文章编号: 1001-0920 (2015) 12-2280-05 DOI: 10.13195/j.kzyjc.2014.1650

基于高斯核函数的朴素贝叶斯分类器依赖扩展

王双成[1a,1b],高 瑞[1a,2], 杜瑞杰[1a]

(1. 上海立信会计学院 a. 数学与信息学院，b. 立信会计研究院，上海 201620；
2. 上海财经大学 统计与管理学院，上海 200433)

摘 要: 朴素贝叶斯分类器不能有效地利用属性之间的依赖信息, 而目前所进行的依赖扩展更强调效率, 使扩展后分类器的分类准确性还有待提高. 针对以上问题, 在使用具有平滑参数的高斯核函数估计属性密度的基础上, 结合分类器的分类准确性标准和属性父结点的贪婪选择, 进行朴素贝叶斯分类器的网络依赖扩展. 使用 UCI 中的连续属性分类数据进行实验, 结果显示网络依赖扩展后的分类器具有良好的分类准确性.

关键词: 朴素贝叶斯分类器；高斯核函数；贝叶斯网络；分类准确性；依赖扩展

中图分类号: TP181 **文献标志码:** A

Dependency extension of naive Bayesian classifiers based on Gaussian kernel function

WANG Shuang-cheng[1a,1b], *GAO Rui*[1a,2], *DU Rui-jie*[1a]

(1a. School of Mathematics & Information，1b. Lixin Accounting Research Institute，Shanghai Lixin University of Commerce，Shanghai 201620，China；2. School of Statistics and Management，Shanghai University of Finance and Economics，Shanghai 200433，China. Correspondent：WANG Shuang-cheng，E-mail：wangsc@lixin.edu.cn)

Abstract: The naive Bayesian classifier can not effectively use the dependency information between attributes. At present, the efficiency of dependency extension is emphasized, which makes the classification accuracy of the extended classifier need to be improved. By using Gaussian kernel function with a smoothing parameter to estimate attribute density, the classification accuracy criterion and the greedy parent node selection of attributes are combined to extend the naive Bayesian classifier. An experiment is done by using data sets in UCI. The results show that extended classifiers have good classification accuracy.

Keywords: naive Bayesian classifiers；Gaussian kernel function；Bayesian network；classification accuracy；dependency extension

0 引 言

朴素贝叶斯分类器 (NBC) 是最简单的贝叶斯网络分类器[1], 具有高效率和较好的分类准确性, 已在许多领域得到了广泛的应用. 但是, 这种分类器受条件独立性的约束, 使属性之间的依赖信息无法得到有效的利用. 围绕这一问题, 研究者们展开了一系列的 NBC 依赖扩展研究. 早期可追溯到 Chow 等[2]所建立的依赖树分类器; Friedman 等[3]在 Chow 等算法的基础上给出了著名的树增强型贝叶斯分类器 (TAN); Jing 等[4]以分类准确性为标准, 对 TAN 分类器进行属性选择和参数集成; Wang 等[1]建立了约束贝叶斯分类网. 这些经过依赖扩展后的分类器均能有效地提高分类准确性, 但所针对的是离散属性的分类器. 关于连续属性的分类器, 可以通过连续属性的离散化, 将其转化为离散属性的情况, 但这样可能导致信息丢失和引入噪声, 从而降低分类器的可靠性. 本文研究不离散化连续属性的 NBC, 并假设所有的属性都是连续的, 其研究结果可推广到混合属性的情况. 处理连续属性的核心是属性密度估计, John 等[5]分别使用高斯函数和高斯核函数估计属性的边缘密度建立了 NBC 和柔性贝叶斯分类器 (FBC), 虽然它们的分类效果并不理想 (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各种贝叶斯分类器的基础; Pérez 等[6-7]在 John 等工作的基础上, 通过连续属性之间的条件互信息计算、排序和选择, 对 NBC 进行依赖扩展, 使分类器的分类准确性有所改进; 夏战

收稿日期: 2014-10-29；**修回日期:** 2015-04-08.

基金项目: 上海市自然科学基金项目(15ZR1429700)；上海市教委科研创新项目(15ZZ099).

作者简介: 王双成(1958−), 男, 教授, 博士, 从事人工智能、机器学习、数据采掘及其应用等研究；高瑞(1980−), 女, 博士生, 从事应用统计与数据采掘的研究.

国等[8]将高斯过程用于具有不均衡类的半监督分类器学习, 取得了较好的分类效果; Bounhas 等[9]和 He 等[10]分别建立了基于高斯函数与高斯核函数估计属性密度的贝叶斯分类器.

