究方向为机器学习和数据挖掘.
E-mail: gaorui@lixin.edu.cn
(**GAO Rui** Ph. D. candidate at the School of Statistics and Management, Shanghai University of Finance and Economics. She is also a lecturer at the School of Mathematics and Information, Shanghai Lixin University of Commerce. She received her master degree from the College of Science, University of Shanghai for Science and Technology in 2006. Her research interest covers machine learning and data mining.)

杜瑞杰 博士, 上海立信会计学院数学与信息学院讲师. 主要研究方向为机器学习和数据挖掘.
E-mail: duruijie@lixin.edu.cn
(**DU Rui-Jie** Ph. D., lecturer at the School of Mathematics and Information, Shanghai Lixin University of Commerce. Her research interest covers machine learning and data mining.)