文章编号: 1001-0920 (2013) 06-0855-06

嵌入非对称拒识代价的二元分类算法

郑恩辉[1a], 徐 欢[1a], 孙 坚[1a], 王 凌[1a], 陆慧娟[1b,2], 李 平[3]

(1. 中国计量学院 a. 机电工程学院, b. 信息工程学院, 杭州 310018; 2. 中国矿业大学 信息与电气工程学院, 江苏 徐州 221008; 3. 浙江大学 工业控制技术国家重点实验室, 杭州 310027)

摘 要: 针对传统分类算法隐含的假设 (相信并且接受每个样本的分类结果) 在医疗/故障诊断和欺诈/入侵检测等领域中并不适用的问题, 提出嵌入非对称拒识代价的二元分类问题, 并对其进行简化. 在此基础上设计出基于支持向量机 (SVM) 的代价敏感分类算法 (CSVM-CRC). 该算法包括训练 SVM 分类器、计算后验概率、估计分类可靠性和确定最优拒识阈值 4 个步骤. 基于 10 个 Benchmark 数据集的实验研究表明, CSVM-CRC 算法能够有效降低平均代价.

关键词: 结构风险最小化; 非对称拒识代价; 分类可靠性; 支持向量机

中图分类号: TP18 **文献标志码:** A

Binary classification algorithm with class-dependent reject cost

ZHENG En-hui[1a], *XU Huan*[1a], *SUN Jian*[1a], *WANG Ling*[1a], *LU Hui-juan*[1b,2], *LI Ping*[3]

(1a. College of Mechanical and Electrical Engineering, 1b. College of Information Engineering, China Jiliang University, Hangzhou 310018, China; 2. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221008, China; 3. State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China. Correspondent: ZHENG En-hui, E-mail: ehzheng@cjlu.edu.cn.)

Abstract: To minimize "0-1" loss, most of conventional classification algorithms non-explicitly assume that all results of classification are accepted. However, the assumption is inapplicability to knowledge extraction in such fields as medical/fault diagnosis and fraud/intrusion detection. Therefore, the binary classification problem with class-dependent reject cost(BCP-CRC) is summarized and is simplified, on basis of which the algorithm based on cost-sensitive support vector machines with CRC(CSVM-CRC) is formulated. The CSVM-CRC algorithm involves training a classifier based on SVM algorithm, computing the post probability of each sample, estimating the classification reliability of each sample, and determining the optimal reject threshold. The experiment results show that the CSVM-CRC algorithm can reduce the average cost effectively.

Key words: structural risk minimization; class-dependent reject cost; classification reliability; support vector machine

0 引 言

传统的分类算法基于经验风险最小化 (ERM) 准则, 以渐近理论和大数定理为依据, 但经验风险最小并不能保证其参数与期望风险最小时的参数相同, 导致分类器在应用中存在过拟合、模型过于复杂等缺点. 结构风险最小化 (SRM) 准则最小化模型复杂度和经验风险的折衷, 在理论和应用中保证了有限样本下分类器的泛化能力[1-2].

在某些实际应用领域 (如医疗诊断、欺诈检测和故障诊断等) 中, 以追求高分类精度为目标的传统分类器并不适用, 提高可靠性和降低平均代价显得尤为重要[3]. 因此, 分类算法的设计应考虑到: 当某样本的分类可靠性低于某个阈值时, 为了避免误分类导致的高代价, 需要对样本进行拒识决策, 即不接受分类可靠性低的模式自动分类的结果. Chow[4]通过给贝叶斯系统增加一个拒识选项, 解决误差率和拒识率的折衷问题. Foggia 等[5]基于 Bayesian 规则提出了一个解决多专家系统误差率和拒识率最优折衷的方法. Claudio 等[6]将 Foggia 的研究扩展到神经网络, 基于分类可靠性和预定义的拒识阈值, 对可靠性低于预设阈值的样本进行拒识操作. 针对拒识代价, Giorgio 等[7]基于 SRM 原则提出拒识代价敏感 SVM 算法. 针对同时包含已知模式和未知模式的分类问题, Thomas 等[8]基于分类阈值 (针对已知模式) 和拒识阈值 (针对未知模式), 解决了已知模式和未知模式分类的最优折衷问

收稿日期: 2012-02-16; **修回日期:** 2012-06-09.

基金项目: 国家自然科学基金项目(60905034, 60842009); 浙江省自然科学基金项目(Y1080950, Y1100376, Y1110342).

作者简介: 郑恩辉(1975–), 男, 副教授, 博士, 从事数据挖掘、复杂系统建模与控制等研究; 徐欢(1988–), 女, 硕士生, 从事图像处理的研究.

题. Zheng 等[9]设计了 CSVM-RC^{2}EC 算法, 在嵌入依赖于类别的误差代价的基础上嵌入拒识代价, 其研究结果能有效降低分类器的平均代价.

