文章编号: 1001-0920(2004)04-0478-03

异常检测中查全率和查准率的控制

陈光英, 张千里, 李 星
(清华大学 电子工程系, 北京 100084)

摘 要: 在使用支持向量机分类技术的异常检测系统中, 提出控制查全率和查准率的方法. 该方法采用遗传算法优化特征选择和训练模型, 其中染色体由特征选择和训练模型组成, 适应度是用 F$_\beta$-estimate 方法计算的查全率和查准率的组合, 通过设置其中一个参数 η 达到控制查全率和查准率的目的. 实验中采用异常检测标准数据分析该方法的使用效果, 结果表明随着 η 增大, 查全率也增大, 而查准率却减小, 使得用户可以通过设置 η 的值控制查全率和查准率.

关键词: 支持向量机; 查全率; 查准率; 优化; 控制
中图分类号: TP18; TP39 **文献标识码**: A

Recall and precision control in anomaly detection

TRAN Quang-Anh, ZHANG Qian-li, LI Xing
(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Correspondent: TRAN Quang-Anh, E-mail: qa00@mails.tsinghua.edu.cn)

Abstract: An approach to control the recall and precision in an anomaly detection system is presented using support vector machine (SVM). Genetic Algorithm (GA) is used to optimize the feature set and to train the model of SVM. The chromosome in GA consists of feature selection and training model. The fitness of chromosome is the formulation of recall and precision which are computed by the F$_\beta$-estimate method. A parameter is given to control the recall and precision in the formulation. The experiment results show that when the parameter increases, the recall increases and the precision decreases. It permits users to control the recall and precision by setting the value of the parameter.

Key words: support vector machine; recall; precision; optimization; control

1 引 言

查全率和查准率是入侵检测的重要性能[1], 任何一个入侵监测系统都希望具有高的查全率和查准率, 但两者难以同时都满足. 入侵检测方法分为误用检测和异常检测两种[2]. 异常检测通常使用机器学习分类技术进行入侵检测, 因而本文选择支持向量机(SVM)[3]作为异常检测的分类技术, 并提出一种控制查全率和查准率的方法, 允许用户能够简单地控制异常检测系统的查全率和查准率.

2 控制方法

本文使用特征选择和训练模型的联合优化方法提高 SVM 性能. 优化方法使用遗传算法(GA)[4,5], 并利用 SVM 性能作为反馈信息来优化特征选择和训练模型. 在优化过程中, 通过控制染色体适应度的表达式来达到优化后查全率和查准率的控制目的. 该方法的流程如图 1 所示.

收稿日期: 2003-01-14; **修回日期**: 2003-05-30.
基金项目: 国家自然科学基金资助项目(90104027); 国家 863 计划资助项目(2001AA 112041).
作者简介: 陈光英(1974—), 男, 越南人, 博士生, 从事网络安全、人工智能的研究; 李星(1957—), 男, 北京人, 教授, 博士生导师, 从事信号与信息处理、信息网络理论及其应用等研究.

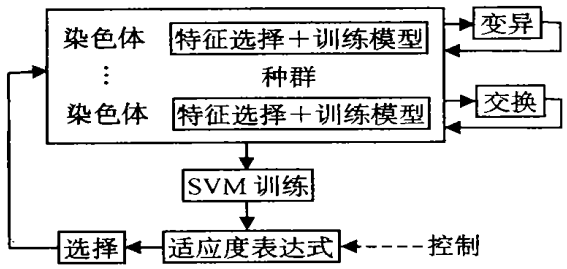

图 1　控制性能流程

图中表示在 GA 方法中, 种群由多个染色体组成, 染色体由特征选择和训练模型表示. 染色体通过 SVM 训练计算其适应度. GA 的交换和变异流程不变, 在选择流程中(即适应度表达式) 加以控制.

设 R 和 P 为查全率和查准率. 由于提高查全率常会导致查准率的降低, 为了控制 R 和 P 之间的折衷, 本文引用一个控制参数 $\eta(0 \leq \eta \leq 1)$, 并用 R 和 P 的组合来描述染色体的适应度, 即

$$G = \eta R + (1 - \eta) P. \qquad (1)$$

实际上 R 和 P 的组合就是对 R 和 P 加上不同权重, 优化过程就是寻找使 $\eta R + (1 - \eta) P$ 的值最大的特征选择和训练模型. 通过优化过程, 最优解不但能使 SVM 有较高的 G, 而且 R 和 P 之间的折衷也应如同所期望的.

3　实现

实现过程要解决: 如何描述特征选择、训练模型及如何快速计算 SVM 性能等问题.

