第 21 卷 第 4 期　　控 制 与 决 策　　2006 年 4 月
Vol. 21 No. 4　　Control and Decision　　Apr. 2006

文章编号: 1001-0920(2006)04-0473-04

代价敏感支持向量机

郑恩辉, 李　平, 宋执环
(浙江大学 a. 工业控制技术国家重点实验室, b. 工业控制技术研究所, 杭州 310027)

摘　要: 以分类精度为目标的传统分类算法通常假定: 每个样本的误分类具有同样的代价且每类样本数大致相等. 但现实数据挖掘中该假定不成立时, 这些算法的直接应用不能取得理想的分类和预测. 针对此缺陷, 并基于标准的 SVM, 通过在 SVM 的设计中集成样本的不同误分类代价, 提出代价敏感支持向量机(CS-SVM)的设计方法. 实验结果表明 CS-SVM 是有效的.

关键词: 分类; 支持向量机; 代价
中图分类号: TP18　　**文献标识码**: A

Cost Sensitive Support Vector Machines

ZHENG En-hui, LI Ping, SONG Zhi-huan
(a. National Laboratory of Industrial Control Technology, b. Institute of Industrial Process Control, Zhejiang University, Hangzhou 310027, China. Correspondent: ZHENG En-hui, E-mail: ehzheng@iipc.zju.edu.cn)

Abstract: Classical methods of designing classifier generally pursue more highly accuracy based on the assumption that all misclassifications have the same cost and the sample number of each class is approximately equal. However, the assumption is not valid in some real applications such as fraud detection and medical diagnosis, so that classification algorithms without taking different misclassification cost into account do not perform well. Based on standard support vector machines (SVM), the algorithm of cost-sensitive SVM (CS-SVM) is proposed by integrating misclassification cost of each sample into standard SVM. Experimental results show that CS-SVM is effective.

Key words: Classification; Support vector machine; Cost

1　引　　言

分类是数据挖掘和机器学习等领域的重要内容之一, 传统的分类算法通常以精度为优化目标, 假定每个样本的误分类代价相等而致力于提高其运行效率和泛化能力[1]. 然而, 在故障诊断、欺诈检测和医疗诊断等领域, 上述假定通常不成立, 忽略样本不同误分类代价的传统分类算法有时不能满足现实数据挖掘的要求[2~5]. 以医疗诊断为例, 把"病人"误诊为"健康人"的代价与把"健康人"误诊为"病人"的代价是不同的. 前者使"病人"失去治疗的机会, 以病情恶化甚至生命为代价, 后者以再次诊断或药物的副作用为代价, 显然前者的误分类代价要大于后者. 若样本集包含 2 个"病人", 98 个"健康人", 分类器把所有样本划分为"健康人"即可得到 98 % 的分类精度, 但这样的诊断不能识别出"病人", 是没有意义的. 在这种情况下, 设计分类器时要考虑样本的不同误分类代价, 实现代价敏感挖掘(CSM). CSM 是普适机器学习所面临的挑战, 研究较少[2].

本文通过在 SVM 的设计中集成样本的不同误分类代价, 提出基于支持向量机(SVM)的 CSM 算法 CS-SVM. 实验结果表明了 CS-SVM 的有效性.

2　代价敏感挖掘

当样本的误分类代价不相等时, 基于精度的传统分类算法通常不能直接用于 CSM 问题. 在数据

收稿日期: 2005-02-02; **修回日期**: 2005-04-25.
基金项目: 国家 863 计划基金项目(2002AA412010-12).
作者简介: 郑恩辉 (1975—), 男, 辽宁新民人, 博士生, 从事人工智能、数据挖掘等研究; 李平 (1954—), 男, 广西北流人, 教授, 博士生导师, 从事工业过程模型化、智能控制等研究.

挖掘和机器学习领域, CSM 的实现集中为以下两类方法: 1) 重构训练样本集, 使用精度最优的标准分类算法实现 CSM; 2) 使用原有样本集, 重新设计分类器, 使其本身可以实现 CSM.