为了提高分类可靠性, 上述研究在分类算法中嵌入拒识选项, 不接受分类可靠性低的分类结果, 但没有考虑到拒识不同样本产生的代价 (损失) 也是不同的. 本文认为被"拒识"的样本需要其他过程处理, 拒识代价相对于类别而言是非对称的. 本文从实际需求中抽象出嵌入非对称拒识代价的二元分类问题 (BCP-CRC), 并对其进行数学描述和算法实现研究. 本文关注的拒识代价是依赖于类别的.

1 支持向量机

SVM 是一种强大的机器学习和数据挖掘算法, 是统计学习理论中 SRM 原则的具体实现. SVM 最小化模型复杂度和经验风险的折衷, 较神经网络等传统算法具有更好的泛化能力[1]. 假设超平面为 $(\omega x)-b=0$, SVM 最小化期望风险, 即

$$\min R(\alpha)=\|\omega\|^2+C\Big(\sum_{i=1}^{n}\xi_i\Big);$$

$$\text{s.t. } y_i(x_i\omega+b)\geqslant 1-\xi_i,\ \xi_i\geqslant 0,\ i=1,2,\cdots,n. \quad (1)$$

其中: $\|\omega\|^2$ 为模型复杂度, ξ_i 为松弛变量 (训练样本误差), C 为松弛因子.

2 BCP-CRC问题的描述与分析

定义 1 嵌入非对称拒识代价的二元分类问题可定义为四元组 BCP-CRC(D,F,L,A). 其中: D 为系统的观测样本集, F 为决策函数的假设空间, L 为与系统和决策相关的代价 (损失) 函数, A 为 BCP-CRC 问题的解决算法. 基于 D 和 L, 算法 A 在 F 中寻找最优决策函数, 以最小化 D 上的期望代价.

定义 2 BCP-CRC 问题中的观测样本集 D 为

$$(x_1,y_1,r_1,m_1),\cdots,(x_i,y_i,r_i,m_i),\cdots,(x_n,y_n,r_n,m_n),$$
$$x_i\in R^l,\ y_i\in\{+1,-1\},\ m_i,r_i\in R,\ i=1,2,\cdots,n. \quad (2)$$

其中: l 为预测属性的维数, m_i 和 r_i 分别为第 i 个样本的误差代价和拒识代价. 在 BCP-CRC 问题中, 定义 $m_i=1$, $r_i\in\{r^+,r^-\}$, $r^-\leqslant r^+\leqslant 1$, $i=1,2,\cdots,n$, r^+ 和 r^- 分别为"+1"类和"−1"类样本的拒识代价.

定义 3 BCP-CRC 问题中的决策函数假设空间为 $F=\{f_r(\alpha,\sigma_+,\sigma_-,x)\}_{(\alpha,\sigma_+,\sigma_-)}$, 不同参数组合$(\alpha,\sigma_+,\sigma_-)$ 对应不同的备选决策函数, 有

$$f_r(\alpha,\sigma_+,\sigma_-,x)=$$
$$\begin{cases}+1,\ \psi(\alpha,\sigma_+,\sigma_-,x)\geqslant\sigma_+,\ P(+1|x)\geqslant p(-1|x);\\ -1,\ \psi(\alpha,\sigma_+,\sigma_-,x)\geqslant\sigma_-,\ P(+1|x)<p(-1|x);\\ 0,\ \psi(\alpha,\sigma_+,\sigma_-,x)<\sigma_+I(y=+1)+\sigma_-I(y=-1).\end{cases} \quad (3)$$

其中: $0\leqslant\psi(\alpha,\sigma_+,\sigma_-,x)\leqslant 1$ 为样本 x 的分类可靠性; $0\leqslant\sigma_+,\sigma_-\leqslant 1$ 分别为"+1"类和"−1"类样本的拒识阈值; α 为模型参数; $P(+1|x)$ 和 $p(-1|x)$ 分别为 x 属于"+1"类和"−1"类的后验概率.

注 1 $I(z)$ 为指示函数, 当 z 为"真" 时, $I(z)=1$; 否则 $I(z)=0$.

注 2 给定拒识阈值 (σ_+,σ_-), 当"$\psi(\alpha,\sigma_+,\sigma_-,x)<\sigma_+$ and $y=+1$" 为 "真" 时, 或 "$\psi(\alpha,\sigma_+,\sigma_-,x)<\sigma_-$ and $y=-1$" 为 "真" 时, 不相信分类器对 x 的分类结果, 即采用"拒识"决策, 此时 $f_r(\alpha,\sigma_+,\sigma_-,x)=0$.