3.1　特征选择

本文所谓的特征选择就是如何从特征集中挑选一个最佳的子集. 给定特征集 $F = \{f_1, f_2, \cdots, f_N\}$, 特征选择可以用一个二进制向量来表示, 即

$$S = \{s_1, s_2, \cdots, s_N\}, s_i \in \{0, 1\}. \qquad (2)$$

其中 S 中的每一位 1 和 0 分别表示 F 中相应位子的特征被选中与否.

3.2　训练模型

在训练 SVM 之前必须设置一些参数, 用这些参数来控制训练过程. 有些参数控制训练算法的运行, 有些参数控制 SVM 的推广能力. 本文所说的训练模型就是后者, 其中包括核函数 K 和折衷参数 C. 本文使用混合核函数[6]:

$$K_{mix} = \lambda K_{poly} + (1 - \lambda) K_{rbf}, 0 \leq \lambda \leq 1. \qquad (3)$$

式中的 K_{poly} 和 K_{rbf} 的表达式如下:

$$K_{poly}(u, v) = (u \cdot v + 1)^d,$$

$$K_{rbf}(u, v) = \exp(- \| u - v \|^2 / \sigma). \qquad (4)$$

另外, Osuna[7] 提出在不均衡的数据集中, 对于不同的类型使用不同的 C 值可以提高 SVM 性能. 对于一般的二类分类问题, 本文使用两个参数 C_+ 和 C_-. 结合核函数和 C, 本文的训练模型如下:

$$M = \{\lambda, d, \sigma, C_+, C_-\}.$$

3.3　计算 SVM 性能

本文使用由 Joachims 提出的对 SVM 性能的估算方法 $\xi\alpha$-estimate[8]. 该方法的优点是只要通过 SVM 的一次训练, 估算用 Leave-one-out 方法[8] 计算得到的分类错误个数的上界, 从而估算查准率和查全率的下界 $R_{\xi\alpha}$ 和 $P_{\xi\alpha}$. $R_{\xi\alpha}$ 和 $P_{\xi\alpha}$ 的计算在文献[8] 中有详细描述. 用 $\xi\alpha$-estimate 方法计算出来的适应度 $G_{\xi\alpha}$ 是实际适应度 G 的下界.

4　实验研究

4.1　数据源

本实验使用由美国国防部高级研究计划机构 (DARPA) 资助的 1999 年知识发现与数据挖掘竞赛提供的一个异常检测的标准数据集[9]. 该数据集包括训练集和测试集, 分别含有大约 50 万和 30 万条数据, 且每条数据包括 41 个特征. 这些数据分别为正常、攻击及攻击类型. 训练集和测试集的攻击类型数目分别为 22 和 37. 本文着重研究查全率和查准率, 因此用攻击和正常两大类来标志数据类型(第 1 类为攻击). 本文使用 SVM ^{light} 源代码[10] 训练 SVM, 并计算查全率和查准率的估计值 ($R_{\xi\alpha}$ 和 $P_{\xi\alpha}$).

为了使遗传算法有效地搜索, 本文缩小优化空间, 即将训练模型 M 的参数约束如下: $0 \leq d \leq 10$; $0 \leq \sigma \leq 20$; $0 \leq C_+, C_- \leq 100$. 遗传算法的种群大小为 200, 并随机初始化, 交换参数 $P_c = 0.3$, $P_m = 0.1$. 这些设置都是经验值.

设折衷参数 η 为 0～1 之间不同的值. 对于每个 η, 本文对训练模型进行优化. 经过 50 代种群后, 计算 $R_{\xi\alpha}$ 和 $P_{\xi\alpha}$. $R_{\xi\alpha}$ 和 $P_{\xi\alpha}$ 随着 η 值变化曲线如图 2 所示.