对第 1 类方法, 重构训练样本集的方法包括[5] 复制样本集中小类别样本和删除部分大类别样本, 缺点是丢失部分有用样本的信息或增加计算和存储开销, 且复制正例样本易引起过学习. 文献[5]依据每类样本的误分类代价给训练样本加权, 然后按权值采样重构训练集.

对第 2 类方法, 文献[3]研究了误分类代价和类分布对决策树的分裂标准及剪枝方法的影响, 直接使用误分类代价来判断分类器的性能. 文献[4]依据代价, 把代表重要性的权值集成进 AdaBoost, 在第 1 次迭代中按此权值进行采样.

本文的研究属于第 2 类方法, 通过在 SVM 的设计中集成样本的不同误分类代价, 提出代价敏感支持向量机(CS-SVM)的设计, 用以解决 CSM 问题.

3 支持向量机

在模式分类中, 与传统算法的经验风险最小化准则不同, SVM 使用结构风险最小化准则构造决策超平面, 控制模型复杂度和经验风险的平衡, 提高了算法的泛化能力[6, 7].

对两类问题, 假定已知观测样本集

$$(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_n, y_n),$$
$$x_i \in R^l, y_i \in \{+1, -1\}, i = 1, \cdots, n \quad (1)$$

能被超平面 $(w \cdot x) - b = 0$ 分类, 学习问题为最小化目标函数

$$R(w, \xi) = \frac{1}{2}\|w\|^2 + C\Big(\sum_{i=1}^{n}\xi_i\Big),$$
$$\text{s.t.}\quad y_i(x_i \cdot w + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \cdots, n. \quad (2)$$

目标函数中各参数的意义和优化问题的求解见文献[6, 7].

4 代价敏感支持向量机

尽管标准 SVM 是一种强大的模式分类算法, 但它依然是基于精度的, 不能直接用于 CSM 问题. 鉴于不同样本的误分类具有不同的代价, 本文把样本的不同误分类代价集成到 SVM 的设计中, 以经验代价和结构代价的线性和最小为优化目标设计 CS-SVM.

考虑每个样本都有不同的误分类代价, 样本集(1) 可重构为

$$(x_1, y_1, co_1), \cdots, (x_i, y_i, co_i), \cdots, (x_n, y_n, co_n),$$
$$x_i \in R^l, y_i \in \{+1, -1\},$$
$$co_i \geq 0, i = 1, \cdots, n. \quad (3)$$

其中 co_i 为第 i 个样本 x_i 的误分类代价, 为正常数, 它依赖于 x_i 或 y_i. 设样本集(3) 能被超平面 $(w \cdot x) - b = 0$ 分类, 那么基于 SVM 的 CSM 问题为最小化目标函数

$$R(w, \xi) = \frac{1}{2}\|w\|^2 + C\Big(\sum_{i=1}^{n} co_i\xi_i\Big),$$
$$\text{s.t.}\quad y_i(x_i \cdot w + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \cdots, n. \quad (4)$$

其中: $\|w\|^2$ 为结构代价, 代表模型复杂度; $\sum_{i=1}^{n} co_i\xi_i$ 为经验代价; C 为松弛因子, 控制结构代价和经验代价之间的平衡. 与式(3) 不同, 函数(4) 的经验代价考虑到不同样本的误差具有不同的误分类代价. 为求解优化问题(4), 构造如下 Lagrange 方程:

$$L_p = \frac{1}{2}w \cdot w + C\sum_{i}^{n} co_i\xi_i - \sum_{i}^{n}\hat{\alpha}_i\{y_i(x_i \cdot w + b) - 1 + \xi_i\} - \sum_{i}^{n}\hat{\mu}_i\xi_i. \quad (5)$$

其中: $\hat{\mu}_i \geq 0$ 和 $\hat{\alpha}_i \geq 0$ 为 Lagrange 系数. 最小化式(5), 分别令

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{n}\hat{\alpha}_i y_i x_i = 0,$$
$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{n}\hat{\alpha}_i y_i = 0,$$
$$\frac{\partial L_p}{\partial \xi_i} = co_i C - \hat{\alpha}_i - \hat{\mu}_i = 0. \quad (6)$$