定义 4 BCP-CRC 问题中代价 (损失) 函数 L 为

$$L_r(x,y,f_r(\alpha,\sigma_+,\sigma_-,x))=$$
$$\begin{cases}0,\ f_r(\alpha,\sigma_+,\sigma_-,x)=y,\\ \quad \psi(\alpha,\sigma_+,\sigma_-,x)\geqslant(\sigma_+I(y=+1)+\sigma_-I(y=-1));\\ 1,\ f_r(\alpha,\sigma_+,\sigma_-,x)\neq y,\\ \quad \psi(\alpha,\sigma_+,\sigma_-,x)\geqslant(\sigma_+I(y=+1)+\sigma_-I(y=-1));\\ r^+,\ f_r(\alpha,\sigma_+,\sigma_-,x)=0,\ y=+1;\\ r^-,\ f_r(\alpha,\sigma_+,\sigma_-,x)=0,\ y=-1.\end{cases} \quad (4)$$

定义 4 给出的 BCP-CRC 问题代价 (损失) 函数满足 SRM 准则要求的"实值有界"条件[1-2], 因此, 可遵循 SRM 准则设计针对 BCP-CRC 问题的分类算法 A. 基于 SRM 准则, 算法 A 最小化期望代价 (风险)

$$R(\alpha,\sigma_+,\sigma_-)=R_{\rm str}+CR_{\rm emp}. \quad (5)$$

其中: $R_{\rm str}$ 为结构代价 (表示模型复杂度), $R_{\rm emp}$ 为经验代价, C 为折衷 $R_{\rm str}$和 $R_{\rm emp}$的正则化系数.

令 P_e^+ 和 P_e^- 分别表示 "+1" 类和 "−1" 类样本的误差率, P_r^+ 和 P_r^- 分别表示 "+1" 类和 "−1" 类样本的拒识率, 则有

$$R_{\rm emp}=P_e^+n^++P_e^-n^-+r^+P_r^+n^++r^-P_r^-n^-, \quad (6)$$

其中 $n^+=\int I(y=+1)\mathrm{d}x$ 和 $n^-=\int I(y=-1)\mathrm{d}x$ 分别为"+1"类和"−1"类样本的样本数. 在 BCP-CRC 问题中, $R_{\rm str}$, P_e^+, P_e^-, P_r^+ 和 P_r^- 均是模型参数 α 和拒识阈值 (σ_+,σ_-) 的函数, 则期望代价 (5) 可改写为

$$R(\alpha,\sigma_+,\sigma_-)=$$
$$R_{\rm str}(\alpha,\sigma_+,\sigma_-)+C(n^+P_e^+(\alpha,\sigma_+,\sigma_-)+$$
$$n^-P_e^-(\alpha,\sigma_+,\sigma_-)+r^+n^+P_r^+(\alpha,\sigma_+,\sigma_-)+$$
$$r^-n^-P_r^-(\alpha,\sigma_+,\sigma_-)). \quad (7)$$

其中

$$P_e^+(\alpha,\sigma_+,\sigma_-)=$$
$$\frac{1}{n^+}\Big(\int I(y=+1)I(f_r(\alpha,\sigma_+,\sigma_-,x)=-1)\times$$

$$I(\psi(\alpha,\sigma_+,\sigma_-,x)\geqslant\sigma_+)\mathrm{d}x\Big),$$

$$P_e^-(\alpha,\sigma_+,\sigma_-)=$$

$$\frac{1}{n^-}\Big(\int I(y=-1)I(f_r(\alpha,\sigma_+,\sigma_-,x)=+1)\times$$

$$I(\psi(\alpha,\sigma_+,\sigma_-,x)\geqslant\sigma_-)\mathrm{d}x\Big),$$

$$P_r^+(\alpha,\sigma_+,\sigma_-)=$$

$$\frac{1}{n^+}\Big(\int I(y=+1)I(\psi(\alpha,\sigma_+,\sigma_-,x)<\sigma_+)\mathrm{d}x\Big),$$

$$P_r^-(\alpha,\sigma_+,\sigma_-)=$$

$$\frac{1}{n^-}\Big(\int I(y=-1)I(\psi(\alpha,\sigma_+,\sigma_-,x)<\sigma_-)\mathrm{d}x\Big).$$

在 BCP-CRC 问题中, 算法 A 最小化期望代价 (7) 以获得分类器参数, 即

$$(\alpha^*,\sigma_+^*,\sigma_-^*)=\underset{\alpha,0\leqslant\sigma_+,\sigma_-\leqslant1}{\arg\min}R(\alpha,\sigma_+,\sigma_-),\tag{8}$$

其中 α^*, σ_+^* 和 σ_-^* 分别为最优模型参数和最优拒识阈值. 当 $r^+=r^-=r$, $\sigma_+=\sigma_-=\sigma=0$ 时, $R(\alpha)$ 对应精度最优问题, 即"0-1"损失问题. 期望代价 (7) 简化为

$$R(\alpha)=R_{\mathrm{str}}(\alpha)+CP_e(\alpha).\tag{9}$$

综上所述可得到以下结论: 1) 算法 A 基于 SRM 准则, 从理论上保证了分类器的泛化能力; 2) 在分类器中嵌入了非对称拒识代价, 扩展了传统分类算法适应实际需求的能力; 3) 当应用领域的情况与上述定义相符时, 只要给出分类可靠性的计算方法与正 (反) 例样本的拒识率和误差率的估计方法, 即可基于 BCP-CRC 设计分类器, 实现知识提取. 但是, 由于 α, σ_+ 和 σ_- 同时影响期望代价中的结构代价和经验代价, 直接根据算法 A 设计分类器较为复杂. 以下定义简化版的 BCP-CRC(SBCP-CRC) 问题.