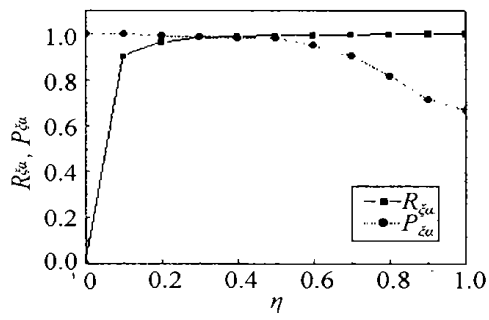

图 2　$R_{\xi\alpha}$ 和 $P_{\xi\alpha}$ 随 η 的曲线

图2表明设置η可以控制查全率和查准率. η越大, $R_{\xi\alpha}$越大而$P_{\xi\alpha}$越小. 当η从0升到1, $R_{\xi\alpha}$也从0升到1, 而$P_{\xi\alpha}$从1减到0.7左右. 可以认为, 当$\eta=0$时, 所有的事件都会被检测为正常类型, 因此$R_{\xi\alpha}=0$而$P_{\xi\alpha}=1$; 当$\eta=1$时, 所有的事件都会被检测为攻击类型, 因此$R_{\xi\alpha}=1$而$P_{\xi\alpha}=0.7$左右. 注意到$R_{\xi\alpha}$的上升和$P_{\xi\alpha}$的减小都有一次突变, 分别在η为0～0.2之间和0.7～1之间. 可见, 选择η在0.2～0.7之间可以保证$R_{\xi\alpha}$和$P_{\xi\alpha}$都比较高.

5 结 论

本文提出一种控制SVM分类的查全率和查准率的方法, 并将其应用于异常检测系统中. 该方法利用在训练时SVM性能的可优化性, 通过GA优化SVM性能的同时, 调整反馈信息(即适应度)的表达式, 达到控制SVM的查全率和查准率的性能. 方法描述和实现包括3点: 1) 优化参数是特征选择和SVM训练模型的混合模型; 2) 查全率和查准率由$\xi\alpha$-estimate方法计算; 3) 期望的查全率和查准率由适应度表达式中的参数η控制. 实验结果表明随着η的增大, 查全率也增大而查准率却减小, 这样用户可以通过设置ρ的值来控制查全率和查准率.

参考文献(References):

[1] Lippman R P, Fried D J. Evaluating intrusion detectionsystems: The 1998 DARPA off-line intrusion detection evaluation[A]. *Proc of the DARPA Information Survivability Conf and Exposition*[C]. Hilton Head, 1999. 12-26.

[2] Denning E D. An intrusion detection Model[A]. *Proc of the IEEE Symposium on Security and Privacy*[C]. Oakland, 1986. 118-133.

[3] Vapnik V N. An overview of statistical learning theory[J]. *IEEE Trans on Neural Networks*, 1999, 10(5): 988-999.

[4] Yao X. Evolving artificial neural networks[J]. *Proc of the IEEE*, 1999, 87(9): 1423-1447.

[5] Holland J H. *Adaptation in Natural and Artificial Systems*[M]. Ann Arbor, Univ: Michigan Press, 1975.

[6] Smits G F, Jordaan E M. Improved SVM regression using mixtures of kernels[A]. *Proc of the* 2002 *Inter Joint Conf on Neural Networks*[C]. Honolulu, 2002. 3: 2785-2790.

[7] Osuna E, Freund R, Girosi F. Support vector machines: Training and applications[R]. Massachusetts Institute Technology, 1997.

[8] Joachims T. Estimating the generalization performance of a SVM efficiently[A]. *Proc of the Seventeenth Int Conf on Machine Learning*[C]. 2000. 431-438.

[9] The UCI KDD Archive. KDD Cup 1999 Data[EB/OL]. URL: http://kdd.ics.uci.edu/databases/kdducp99/kddcup99.html, 1999.

[10] Joachims T. SVM^{light} Support Vctor Machine[EB/OL]. URL: http://svmlight.joachims.org, 2002.

(上接第477页)

(YANG Junhui, DAI Zongduo, YANG Dongyi, et al. An elliptic curve signature scheme and an identity-based signature agreement[J]. *J of Software*, 2000, 11(10): 1303-1306.)

[5] 罗皓, 乔秦宝, 刘金龙, 等. 椭圆曲线签名方案[J]. 武汉大学学报(理学版), 2003, 149(11): 095-098.
(LUO Hao, QIAO Qinbao, LIU Jinlong, et al. Signing schedules with elliptic curve cryptography[J]. *J of Wuhan University*, 2003, 149(11): 095-098.)

[6] Microprocessor, Microcomputer Standards Committee of the IEEE Computer Society. IEEE Standard Specifications for Public Key Cryptography[DB/OL]. http://intl.ieeexplore.ieee.org. 2002-01-30.

[7] Sarbari Guta, Stephen M, Matyas J. Public key infrastructure analysis of existing and needed protocols and object formats for key recovery[J]. *Computers and Security*, 2000, 19: 56268.

[8] Park C-S. On certificate-based security protocols for wireless mobile communication systems[A]. *IEEE Network*[C]. 1997. 50-55.

[9] Beller M J, Chang L-F, Yacobi J. Privacy and authentication on a portable communications systems[J]. *IEEE J on Selected Areas in Communications*, 1993, 11(6): 821-829.