基于 SVM的多类代价敏感学习及其应用

程学云[1,2],吉根林[1],凌霄汉[1]

(1. 南京师范大学 数学与计算机科学学院,江苏 南京 210097; 2. 南通大学 计算机科学与技术学院,江苏 南通 226007)

[摘要] 标准的分类器设计一般基于最小化错误率. 在入侵检测等问题中,不同类型的错分往往具有不等的代价. 通过在支持向量机的类概率输出中引入代价敏感机制,提出了 3种基于最小化总体错分代价设计分类器的方法. 实验结果表明通过改变代价矩阵,能在漏报率、误报率及稀有类样本的错误率之间调节,从而保证在误报率尽可能小的情况下降低漏报率和稀有类样本的错误率,以减少总体错分代价.

[关键词] 代价敏感学习,支持向量机,入侵检测,漏报率,误报率

[中图分类号] TP391 [文献标识码] A [文章编号] 1672-1292(2006)04-0079-04

SVM-Based Multiclass Cost-Sensitive Learning and its Application

CHENG Xueyun[1,2], JI Genlin[1], LIN Xiaohan[1]

(1. School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China;
2. School of Computer Science and Technology, Nantong University, Nantong 226007, China)

Abstract: The standard classifier is usually based on minimizing the error rate, but in intrusion detection and some practical problems, different errors have different costs. Three kinds of support vector machine (SVM) learning methods based on minimizing the total misclassification cost are proposed, which introduce the cost-sensitive mechanism into the probabilistic outputs of SVM. The results show that we can trade off among false negative, false positive and error rate of rare class by changing cost matrix, which can minimize false negatives and error rate of rare class while constraining false positives at a low level so as to minimize the total misclassification cost.

Key words: cost-sensitive learning, support vector machine (SVM), intrusion detection, false negative, false positive

0 引言

在机器学习和数据挖掘中,分类精度是衡量分类器性能的一个重要指标,这往往与假设不同类型的错分代价相同. 而在实际应用中,如入侵检测,将攻击漏报为正常和将正常误报为攻击,所引起的代价是截然不同的. 因此,需要对不同类型的误分引入不同的惩罚代价,使分类器的设计目标由原来的最小化错误率变为最小化总体错分代价.

代价敏感学习作为机器学习领域的一个新的研究热点,正是研究了对不同类型的错分引入不同的惩罚代价,并研究在什么机制下保证得到的分类器使总体错分代价最小. 基于结构风险最小化理论的 SVM,以其维数无关、避免过拟合、全局最优、分类精度高等特性在分类领域得到越来越广泛的应用. 传统的 SVM是基于最小化错误率实现的,在错分代价不相等问题中不能达到最优. 为此本文探讨了如何在多类 SVM的学习中实现代价敏感,并将其用于入侵检测问题. 在文献[1]中概括了 C-SVM和 ν-SVM的代价敏感扩展 2C-SVM和 2ν-SVM,并探讨了它们之间的联系. 这两种方法是通过对正类和负类的松弛变量 ξ_i 分别引入不同的惩罚系数 C^+和 C^-实现的. 本文提出了 3种使 SVM代价敏感的方法. 其一是在 Libsvm[2]提

收稿日期: 2006-06-01.
基金项目: 江苏省自然科学基金资助项目(BK2005135)和南通大学校级自然科学研究基金资助项目(05Z053).
作者简介: 程学云(1978-),女,助教,硕士研究生,主要从事数据挖掘、模式识别的教学与研究. E-mail: chen_xy@ntu.edu.cn
通讯联系人:吉根林(1964-),教授,博士生导师,主要从事数据库、数据挖掘与入侵检测技术的教学与研究. E-mail: glji@njnu.edu.cn

供的类概率输出中根据错分代价矩阵对输出加权,称为 OW-SVM方法;其二是基于 Bayes风险最小化的思想,将样本划分到总体错分代价最小的类;其三是基于 Meatcost[3]方法,根据最小化总体错分代价的思想,改变训练样本的类别标签,再利用新的类别标签进行训练,使训练出的分类器达到减小总体错分代价的目的,称之为 MC-SVM方法. 将 3种方法在 8种不同代价矩阵下做了对比实验,结果表明可以通过调节代价矩阵,实现在漏报率、误报率和高代价样本的错误率之间调节,并且在 3种方法中,B-SVM方法效果较好.

1 基于 SVM的代价敏感学习

1.1 代价敏感学习

在实际问题中,不同类型的错分所引起的代价不一定相同,所以基于最小化错误率原则设计的分类器,在错分代价不等问题中不能满足总体错分代价最小的要求. 代价敏感学习正是考虑了不同类型错分的代价,并基于最小化总体误分代价的原理来设计分类器,从而能更好的满足错分代价不同的情形.