3　SBCP-CRC 问题的描述与分析

定义 5　BCP-CRC 问题可以定义为一个四元组 SBCP-CRC(D,F_s,L_s,A_s). 其中: 令 L 中 $f_r(\alpha,\sigma_+,\sigma_-,x)=f(\alpha,x)$, 则 $L_s=L$, $f(\alpha,x)$ 对应以分类精度为目标的无拒识问题; $F_s=\{f(\alpha,x),(\sigma_+,\sigma_-)\}_{(\alpha,\sigma_+,\sigma_-)}$; A_s 为与 D, F_s 和 L_s 相对应的算法.

首先, 令 $\sigma_+=\sigma_-=0$, 则有 $P_r^+=P_r^-=0$, $P_e=P_e^++P_e^-$, 由此期望代价 (7) 可转化为

$$R(\alpha,\sigma_+=\sigma_-=0)=$$

$$R_{\mathrm{str}}(\alpha,\sigma_+=\sigma_-=0)+CnP_e(\alpha,\sigma_+=\sigma_-=0).\tag{10}$$

期望代价 (10) 对应的是以分类精度为目标的无拒识问题, 其相应算法是基于 SRM 准则的, 记为 $A_s(\alpha,\sigma_+=\sigma_-=0)$ (简记为 $A_s(\alpha)$), 算法 $A_s(\alpha)$ 最小化期望代价 (10), 即

$$\min R(\alpha,\sigma_+=\sigma_-=0),\tag{11}$$

基于 $A_s(\alpha)$ 算法的分类器参数为

$$\hat{\alpha}^*=\arg\min_{\alpha}R(\alpha,\sigma_+=\sigma_-=0).\tag{12}$$

然后, 将分类器参数 $\hat{\alpha}^*$ 代入式 (7), 并假设 $R_{\mathrm{str}}(\hat{\alpha}^*,\sigma_+,\sigma_-)=R_{\mathrm{str}}(\hat{\alpha}^*,\sigma_+=\sigma_-=0)$, 可得到

$$R(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$

$$R_{\mathrm{str}}(\hat{\alpha}^*,\sigma_+=\sigma_-=0)+C(n^+P_e^+(\hat{\alpha}^*,\sigma_+,\sigma_-)+$$

$$n^-P_e^-(\hat{\alpha}^*,\sigma_+,\sigma_-)+r^+n^+P_r^+(\hat{\alpha}^*,\sigma_+,\sigma_-)+$$

$$r^-n^-P_r^-(\hat{\alpha}^*,\sigma_+,\sigma_-)).\tag{13}$$

最优拒识阈值为

$$(\hat{\sigma}_+^*,\hat{\sigma}_-^*)=\underset{0\leqslant\sigma_+,\sigma_-\leqslant1}{\arg\min}R(\hat{\alpha}^*,\sigma_+,\sigma_-).\tag{14}$$

与算法 A 根据式 (8) 确定分类器参数 $(\alpha^*,\sigma_+^*,\sigma_-^*)$ 不同, 算法 A_s 根据式 (12) 和 (14) 确定分类器参数 $(\hat{\alpha}^*,\hat{\sigma}_+^*,\hat{\sigma}_-^*)$. 算法 $A_s(\alpha)$ 通过 2 个步骤实现分类器参数的获取: 1) 假设 $\sigma_+=\sigma_-=0$, 对应的是基于"0-1"损失的无"拒识"问题; 2) 在模型参数 $\hat{\alpha}^*$ 已知的情况下, 引入非对称拒识代价以确定最优拒识阈值 (σ_+^*,σ_-^*). 在 $A_s(\alpha)$ 中, 结构代价 $R_{\mathrm{str}}(\alpha,\sigma_+,\sigma_-)$ 与 (σ_+,σ_-) 无关, 即 $R_{\mathrm{str}}(\alpha,\sigma_+,\sigma_-)=R_{\mathrm{str}}(\alpha)$. 以下基于 SVM 给出算法 A_s 的一个实现方法, 称为 CSVM-CRC 算法.

4　CSVM-CRC 算法

4.1　一类 $A_s(\alpha)$ 和 SVM 的等效性

假设 1　算法 $A_s(\alpha)$ 的决策超平面为 $(\omega x)-b=0$, 结构代价为

$$R_{\mathrm{str}}(\alpha,\sigma_+=\sigma_-=0)=R_{\mathrm{str}}(\alpha)=\frac{1}{2}\|\omega\|^2.\tag{15}$$

引入参数 $\xi_x(y(x\omega+b)=1-\xi_x,\xi_x\geqslant0)$ 表示样本 x 的分类误差, 则总误差为

$$S_e=\int\xi_x\mathrm{d}x.\tag{16}$$

假设 2　存在一个正实数 $d(d\in|R|)$ 使得

$$P_e(\alpha,\sigma_+=\sigma_-=0)=dS_e.\tag{17}$$

将式 (15) ~ (17) 代入期望代价 (10) 得到

$$R'(\alpha,\sigma_+=\sigma_-=0)=\frac{1}{2}\|\omega\|^2+Cd\int\xi_x\mathrm{d}x,\tag{18}$$

则算法 $A_s(\alpha)$ 对应的优化问题 (11) 转化为

$$\min R'(\alpha,\sigma_+=\sigma_-=0);$$

$$\text{s.t. } y(x\omega+b)\geqslant1-\xi_x,\ \xi_x\geqslant0.\tag{19}$$

定理 1　如果假设 1 和假设 2 同时成立, 则优化问题 (11) 等价于优化问题 (19).