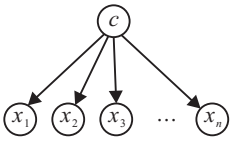
本文采用引入平滑参数的高斯核函数来估计属性密度, 结合使用均方差(MISE)设置平滑参数、属性排序、分类准确性标准和属性父结点的贪婪搜索, 建立扩展的朴素贝叶斯分类器(ENBC), 并利用 UCI[11]中连续属性分类数据集进行实验与分析, 验证了依赖扩展的必要性和扩展方法的有效性.

1　NBC 的依赖扩展

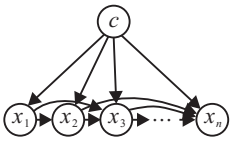
对 NBC 进行依赖扩展, 即属性除类之外还可以具有属性父结点, 依赖扩展的目的是使属性之间的条件依赖信息得到有效的利用. 用 $X_1, X_2, \cdots, X_n$ 和 C 表示连续属性和类, $x_1, x_2, \cdots, x_n$ 和 c 表示其值, D 表示具有 N 个记录的数据集, 数据随机产生于概率分布 P, x_{im} 和 c_m 表示 $X_i(1 \leqslant i \leqslant n)$ 和 C 在数据集 D 中第 $m(1 \leqslant m \leqslant N)$ 个记录的观测值.

1.1　分类器结构和表示形式

NBC 具有星形结构(用 S 表示), ENBC 一般不再具有星形结构(用 G 表示), 两种分类器的结构如图 1 所示.



(a) NBC



(b) ENBC

图 1　NBC 和 ENBC 的结构

依据概率公式、贝叶斯网络理论和图 1(b) 中变量之间的依赖关系, 可以得到

$$p(c|x_1, x_2, \cdots, x_n) = \frac{p(c)f(x_1, x_2, \cdots, x_n|c)}{f(x_1, x_2, \cdots, x_n)} = \alpha p(c)\prod_{i=1}^{n} f(x_i|\pi_i, c, G). \quad (1)$$

其中: α 是与 C 无关的量, $p(c)$ 是类先验概率, $f(x_i|\pi_i, c, G)$ 是属性条件密度, π_i 是在 G 中 X_i 的属性父结点集 Π_i 的配置.

依据分解式 (1) 可以得到 ENBC 的表示形式为

$$\arg\max_{c(x_1, x_2, \cdots, x_n)} \Big\{p(c)\prod_{i=1}^{n} f(x_i|\pi_i, c, G)\Big\}. \quad (2)$$

从分类器 (2) 可以看出, 建立 ENBC 需要解决两个主要问题, 一个是估计 $f(x_i|\pi_i, c, G)$ (属性条件密度估计), 另一个是如何确定 Π_i (分类器结构学习).

1.2　属性条件密度估计

为了使所估计的属性条件密度更好地拟合数据, 采用多元核函数方法. 用 $\hat{p}(x_1, x_2, \cdots, x_n|c, D)$ 表示 $p(x_1, x_2, \cdots, x_n|c)$ 的估计, 基于数据集 D 的多元核函数估计一般形式为

$$\hat{p}(x_1, x_2, \cdots, x_n|c, D) = \frac{1}{N(c)h_1\cdots h_n}\sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\prod_{i=1}^{n} K_i\Big(\frac{x_i - x_{im}}{h_i}\Big)\Big]. \quad (3)$$

其中: $N(c)$ 是数据集 D 中 $c_m = c$ 的情况数量; 而

$$\mathrm{sign}(c_m) = \begin{cases} 1, & c_m = c; \\ 0, & c_m \neq c \end{cases}$$

是示性函数; $K_i(\cdot)$ 是 X_i 的核函数; $h_i(i = 1, 2, \cdots, n)$ 是平滑参数.