将式(6) 代入式(5), 得到优化问题的对偶形式

$$L_D = \frac{1}{2}\sum_{i,j=1}^{n}\hat{\alpha}_i\hat{\alpha}_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n}\hat{\alpha}_i,$$
$$\text{s.t.}\quad \sum_{i=1}^{n}\hat{\alpha}_i y_i = 0,$$
$$0 \leq \hat{\alpha}_i \leq co_i C, i = 1, \cdots, n. \quad (7)$$

求解二次规划(7), 得到的 Lagrange 系数 $\hat{\alpha}_i$, 带入式(6) 得最优权值

$$\hat{w}_o = \sum_{\text{支持向量}}\hat{\alpha}_i y_i x_i. \quad (8)$$

定义对应于 $0 < \hat{\alpha}_i \leq co_i C$ 的样本 x_i 为代价敏感支持向量(CS-SV). 有两种类型的 CS-SV 来决定 CS-ODH 的位置. 当 $0 < \hat{\alpha}_i < co_i C$ 时, 相应的 CS-SV 落在超平面的间隔上; 当 $\hat{\alpha}_i = co_i C$ 时, 相应的 CS-SV 为误分类的样本.

最优偏置 $\hat{b}_o$ 按如下方法确定, 由 KKT 条件得:

$$\bar{\alpha}_i(y_i(\bar{w}\cdot x_i+\bar{b})-1+\bar{\xi}_i)=0,$$
$$\mu_i\bar{\xi}_i=0,\ i=1,\cdots,n. \tag{9}$$

取训练集中满足 $0<\hat{\alpha}_{o,i}<co_iC$ 的任意一个样本 (x_i,y_i,co_i), 结合式(6) 和式(9) 得到最优偏置 $\hat{b}_o$.

定义代价敏感最优决策超平面(CS-ODH) 为误分类代价最小的超平面, 相应的决策函数定义为代价敏感决策函数(CS-DF). 根据优化得到的 $\hat{w}_o$ 和 $\hat{b}_o$, CS-ODH 和相应的 CS-DF 分别为

$$\sum_{\text{支持向量}} y_i\hat{\alpha}_i(x\cdot x_i)+\hat{b}_o=0,$$
$$f_o(x)=\operatorname{sign}\Big(\sum_{\text{支持向量}} y_i\hat{\alpha}_i(x\cdot x_i)+\hat{b}_o\Big). \tag{10}$$

以上为线性情况, 引入核函数得非线性情况下的 CS-ODH 和相应的 CS-DF 分别为

$$\sum_{\text{支持向量}} y_i\hat{\alpha}_iK(x\cdot x_i)+\hat{b}_o=0,$$
$$f_o(x)=\operatorname{sign}\Big(\sum_{\text{支持向量}} y_i\hat{\alpha}_iK(x\cdot x_i)+\hat{b}_o\Big). \tag{11}$$

CS-SVM 和 SVM 的主要不同之处是: 1) 由于引入代价 co_i, SVM 中具有同样 Lagrange 系数的样本在 CS-SVM 中可能表示不同类型的支持向量; 2) 与 SVM 相比, CS-SVM 考虑了样本不同的误分类代价, 可直接应用于 CSM 问题

5　实验结果

5.1　二维虚拟数据实验

为观察 ODH 和 CS-ODH 的位置, 随机生成二维两类样本 X. 设类 1 样本的误分类代价为 1, 类 -1 的为 3. SVM 和 CS-SVM 不使用核函数, C 值设为 100, 实验结果见图 1.