根据 $\int\xi_x\mathrm{d}x=\sum_{i=1}^{n}\xi_i$, 令 $C'=Cd$, 优化问题 (19) 可转化为

$$\min\frac{1}{2}\|\omega\|^2+C'\sum_{i=1}^{n}\xi_i;$$

$$\text{s.t. } y(x\omega+b)\geqslant1-\xi_x,\ \xi_x\geqslant0.\tag{20}$$

显然, 优化问题 (19) 与式 (1) 一致, 由此得到定理 2.

定理 2 如果假设 1 和假设 2 同时成立, 则优化问题 (11) 等价于式 (1), 即 $A_s(\alpha)$ 算法等效于 SVM.

根据定理 2, 可以将 $A_s(\alpha)$ 转化为 SVM 进行分类器的设计, 解决以分类精度为目标的分类问题.

4.2 分类可靠性

SVM 算法的决策超平面为 $\omega x+b=0$, 引入 S 型函数估计 x 属于类 "+1" 的后验概率为

$$p(+1|x)=1/(1+\exp(-(\omega x+b)/\|\omega\|)),$$

则 x 属于类 "−1" 的概率为

$$p(-1|x)=1-p(+1|x). \tag{21}$$

根据式 (21) 可计算后验概率矩阵 $P(n,2)$, 其元素 $P(i,1)=p(+1|x_i)$ 和 $P(i,2)=p(-1|x_i)$ 规范化到区间 [0 1], $i=1,2,\cdots,n$.

SVM 的决策函数为

$$f(\alpha,x)=\begin{cases}+1,\ p(+1|x)\geqslant(-1|x);\\ -1,\ p(+1|x)<(-1|x).\end{cases} \tag{22}$$

某些样本分类可靠性低的原因在于: 这些样本在两个类别的重叠区域, 或者离任意类别的中心均很远[13-14]. 对于样本 x, 其最大后验概率 π_1 (属于获胜类的概率) 和次大后验概率 π_2 定义为

$$\pi_1=\max(p(+1|x),p(-1|x)),$$
$$\pi_2=\min(p(+1|x),p(-1|x)). \tag{23}$$

显然, $0\leqslant\pi_2\leqslant\pi_1\leqslant 1$. 若定义分类可靠性的两个影响因素为

$$\psi_a=\pi_1,\ \psi_b=1-\pi_2/\pi_1, \tag{24}$$

则样本 x 的分类可靠性由 ψ_a 和 ψ_b 共同确定. ψ_a 小表示该样本离两个类别的中心均很远, ψ_b 小表示该样本在两个类别的重叠区域附近. 因此, 样本 x 分类可靠性 ψ 可定义为 ψ_a 和 ψ_b 的函数[13-14], 即

$$\psi=\psi(\alpha,x)=\psi(\psi_a,\psi_b)=\frac{1}{2}(\psi_a+\psi_b). \tag{25}$$

根据式 (25) 可以估计描述样本分类可靠性的矩阵 $R(n,1)$, 其元素 $R(i,1)=\psi(x_i)$, $i=1,2,\cdots,n$, 规范化到 [0 1] 区间.

4.3 最优拒识阈值

根据优化问题 (1) 得到分类器参数

$$\hat{\alpha}^*=\{\omega^*,b^*,C^*,o^*\}, \tag{26}$$

其中 o^* 为核函数参数等其他参数. 结构代价为

$$R_{\rm str}(\hat{\alpha}^*,\sigma_+=\sigma_-=0)=\|\omega^*\|^2/2. \tag{27}$$

将式 (26) 和 (27) 代入 (13) 得

$$R''(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$
$$\|\omega^*\|^2/2+C^*(n^+P_e^+(\hat{\alpha}^*,\sigma_+,\sigma_-)+$$
$$n^-P_e^-(\hat{\alpha}^*,\sigma_+,\sigma_-)+r^+n^+P_r^+(\hat{\alpha}^*,\sigma_+,\sigma_-)+$$
$$r^-n^-P_r^-(\hat{\alpha}^*,\sigma_+,\sigma_-)). \tag{28}$$