选择 $K_i(\cdot)$ 为高斯函数, 为了简单起见, 取 $h = h_1 = \cdots = h_n$, 则

$$K_i\Big(\frac{x_i - x_{im}}{h_i}\Big) = \frac{1}{\sqrt{2\pi}h}\exp\Big[-\frac{(x_i - x_{im})^2}{2h^2}\Big].$$

可以得到

$$\hat{p}(x_1, x_2, \cdots, x_n|c, D) = \frac{\sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}h}\exp\Big[-\frac{(x_i - x_{im})^2}{2h^2}\Big]\Big]}{N(c)h^n} = \frac{\sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\prod_{i=1}^{n}\exp\Big[-\frac{(x_i - x_{im})^2}{2h^2}\Big]\Big]}{N(c)(\sqrt{2\pi}h^2)^n}. \quad (4)$$

设 $\Pi_i = \{X_i^1, X_i^2, \cdots, X_i^{t(i)}\}$, $t(i)$ 是 X_i 父结点的个数, 用 $\hat{p}(x_i|\pi_i, c, D, G)$ 表示 $p(x_i|\pi_i, c, G)$ 的估计, 则有

$$\hat{p}(x_i|\pi_i, c, D, G) = \frac{\hat{p}(x_i, \pi_i|c, D, G)}{\hat{p}(\pi_i|c, D, G)} = \frac{\frac{1}{N(c)(\sqrt{2\pi}h^2)^{t(i)+1}}}{\frac{1}{N(c)(\sqrt{2\pi}h^2)^{t(i)}}}\times \sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\exp\Big[-\frac{(x_i - x_{im})^2}{2h^2}\Big]\times \prod_{s=1}^{t(i)}\exp\Big[-\frac{(x_i^s - x_{im}^s)^2}{2h^2}\Big]\Big]\Big/ \sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\prod_{s=1}^{t(i)}\exp\Big[-\frac{(x_i^s - x_{im}^s)^2}{2h^2}\Big]\Big] =$$

$$\sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\exp\Big[-\frac{(x_i-x_{im})^2}{2h^2}\Big]\times \prod_{s=1}^{t(i)}\exp\Big[-\frac{(x_i^s-x_{im}^s)^2}{2h^2}\Big]\Big]\Big/ \sqrt{2\pi}h^2\sum_{m=1}^{N}\Big[\mathrm{sign}(c_m)\prod_{s=1}^{t(i)}\exp\Big[-\frac{(x_i^s-x_{im}^s)^2}{2h^2}\Big]\Big]. \quad (5)$$

1.3 结构学习

ENBC的结构学习是在NBC结构的基础上, 发现每一个属性结点的新父结点的过程. 首先根据Quinlan[12]的信息增益率为属性排序, 其中的属性条件密度计算采用高斯核函数, 平滑参数使用MISE进行设置. 通过实验发现具有三阶父结点的属性非常少, 而且三阶父结点对分类的贡献也很小, 因此只进行到三阶父结点选择(如果综合考虑学习效率和分类准确性, 也可以只进行到二阶属性父结点选择). h^*表示使用MISE方法设置的平滑参数; $\mathrm{accuracy_nbc}(h^*|D,S)$表示NBC的分类准确率; $\mathrm{accuracy_bnc_}f(j|D,G_j), j=1,2,3$和$\mathrm{accuracy_bnc_}b(j|D,G_j)$分别表示进行第$j$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EBNC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表示属性X_i的父结点候选集. 具体的属性父结点选择算法如表1所示.

表1 属性父结点选择算法

1. 输入: D, h^*和$\mathrm{accuracy_nbc}\,(h^*|D,S)$
2. 设置$\Pi_i^{(1)}=\Pi_i^{(2)}=\Pi_i^{(3)}=\phi$ and $\Gamma_i=\{X_1,X_2,\cdots,X_{i-1}\}$
3. for $j=1,2,3$
4. if $j=1$
5. $\mathrm{accuracy\_bnc\_}f(j|D,G_1)=\mathrm{accuracy\_nbc}(h^*|D,S)$
6. 用S初始化G_1
7. else
8. $\mathrm{accuracy\_bnc\_}f(j|D,G_j)=\mathrm{accuracy\_bnc\_}b(j-1|D,G_j)$
9. 用G_{j-1}初始化G_j
10. for $i=1,\cdots,n$
11. 计算Γ_i, 设$s=0$
12. for $k=1,\cdots,i-1$
13. if $X_k\in\Gamma_i$
14. 计算$\mathrm{accuracy_bnc_}b(j|D,G_j)$
15. if $\mathrm{accuracy\_bnc\_}b(j|D,G_j)>\mathrm{accuracy\_bnc\_}f(j|D,G_j)$
16. $\mathrm{accuracy\_bnc\_}f(j|D,G_j)=\mathrm{accuracy\_bnc\_}b(j|D,G_j)$
17. $s=k$
18. if $s\neq 0$
19. if $\Pi_i^{(1)}=\phi$
20. $\Pi_i^{(1)}=\{X_s\}$, 更新Γ_i和G_1
21. if $\Pi_i^{(1)}\neq\phi$ and $\Pi_i^{(2)}=\phi$
22. $\Pi_i^{(2)}=\{X_s\}$, 更新Γ_i和G_2
23. if $\Pi_i^{(1)}\neq\phi$ and $\Pi_i^{(2)}\neq\phi$ and $\Pi_i^{(3)}=\phi$
24. $\Pi_i^{(3)}=\{X_s\}$, 更新Γ_i和G_3
25. if $J>1$ and $\mathrm{accuracy\_bnc\_}b(j|D,G_j)=\mathrm{accuracy\_bnc\_}b(j-1|D,G_j)$
26. 退出j循环
27. 输出: $\mathrm{accuracy\_bnc\_}b(j|D,G_j),(j=1,2,3)$、$G_j$和$G$.