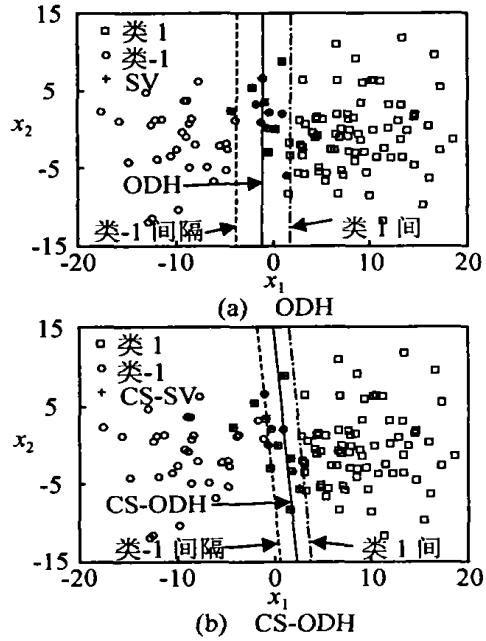

图 1　样本 X 上的 ODH 和 CS-ODH

由图 1 可见, 相比于 ODH, CS-ODH 向右偏移, 靠近误分类代价较小的类 1 样本, 尽管总的误分类数有所提高, 但由于提高了误分类代价较高的类 -1 样本的分类精度, 使误分类的代价大大降低

5.2　Benchmark 数据实验

采用两个包含代价矩阵的 Benchmark 数据集 German Credit 和 Heart Disease, 两个数据集的符号属性被转换为整型, 样本数较少的"欺诈"和"病人"样本称为正例, 相应的另一类别称为反例, 其他信息参见文献[8]. 对两个数据集, 代价矩阵设定如下: 正确分类的代价为 0, 误分类反例的代价为 1, 误分类正例的代价为 5. 本实验在计算测试集平均误分类代价时使用上面代价矩阵, 在训练 CS-SVM 时, 取误分类正例的代价为 R, R 从 1 按步长 0.5 递增到 20. 随机选择 German 数据集 1 000 个样本中的 600 个组成训练集, 其余 400 个组成测试集. 随机选择 Heart 数据集 270 个样本中的 200 个组成训练集, 其余 70 个组成测试集. 图 2 为 20 次试验的平均结果

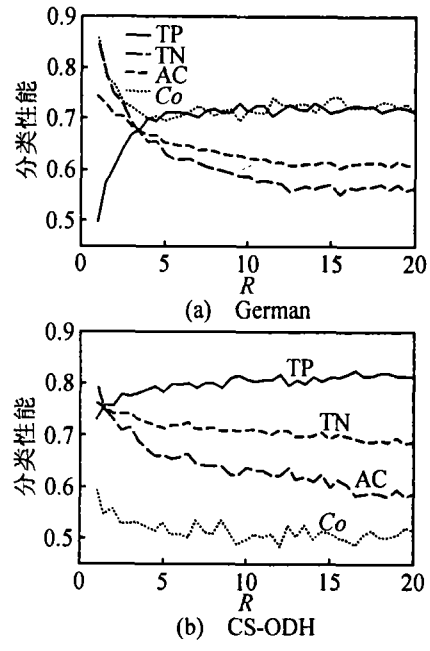

图 2　CS-SVM 的分类性能

图 2 中 TP, TN 和 AC 分别表示正例的分类精度(正确识别的正例数 / 正例的总数)、反例的分类精度和全局精度, Co 表示测试集的平均误分类代价. 对应 $R=1$ 时, CS-SVM 退化为标准 SVM, 即 $R=1$ 对应的性能指标值和 SVM 相同. 随 R 的增大, TP 增加, TN 和 AC 减小, Co 先急剧减小后缓慢增加. 可见, 在一定范围内, 尽管 CS-SVM 降低全局分类精度 AC, 但平均误分类代价也降低, 能够实现 CSM.

6 结 论

当样本的误分类代价不相等时,传统的基于精度的分类器有时不能满足现实数据CSM 的要求. 基于标准的SVM 算法通过把样本的不同误分类代价集成到SVM 的设计中,提出CS-SVM 的设计方法.

CS-SVM 和SVM 的主要不同之处是: 1) CS-SVM 考虑了样本的不同误分类代价,以结构代价和经验代价的线性和最小为目标,能够实现CSM; 2) 由于引入代价 co_i, SVM 中具有同样Lagrange 系数的样本在CS-SVM 中可能表示不同类型的支持向量,它们决定了ODH 和CS-ODH 的不同位置.

实验结果表明,尽管基于CS-SVM 的CS-ODH 和基于SVM 的ODH 相比有较高的误分类数,但平均误分类代价大大降低,实现了CSM.