其中

$$P_e^+(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$
$$\frac{1}{n^+}\Big(\int I(y=+1)I(f(\hat{\alpha}^*,x)=-1)I(\psi(\hat{\alpha}^*,x)\geqslant\sigma_+)\mathrm{d}x\Big),$$
$$P_e^-(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$
$$\frac{1}{n^-}\Big(\int I(y=-1)I(f(\hat{\alpha}^*,x)=+1)I(\psi(\hat{\alpha}^*,x)\geqslant\sigma_-)\mathrm{d}x\Big),$$
$$P_r^+(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$
$$\frac{1}{n^+}\Big(\int I(y=+1)I(\psi(\hat{\alpha}^*,x)<\sigma_+)\mathrm{d}x\Big),$$
$$P_r^-(\hat{\alpha}^*,\sigma_+,\sigma_-)=$$
$$\frac{1}{n^-}\Big(\int I(y=-1)I(\psi(\hat{\alpha}^*,x)<\sigma_-)\mathrm{d}x\Big).$$

如果给定拒识阈值 (σ_+,σ_-) 可以计算相应的期望代价 $R''(\hat{\alpha}^*,\sigma_+,\sigma_-)$, 则最优拒识阈值为

$$(\hat{\sigma}_+^*,\hat{\sigma}_-^*)=\mathop{\arg\min}_{0\leqslant\sigma_+,\sigma_-\leqslant1}R''(\hat{\alpha}^*,\sigma_+,\sigma_-). \tag{29}$$

4.4 CSVM-CRC 算法步骤

CSVM-CRC 算法是 SBCP-CRC 问题中 A_s 算法基于 SVM 的实现, 其输入为训练样本集 (2), 输出为模型参数 $\hat{\alpha}^*$ (即决策超平面) 和最优拒识阈值 $(\hat{\sigma}_+^*,\hat{\sigma}_-^*)$. CSVM-CRC 算法步骤如下: 1) 根据训练样本集 (2) 和优化问题 (1) 确定分类器参数 $\hat{\alpha}^*=\{\omega^*,b^*,C^*,o^*\}$; 2) 根据式 (21) 计算后验概率矩阵 $P(n,2)$; 3) 根据式 (25) 计算分类可靠性矩阵 $R(n,1)$; 4) 根据式 (29) 确定最优拒识阈值 $(\hat{\sigma}_+^*,\hat{\sigma}_-^*)$.

5 实验研究

基于表 1 给出的 10 个 UCI Machine Learning 数据集[20]研究 CSVM-CRC 算法的有效性. 对于每个数

表 1 10 个 UCI Machine Learning 数据集

序号	名称	类别数	类分布(+1/−1)	属性个数
1	Australian	2	307/383	13
2	BCW	2	241/458	9
3	WDBC	2	212/357	30
4	Heart	2	120/150	13
5	Ionosphere	2	126/225	34
6	WPBC	2	47/151	33
7	BUPA	2	74/271	5
8	German	2	300/700	24
9	Hepatitis	2	70/85	19
10	Pima	2	268/500	8

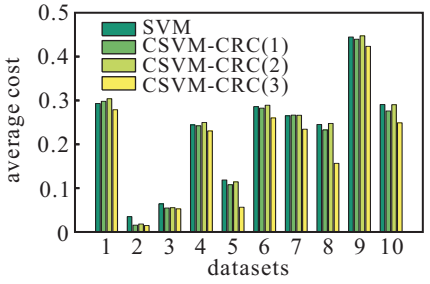


图 1 拒识代价设定对平均代价的影响

据集, 其误差代价和正例与反例的拒识代价设定见表2. 每次随机选择其中2/3的样本构成训练集, 其余的1/3构成测试集, 取20次实验的平均值作为实验结果. SVM采用径向基核函数 $K(x_i,x_j)=\mathrm{e}^{-\|x_i-x_j\|^2/(2\sigma)}$, 参数 $\sigma=0.001$.

5.1　拒识代价设定对平均代价的影响

对于表1中的每个数据集, 其代价设定见表2. 图1给出了不同拒识代价设定对CSVM-CRC分类器平均代价的影响和SVM分类器的平均代价. 在图1中, CSVM-CRC(1), CSVM-CRC(2)和CSVM-CRC(3)分别与表2中的拒识代价设定 $r^+/r^-(1)$, $r^+/r^-(2)$ 和 $r^+/r^-(3)$ 相对应. 对于每个数据集, CSVM-CRC(3)分类器得到的平均代价最小, CSVM-CRC(2)分类器得到的平均代价最大; 对比CSVM-CRC(3)和CSVM-CRC(1)可见, 较大的正例拒识代价设定可增加分类器的平均代价; 对比CSVM-CRC(2)和CSVM-CRC(1)可见, 较小的反例拒识代价设定可减小分类器的平均代价; 对比CSVM-CRC(3)和CSVM-CRC(2)可见, 正例和反例的拒识代价设定越大, 分类器平均代价越大.