属性父结点选择是ENBC学习的主要部分, 也是运算的主体. 在属性父结点选择的过程中, 所进行的主要运算又是分类器的分类准确率估计, 一阶、二阶和三阶依赖扩展最多需要$3(n-1)n$次的分类准确率运算, 因此属性父结点选择算法关于分类准确率运算的时间复杂度是$O(n^2)$.

2 实验与分析

选择12个分类器与ENBC进行分类准确性比较实验并分析(使用UCI数据库中的28个分类数据集), 其中前3个是离散属性分类器(采用基于熵的方法对连续属性进行离散化), 后10个是连续属性分类器. 用于比较实验的分类器是: 离散属性NBC (DNBC); 离散属性TAN (DTAN); 结合MDL (minimal description length)标准、属性排序(使用信息增益率)和贪婪搜索建立的离散属性约束贝叶斯网络分类器(RBNC); 使用具有平滑参数的高斯核函数估计属性边缘密度, 并基于wrapper方法优化平滑参数的NBC (ONBC); 使用具有平滑参数的高斯核函数估计属性联合密度, 并基于MISE标准设置平滑参数的完全贝叶斯分类器(JGK); Pérez等[7]给出的Flexible naive Bayes classifier (FNBC); 采用Pérez等[6-7]方法, 分别基于高斯函数和高斯核函数估计属性密度, 对NBC进行一阶依赖扩展的分类器(OGNB, OKNB); Bounhas等[9]和He等[10]给出的贝叶斯分类器(GBC, GKBC); Quinlan[12]的决