在代价敏感学习中,假设有 c类样本,定义代价矩阵为 $\mathrm{cost}[i, j]$,表明将第 j类样本错分为第 i类的代价,其中 $\mathrm{cost}[i, i] = 0$,即正确分类的代价为 0. 定义 $\mathrm{cost}[j]$为第 j类样本的期望代价:

$$\mathrm{cost}[j] = \sum_{i=1}^{c} \mathrm{cost}[i, j] \tag{1}$$

目前常用的代价敏感学习方法有:估计出样本的类概率后,通过 Bayes理论将样本划分到风险最小的类别中;改变样本的原始分布,如过采样、欠采样方法,将已有的任何分类算法转化为代价敏感学习算法;使某种基于最小化错误率的学习算法获得代价敏感性,如基于神经网络的方法[4]. 本文提出了基于 SVM的代价敏感方法.

1.2 支持向量机

设样本集为 $(x_1, y_1), \cdots, (x_l, y_l)$, $x_i \in \mathbf{R}^n$, $y_i \in \{-1, +1\}$, $i = 1, 2, \cdots, l$ 当样本线性可分时,找到一个最优超平面 $w \cdot x + b = 0$将两类点分离,且使两类点间的间距 $2 / \| w \|^2$最大,以得到最强的泛化能力. 为了允许少量样本错分,引入了松弛因子 $\xi_i \geq 0$, $i = 1, 2, \cdots, l$,用 $\sum_{i=1}^{l} \xi_i$表示样本允许被错分的程度,且使其最小,即求解最优化问题:

$$\begin{aligned} &\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} \xi_i \\ &\text{s.t.} \quad y_i ((w \cdot x_i) + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \quad i = 1, 2, \cdots, l \end{aligned} \tag{2}$$

式中,C为惩罚系数,在错分样本与模型复杂性之间的折衷. 在文献[4]提出的代价敏感方法中,分别用 C^+和 C^-对正类错分为负类和负类错分为正类进行惩罚,从而实现对不同的错分以不同的代价惩罚,一般用较高的代价惩罚最不期望得到的误分. 决策面向误分代价相对较低的一类偏移,从而使错分代价高的样本更倾向于被正确分类.

SVM主要适用于两类问题,可通过“一对一”、“一对多”、“DAGSVM”等方法解决多分类问题. 本文用“一对一”方法实现多类,即在每两类间训练一个分类器,因此对于 k类问题,将有 $k(k-1)/2$个分类函数. 对新样本 x进行预测时,用投票方法将 x划分到投票最多的类中. 在文献[5]中提出了多类 SVM的类概率输出方法,并在 Libsvm中得到实现.

1.3 基于输出加权的代价敏感 SVM (OW-SVM)

在 SVM的类概率输出中,对每类的概率输出加权,并将代价矩阵引入权向量中. 设每类样本的个数为 $N[i]$,总的样本个数为 n,W为权向量,W$[i]$表示第 i类样本的权向量,根据公式(1)中定义的 $\mathrm{cost}[i]$,权向量公式为:

$$\mathrm{W}[i] = \frac{\mathrm{cost}[i] \times n}{\sum_{j=1}^{c} \mathrm{cost}[j] \times N[j]} \tag{3}$$

设 SVM的类概率输出为 $l \times c$矩阵 P,其中 l为训练样本个数,c为训练样本类别数,则加权后的概率输出为 $\mathrm{P}'(j \mid x) = \mathrm{P}(j \mid x) \times \mathrm{W}[j]$,最终待测样本所属类别为: $j' = \arg\max_j \mathrm{P}'(j \mid x)$, $j = 1, \cdots, c$

从推导过程可见,当某类的错分代价较高时,最终的输出偏向该类,以低代价样本的错分增加为代价降低了高代价样本的错分,在两类相对平衡的问题中能使总体错分代价减小.

1.4 基于 *Bayes*风险最小化的代价敏感 *SVM (B-SVM)*

Bayes风险最小化公式为:

$$\arg\min_i \sum_j P(j|x)C(i,j) \tag{4}$$

即在样本的类概率输出中引入代价矩阵. 在对样本 x预测时,将其归为错分代价最小的类,从而使总体错分代价最小. 假设 c类样本中,当第 j类的错分代价较高时,即 $C(i,j)\ (j\neq i)$的值远远大于 $C(j,i)\ (j\neq i)$,则分类的结果偏向于第 j类,即偏向于错分代价较高的类.