表 2　Consideration set

名称	m^+/m^-	$r^+/r^-(1)$	$r^+/r^-(2)$	$r^+/r^-(3)$
Australian	1/1	0.4/0.3	0.4/0.4	0.3/0.3
BCW	1/1	0.06/0.03	0.06/0.06	0.03/0.03
WDBC	1/1	0.12/0.06	0.12/0.12	0.06/0.06
Heart	1/1	0.4/0.3	0.4/0.4	0.3/0.3
Ionosphere	1/1	0.4/0.1	0.4/0.4	0.1/0.1
WPBC	1/1	0.6/0.3	0.6/0.6	0.3/0.3
BUPA	1/1	0.8/0.4	0.8/0.8	0.4/0.4
German	1/1	0.6/0.2	0.6/0.6	0.2/0.2
Hepatitis	1/1	0.6/0.4	0.6/0.6	0.4/0.4
Pima	1/1	0.5/0.3	0.5/0.5	0.3/0.3

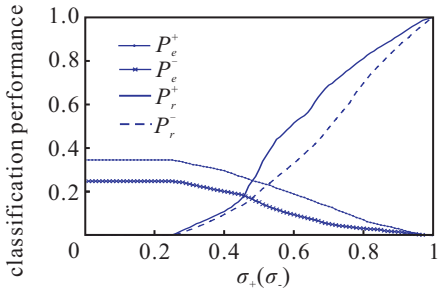


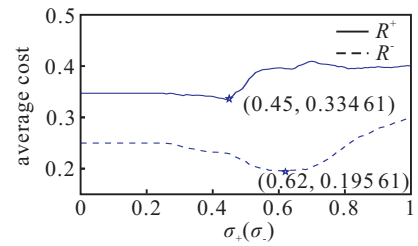
图 2　Australian 数据集拒识阈值对误差率和拒识率的影响

5.2　正(反)例拒识阈值对误差率和拒识率的影响

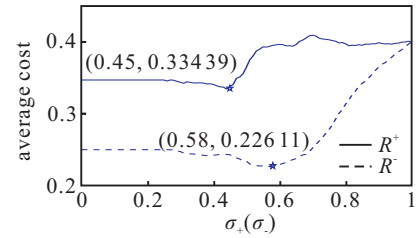
当误差代价一定时, 图2给出了数据集1上的正(反)例拒识阈值 $\sigma_+(\sigma_-)$ 对于正(反)例误差率 $P_e^+(P_e^-)$ 和拒识率 $P_r^+(P_r^-)$ 的影响. 随着拒识阈值 $\sigma_+(\sigma_-)$ 从0增加到1, 正(反)例拒识率 $P_r^+(P_r^-)$ 从0到1递增, 而正(反)例误差率 $P_e^+(P_e^-)$ 递减. 当正(反)例拒识阈值 $\sigma_+(\sigma_-)$ 为0时, CSVM-CRC分类器等效于传统SVM分类器; 当正(反)例的拒识阈值 $\sigma_+(\sigma_-)$ 为1时, 误差率降至0, 此时分类器拒绝接受所有样本的分类结果. 若领域知识明确给定误差代价和非对称拒识代价, 则可以得到正(反)例平均代价随正(反)例拒识阈值 $\sigma_+(\sigma_-)$ 的变化规律, 进而确定最优拒识阈值.

5.3　正(反)例拒识阈值对正(反)例平均代价的影响

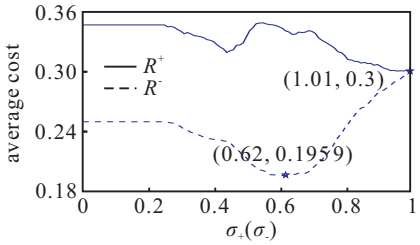
本节研究当拒识代价设定分别为 $r^+/r^-(1)$, $r^+/r^-(2)$ 和 $r^+/r^-(3)$ 时, 正(反)例拒识阈值 $\sigma_+(\sigma_-)$ 对于正(反)例平均代价 $R^+(R^-)$ 的影响. 限于篇幅, 图3给出了1个数据集的实验结果. 正(反)例不同的拒识代价设定并不影响CSVM-CRC分类器零拒识时(此时CSVM-CRC等效于SVM)的分类性能. 当正(反)例拒识阈值 $\sigma_+(\sigma_-)$ 增大到一定数值后, 不同代价设定开始对正(反)例的平均代价 $R^+(R^-)$ 产生明显的影响: 1) 较大的拒识代价设定从整体趋势上会使正(反)例的平均代价有所增加; 2) 较大的拒识代价设定会使正(反)例的最优拒识阈值减小.