表 2　分类错误率实验结果

数据集	DNBC	DTAN	RBNC	ONBC	JGK	FNBC	OGNB	OKNB	GBC	GKBC	C4.5	SVM	ENBC
Arabic_ Digit	0.274 0	0.279 5	0.235 7	0.297 3	0.371 3	0.349 4	0.263 1	0.278 9	0.297 3	0.264 7	0.323 3	0.232 7	0.215 1
Breast_ cancer	0.035 8	0.048 6	0.040 6	0.027 2	0.052 9	0.027 2	0.038 6	0.027 2	0.030 2	0.021 5	0.060 1	0.043 1	0.022 9
Breast_ tissue	0.290 0	0.320 0	0.311 4	0.360 0	0.430 0	0.560 0	0.335 0	0.455 0	0.383 6	0.341 1	0.349 1	0.322 5	0.310 0
Cardiotocography	0.254 2	0.238 9	0.260 4	0.341 7	0.480 6	0.582 0	0.399 3	0.473 4	0.351 5	0.306 4	0.211 8	0.502 6	0.207 3
Cmc	0.328 5	0.327 0	0.345 8	0.340 2	0.344 6	0.347 5	0.333 6	0.340 2	0.351 4	0.322 5	0.317 6	0.308 0	0.314 9
Connectionist_Bench	0.403 9	0.386 6	0.300 0	0.184 7	0.028 9	0.263 5	0.187 5	0.264 5	0.025 6	0.271 5	0.210 2	0.413 5	0.028 9
Echocardiogram	0.638 5	0.515 4	0.353 9	0.338 5	0.438 5	0.353 9	0.400 6	0.338 5	0.307 7	0.343 7	0.447 0	0.375 4	0.338 5
Ecoli	0.103 5	0.100 0	0.079 4	0.048 3	0.058 7	0.058 7	0.062 1	0.075 9	0.064 2	0.055 0	0.0994	0.051 5	0.048 3
Glass	0.342 9	0.300 0	0.287 9	0.290 5	0.438 1	0.462 0	0.409 6	0.404 8	0.337 3	0.317 6	0.359 9	0.301 2	0.271 5
Heart_disease	0.133 4	0.144 5	0.151 9	0.163 0	0.244 5	0.166 7	0.161 1	0.155 5	0.122 3	0.155 8	0.226 0	0.171 9	0.140 8
Horse_colic	0.283 4	0.283 4	0.283 4	0.316 7	0.333 4	0.340 0	0.335 0	0.286 7	0.312 4	0.243 3	0.313 4	0.310 0	0.176 7
Image_Segmentation	0.145 0	0.135 0	0.112 6	0.155 0	0.175 0	0.230 0	0.358 7	0.205 0	0.222 5	0.169 2	0.114 5	0.117 2	0.110 0
Ionosphere	0.257 2	0.285 8	0.294 3	0.320 0	0.434 3	0.325 8	0.438 6	0.272 9	0.312 9	0.251 4	0.234 9	0.220 9	0.217 2
Iris	0.040 0	0.040 0	0.046 7	0.033 4	0.046 7	0.046 7	0.033 4	0.043 4	0.026 7	0.046 7	0.046 7	0.040 0	0.033 4
Japanese_vowels	0.265 0	0.262 4	0.259 8	0.201 3	0.098 8	0.196 2	0.120 8	0.203 3	0.240 8	0.158 5	0.207 6	0.056 2	0.097 5
Liver_disease	0.300 0	0.288 3	0.267 7	0.344 2	0.394 2	0.353 0	0.385 3	0.375 1	0.357 6	0.302 9	0.327 6	0.287 4	0.323 6
MAGIC_Gamma	0.252 2	0.281 7	0.293 0	0.245 2	0.300 0	0.286 2	0.305 7	0.245 2	0.261 9	0.255 3	0.189 4	0.190 7	0.189 4
New_thyroid	0.045 5	0.045 5	0.045 5	0.054 6	0.059 1	0.086 4	0.050 0	0.093 3	0.066 7	0.057 1	0.093 1	0.056 2	0.045 5
Parkinsons	0.165 0	0.120 0	0.150 0	0.095 0	0.220 0	0.160 0	0.178 1	0.140 0	0.062 5	0.035 0	0.128 3	0.051 5	0.070 0
Pima	0.239 0	0.235 6	0.231 5	0.239 0	0.344 2	0.240 3	0.235 6	0.240 3	0.251 8	0.238 7	0.265 7	0.232 9	0.229 9
Red_wine	0.425 0	0.451 3	0.428 2	0.435 0	0.557 5	0.435 0	0.418 8	0.430 4	0.471 3	0.441 5	0.378 4	0.472 0	0.418 8
Sensor_readings	0.155 6	0.124 5	0.148 9	0.180 0	0.340 0	0.306 7	0.353 3	0.224 5	0.246 2	0.198 3	0.057 1	0.190 2	0.111 2
Statlog	0.165 7	0.123 0	0.116 0	0.070 3	0.039 0	0.093 2	0.114 9	0.082 1	0.079 5	0.102 4	0.049 6	0.041 2	0.034 4
Transfusion	0.241 4	0.204 0	0.201 4	0.230 7	0.272 0	0.234 7	0.274 0	0.234 7	0.224 0	0.245 6	0.230 0	0.220 7	0.229 4
Wdbc	0.044 7	0.037 5	0.053 0	0.059 0	0.050 0	0.052	0.047 4	0.054 5	0.031 3	0.051 4	0.073 9	0.017 6	0.005 9
Wine	0.141 2	0.129 5	0.046 3	0.029 5	0.058 9	0.041 2	0.011 8	0.029 5	0.037 4	0.011 8	0.101 2	0.025 0	0.029 1
Wpbc	0.420 0	0.365 0	0.173 7	0.230 0	0.365 0	0.275 0	0.201 6	0.240 0	0.230 0	0.228 7	0.303 1	0.228 7	0.230 0
Yeast	0.454 8	0.400 7	0.389 2	0.404 1	0.416 9	0.411 5	0.385 7	0.451 4	0.421 7	0.411 5	0.396 3	0.389 2	0.365 8
平均	0.244 5	0.231 2	0.211 0	0.215 5	0.264 0	0.260 4	0.244 3	0.238 1	0.218 9	0.208 9	0.218 4	0.209 6	0.172 0
Wilcoxon SR test	※−3.96	※−3.91	※−3.34	※−4.25	※−4.62	※−4.59	※−3.94	※−4.41	※−3.78	※−3.66	※−3.71	※−2.50	ENBC
Friedman/Bonferroni test	※−4.32	※−5.11	※−4.41	※−3.27	※−7.64	※−7.91	※−6.82	※−6.54	※−4.36	※−5.63	※−3.14	※−3.26	ENBC