1.5 基于 *Metacost*的代价敏感 *SVM (MC-SVM)*

文[3]中提出了元代价学习 (Metacost)的思想,通过"元学习"过程,估计出样本的类概率 $P(j|x)$,再根据公式 (4) 修改训练样本的类别标记,使用修改过的训练集重新学习得到新的模型,再用新的模型对测试集进行预测.

2 在入侵检测中的应用

入侵检测作为网络安全中一项积极主动的安全防护技术,越来越受到广泛关注. 本文实验基于KDD99数据集[6],共 41个连接属性和 1个类别属性. 数据分为 5类,分别为正常连接 Normal、DOS攻击、Probe攻击、U2R攻击和 R2L攻击,分别标记为 1、2、3、4、5类. 从数据集中随机选取 20 000条正常连接,1 000条 DOS攻击,1 000条 Probe攻击,50条 U2R攻击和 1 000条 R2L攻击进行预处理,归一化. 再从选取的数据集中随机取 10组同分布的训练集和测试集,分别含 5 000条正常连接,250条 DOS攻击,150条 Probe攻击,20条 U2R攻击和 80条 R2L攻击. 对每种代价矩阵,用如上 3种方法在每组数据集上训练和测试,结果取 10组数据集的平均.

实验中,惩罚系数 C取 1 000, RBF核函数 $e^{(-\gamma\times\|x_i-x_j\|^2)}$ 中 γ取 0.2 设训练集中第 i类样本的频度为 $p(i)$,则在代价矩阵中, $cost(i,i)=0$, $cost(i,j)\ (i\neq j)$为 $(0,(p(i)/p(j))^{1/(11-k)})$之间的随机数,其中代价系数 k取 1, 2, …, 8 可见当第 j类为稀有类时,相应的将 j类错分为其它类的代价较大,从而减少了多类中稀有类的错分. 当 $k=1$时,各类的错分代价比较接近,随着 k值的变大,多数类和稀有类之间的错分代价差距越大,则高代价样本的错分数越来越少. 所以,以该方法产生的代价矩阵,对于入侵检测问题中减少稀有攻击类的错分是有效的.

测试集在 8种代价矩阵下的错误率和误报率如图 1、图 2所示.

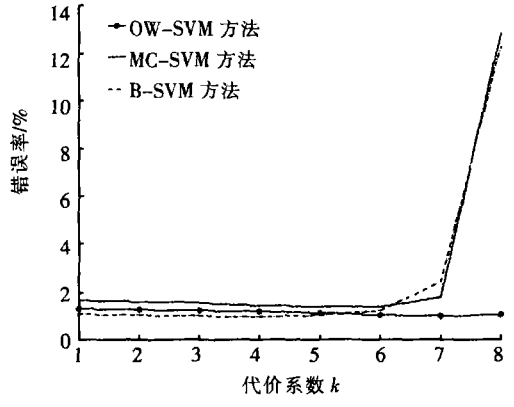

图 1 测试集在 8种代价矩阵下的错误率

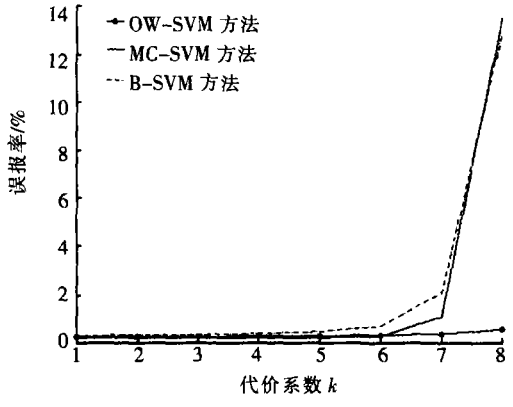

图 2 测试集在 8种代价矩阵下的误报率

从图 1和图 2可见,在前几种代价矩阵下结果差别不大,且错误率和误报率都很低,但随后错误率和误报率急剧增加,且错误率主要受误报率的影响.

测试集在 8种代价矩阵下的漏报率及稀有类 (U2R类)在 8种代价矩阵下的错误率分别如图 3和图 4所示.

从图 3和图 4可见,当 $k=1$时,稀有类样本的错误率非常高,但随着 $(p(i)/p(j))^{1/(11-k)}$中 k值的增加,

漏报率和稀有类样本的错误率均明显下降. 在同种代价矩阵下, B-SVM方法的漏报率及高代价样本的错报率较低.