(a)　$r^+/r^-(1)$=0.4/0.3



(b)　$r^+/r^-(2)$=0.4/0.4



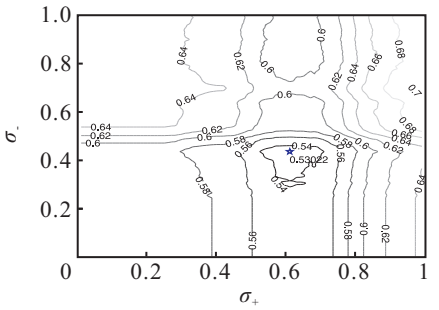
(c)　$r^+/r^-(3)$=0.3/0.3

图 3　Australian 数据集拒识阈值对正(反)例平均代价的影响

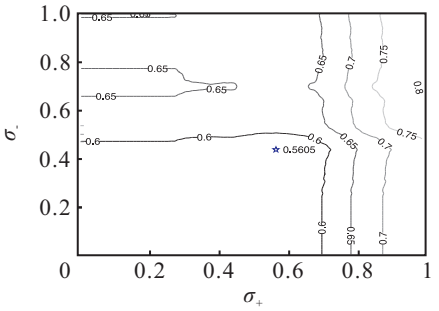
5.4　正例和反例拒识阈值对全局平均代价的影响

本节研究不同拒识代价设定下的正例和反例拒识阈值对于全局平均代价的影响. 图4给出了1个数据集正例、反例拒识阈值 (σ_+,σ_-) 与总平均代价之间的对应关系和全局平均代价最小点的分布情况, 星号标注点为全局平均代价最低点. 随着正例或反例拒识

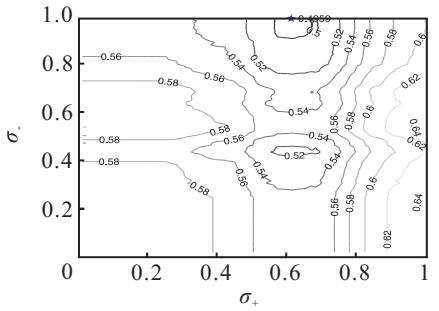
代价的增大, 全局平均代价的最小点向正例或反例拒识阈值小的方向移动.



(a) $r^+/r^-(1)=0.4/0.3$



(b) $r^+/r^-(2)=0.4/0.4$



(c) $r^+/r^-(3)=0.3/0.3$

图 4 Australian 数据集拒识阈值对全局平均代价的影响

6 结 论

当分类算法应用到具有非对称拒识代价的领域时, 平均代价和分类可靠性成为评价分类算法性能的重要标准. 传统分类器以高分类精度为目标, 显然不能满足上述领域的实际需求. 本文研究在分类器设计中嵌入非对称拒识代价的分类算法, 旨在提高分类可靠性和降低平均代价. 从医疗/故障诊断和欺诈/入侵检测等领域背景中提炼出嵌入非对称拒识代价的二元分类问题 BCP-CRC, 并在提出 BCP-CRC 简化版本的基础上, 基于 SRM 准则设计一个“两步化”的简化算法 CSVM-CRC. 最后基于 10 个 Benchmark 数据集通过实验研究其分类性能.

进一步的研究方向包括:

1) 与 BCP-CRC 的“两步化”算法 CSVM-CRC 不同, 需要研究 BCP-CRC 的直接实现方法, 即同时求解模型参数和最优拒识阈值;

2) 根据领域知识确定误差代价和非对称拒识代价的具体数值, 进而实现 BCP-CRC 及其实现算法在特定领域的应用;

3) 在误差代价和拒识代价均依赖于类别的情况下研究科学问题的抽象和具体的算法实现;

4) 寻找不同的分类可靠性计算方法和不同的后验概率估计方法及其对分类性能的影响.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1999: 20-21.

[2] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Knowledge Discovery and Data Mining, 1998, 2(2): 121-167.

[3] 郑恩辉. 基于支持向量机的代价敏感数据挖掘研究与应用[D]. 杭州: 浙江大学信息学院, 2006.
(Zheng E H. Cost sensitive data mining based on support vector machines: Theories and applications[D]. Hangzhou: College of Information, Zhejiang University, 2006.)

[4] Chow C K. On optimum recognition error and reject tradeoff[J]. IEEE Trans on Information Theory, 1970, 16(1): 41-46.

[5] Foggia P, Sansone C, Tortorella F, et al. Multiclassification: Reject criteria for the Bayesian combiner[J]. Pattern Recognition, 1999, 32(8): 1436-1447.

[6] Claudio De Stefano, Carlo Sansone, Mario Vento. To reject or not to reject: That is the question-an answer in case of neural classifiers[J]. IEEE Trans on Systems, Man and Cybernetics, 2000, 30(1): 84-94.

[7] Giorgio F, Fabio R. Cost-sensitive learning insupport vector machines[DB/OL]. (2002-03-05)[2011-08-23]. http://www.diee.unica.it/informatica/en/publications/papers-prag/rel-conference-06.pdf, 2002.

[8] Thomas C Landgrebe, David M Tax, Pavel Paclik, et al. The interaction between classification and reject performance for distance-based reject-option classifiers[J]. Pattern Recognition Letters, 2006, 27(8): 908-917.

[9] Zheng En-hui, Zou Chao, Sun Jian, et al. SVM-based credit card fraud detection with reject cost and class-dependent error cost[C]. The 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Bangkok, 2009: 50-58.

[10] Elkan C. The foundation of cost-sensitive learning[C]. Proc of the 17th Int Joint Conf on Artificial Intelligence. Washington, 2001: 239-246.