表 3　ENBC和其它分类器的分类准确性差异比较　%

关系	DNBC	DTAN	RBNC	ONBC	JGK	FNBC	OGNB	OKNB	GBC	GKBC	C4.5	SVM
$[0.5, \infty)$	85.72	85.72	75.00	75.00	89.28	96.44	82.14	89.28	75.00	82.14	71.42	60.71
$(-0.5, 0.5)$	3.56	7.14	10.72	25.00	10.72	3.56	10.72	10.72	7.14	7.14	21.44	21.43
$(-\infty, -0.5]$	10.72	7.14	10.428	0.00	0.00	0.00	7.14	0.00	17.86	10.72	7.14	17.86

策树分类器; 支持向量机分类器 (SVM, libsvm, http://www.csie.ntu.edu.tw/); 使用具有平滑参数的高斯核函数估计属性密度, 并结合属性排序、分类准确性标准与属性父结点贪婪选择而建立的扩展的NBC (ENBC).

采用10折交叉有效性验证方法[13]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[14]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ENBC与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ENBC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, ENBC与12个分类器在错分率方面差异显著. 再考察总体平均值, ENBC优于其他12个分类器的程度依次是9.61%、7.70%、4.94%、5.54%、12.50%、11.95%、9.57%、8.68%、6.00%、4.66%、5.94%和4.76%, 可见ENBC具有明显的优势.

ENBC相对于其他12个分类器的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)3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显著性检验、分类准确性平均值比较和分类准确性差异百分比计算3方面的结果显示了相对于其他12个分类器, ENBC在分类准确性方面具有明显的优势.

3 结 论

本文在使用通过边缘高斯函数的乘积进行叠加的多元高斯核函数估计属性密度的基础上, 对朴素贝叶斯分类器进行了网络依赖扩展, 使扩展后的分类器能够充分利用属性之间的依赖信息. 从分类器之间的分类错误率显著性检验、分类准确性平均值比较和分类准确性差异百分比计算3个方面进行实验, 实验结果显示了ENBC具有良好的分类准确性. 进一步的工作是提高ENBC的效率, 并在ENBC的基础上建立回归模型.

参考文献(References)

[1] Wang S C, Xu G L, Du R J. Restricted Bayesian classification networks[J]. Science China Information Sciences, 2013, 56(7): 078105(1)-078105(15).

[2] Chow C K, Liu C N. Approximating discrete probability distributions with dependence trees[J]. IEEE Trans on Information Theory, 1968, 14(3): 462-467.

[3] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3): 131-161.

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

[5] John G H, Langley P. Estimating Continuous Distributions in Bayesian Classifiers[C]. Proc of the Eleventh Conf on Uncertainty in Artificial Intelligence(UAI-1995). San Mateo: 1995: 338-345.

[6] Pérez A, Larrañagaa P, Inzaa I. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes[J]. Int J of Approximate Reasoning, 2006, 43(1): 1-25.

[7] Pérez A, Larrañagaa P, Inza I. Bayesian classifiers based on kernel density estimation: Flexible classifiers[J]. Int J of Approximate Reasoning, 2009, 50(2): 341-362.

[8] 夏战国, 夏士雄, 蔡世玉, 等. 类不均衡的半监督高斯过程分类算法[J]. 通信学报, 2013, 34(5): 42-51.
(Xia Z G, Xia S X, Cai S Y ,et al. Semi-supervised Gaussian process classification algorithm addressing the class imbalance[J]. J on Communications, 2013, 34(5): 42-51.)

[9] Bounhas M, Mellouli K, Prade H, et al. Possibilistic classifiers for numerical data[J]. Soft Computing, 2013, 17(5): 733-751.

[10] He Y L, Wang R, Kwong S, et al. Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis[J]. Information Sciences, 2014, 259(2): 252-268.

[11] Murphy S L, Aha D W. UCI repository of machine learning databases[EB/OL]. [2014-9-17]. http://www.ics.uci.edu/ mlearn/MLRepository.Html.

[12] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[13] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]. Proc of the 14th Int Joint Conf on Artificial Intelligence(IJCAI). Montréal: Morgan Kaufmann, 1995: 1137-1143.

[14] Demsar J. Statistical comparisons of classifiers over multiple data sets[J]. J of Machine Learning Research, 2006, 7(1): 1-30.

(责任编辑: 齐 霁)