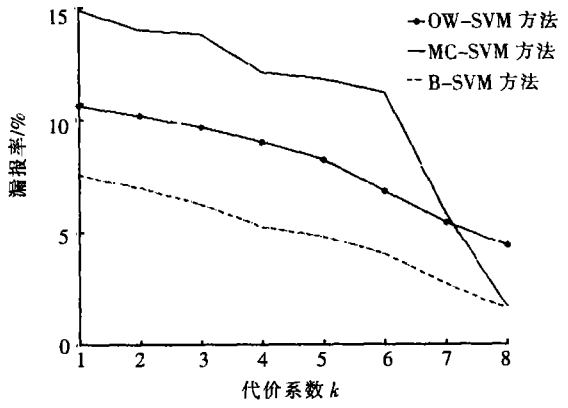

图 3 测试集在 8种代价矩阵下的漏报率

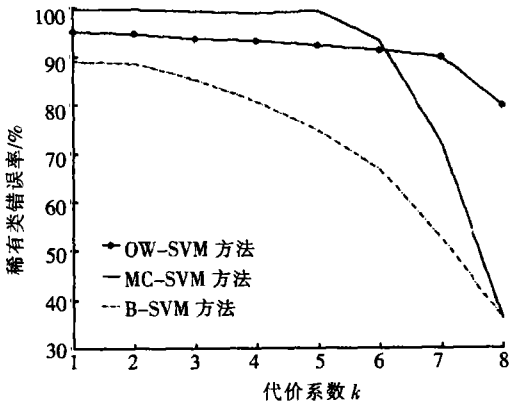

图 4 稀有类(U2R 类)在 8种代价矩阵下的错误率

代价敏感学习的主要目的是最小化总体错分代价或平均错分代价,对于 3种方法,在 8种代价矩阵下的平均错分代价如图 5所示.

在入侵检测中,各类分布极不平衡. 如本实验中,攻击样本个数相对较少,所以当稀有类样本由于错分代价较高而错分个数减少时,会以多数类样本错分个数的大量增加为代价. 所以对样本极不平衡的问题,总体错分代价不一定减少,但能够有效的降低漏报率和高代价样本的错误率. 由图 5可知,从最小化总体错分代价考虑,在相同代价下, B-SVM方法的性能较优. 在实际问题中,可以根据 Neyman-Pearson学习理论[7]调节代价矩阵,将误报率限制在某一个可接受的范围,在此范围下保证漏报率和高代价样本的错误率较小.

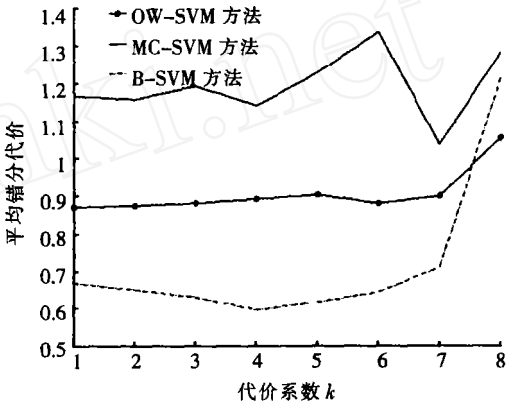

图 5 测试集在 8种代价矩阵下的平均错分代价

3 结语

本文提出了 3种基于 SVM的代价敏感学习方法,并在入侵检测中得到应用. 实验表明,通过改变代价矩阵,可在漏报率、误报率及高代价样本的错误率之间进行调节,并尽量减少总体错分代价.

[参考文献] (References)

[1] MARK A DAVENPORT. The 2nu - SVM: A Cost-Sensitive Extension of the nu - SVM [R]. Rice University ECE Technical Report TREE 0504, 2005.

[2] CHANG Chihchung, LIN Chihjen. LIBSVM: A library for support vector machines [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2005.

[3] DOMINGOS P. MetaCost: a general method for making classifiers cost-sensitive [C]// Proc of the 5th International Conference on Konwledge Discovery and Data Mining. San Diego: ACM Press, 1999. 155-164.

[4] OSUNA E, FREUND R, GIROSI F. Support vector machines: Training and applications [R]. AIMemo 1602, MIT AI Lab, 1997.

[5] WU Tingfan, LIN Chihjen, RUBY C W eng. Probability estimates for multi-class classification by pairwise coupling [J]. Journal of Maching Learning Research, 2004 (5): 975-1 005.

[6] KDD Cup 1999 Data [DB/OL]. [1999 - 10 - 28] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[7] SCOTT C, NOWAK R. A neyman-pearson approach to statistical learning [J]. IEEE Transactions on Information Theory, 2005 (51): 3 806-3 819.

[责任编辑:刘 健]