对CS-SVM,本文给出了一个虚拟数据集和两个Benchmark 数据集的实验结果. 将CS-SVM 应用到大型数据集或者工业过程故障诊断是进一步的主要工作. 另外,把其他的基于精度的分类算法转化为代价敏感挖掘算法也是有意义的研究工作.

参考文献(References)

[1] Han J, Kamber M. *Data Mining: Concepts and Techniques* [M]. San Francisco CA: Morgan Kaufmann, 2001.

[2] 周志华. *普适机器学习*[EB/OL]. http://www.intsci.ac.cn/research/zhouzh04.ppt, 2003.
(Zhou Z H. *Pervasive Machine Learning* [EB/OL]. http://www.intsci.ac.cn/research/zhouzh04.ppt, 2003.)

[3] Drummond C, Holte R. Exploiting the Cost (in) Sensitivity of Decision Tree Splitting Criteria[A]. *Proc of the 17th Int Conf on Machine Learning* [C]. San Francisco, 2000: 239-246.

[4] Fan W, Stolfo S, Zhang J, et al. AdaCost: Misclassification Cost-sensitive Boosting[A]. *Proc of the 16th Int Conf on Machine Learning* [C]. Bled, 1999: 97-105.

[5] Zadrozny B, Langford J, Abe N. Cost-sensitive Learning by Cost-proportionate Example Weighting [A]. *Proc of the 3rd IEEE Int Conf on Data Mining* [C]. Melbourne, 2003.

[6] Vapnik V N. An Overview of Statistical Learning Theory [J]. *IEEE Trans on Neural Networks*, 1999, 10(5): 988-999.

[7] Burges C. A Tutorial on Support Vector Machines from Pattern Recognition [J]. *Data Mining and Knowledge Dlscovery*, 1998, 2(2): 121-167.

[8] Michie, Spiegelhalter D J, Taylor C C. *Machine Learning, Neural and Statistical Classification* [EB/OL]. http://www.ncc.up.pt/liacc/ML/statlog/data.html, 2004.

(上接第 472 页)

参考文献(References)

[1] 张金水. *经济控制论——动态经济系统分析方法与应用* [M]. 北京: 清华大学出版社, 1999: 152-155.
(Zhang J S. *Economical Cybernetics—The Analyzing Method and Application for Dynamic Economic Systems* [M]. Beijing: Tsinghua University Press, 1999: 152-155.)

[2] Lambert J P. *Disequilibrium Macroeconomic Models: Theory and Estimation of Rationing Models Using Business Survey Data* [M]. Cambridge: Cambridge University Press, 1988.

[3] 陈其坤. 非均衡微观市场价格调节的鲁棒控制策略[J]. *厦门大学学报*, 2001, 40(1): 215-218.
(Chen Q K. Robust Control Policies on Price Adjustment for Non-equilibrium Micro-market [J]. *J of Xiamen University*, 2001, 40(1): 215-218.)

[4] 肖冬荣, 陆振宇. 鲁棒控制理论应用于宏观经济系统分析[J]. *控制与决策*, 2002, 17(5): 629-630.
(Xiao D R, Lu Z Y. Analysis of Macroeconomic System Using Robust Control Theory [J]. *Control and Decision*, 2002, 17(5): 629-630.)

[5] 雷勇. 非均衡市场价格调节的纯增益反馈控制问题研究[J]. *中国管理科学*, 2000, 8(1): 51-55.
(Lei Y. Pure Gain Feedback Control Research on Price Adjustment for Disequilibrium Micro Market [J]. *Chinese J of Management*, 2000, 8(1): 51-55.)

[6] Clarke D W, Mohtadi C. Generalized Predictive Control, Part I and part II[J]. *Automatic*, 1987, 23(2): 137-160.

[7] 胡耀华, 贾欣乐. 广义预测控制综述[J]. *信息与控制*, 2000, 29(3): 248-256.
(Hu Y H, Jia X L. Summarization of Generalized Predictive Control[J]. *Information and Control*, 2000, 29(3): 248-256.)