基于支持向量机和最小二乘支持向量机的入侵检测比较*)

任勋益[1] 王汝传[1,2] 谢永娟[1]
(南京邮电大学计算机学院 南京 210003)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[2]

摘 要 将支持向量机和最小二乘支持向量机用于入侵检测之中,利用主元分析对数据进行约简,然后使用 SVM 和 LS-SVM 对数据进行训练和测试。基于 KDDCUP'99 做了三组对比实验,对支持向量机和最小二乘支持向量机的性能做了统计。实验结果表明,SVM 比 LS-SVM 分类能力强,但是 LS-SVM 耗时较少。

关键词 支持向量机,最小二乘支持向量机,入侵检测,主元分析

Comparisons of SVM and LS-SVM for Intrusion Detection

REN Xun-yi[1] WANG Ru-chuan[1,2] XIE Yong-juan[1]
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[2]

Abstract This paper utilizes support vector machine and least square-support vector machine for intrusion detection. We normalizae data, reduce the data with principal component analysis, train and test reduced data with support vector and least square support vector machine. We do three experiments on KDDCUP'99 data set, and utilize Receiver Operating Characteristics curves to evaluate classifier's ability of SVM and LS-SVM, and statistic time cost. Experimental results show SVM has more classifying ability than LS-SVM, but LS-SVM spends less time than SVM.

Keywords Support vector machine, Least square support vector machine, Intrusion detection, Principal component analysis

1 引言

自 1980 年 4 月,James P. Anderson 首次提出入侵检测的概念至今,入侵检测技术越来越向智能化的方向发展。神经网络、遗传算法、K-临近算法等相继应用到入侵检测之中。但是这些智能化方法本身具有一定局限性,比如局部最优问题,假定样本无穷、维数低等,这使得其实际应用效果不佳。

支持向量机是 Vapnik 在 1995 年提出的一个新的学习方法,它建立在统计学习与结构风险最小化理论之上,经过小样本学习,能够得到具有良好泛化能力的模型。SVM 最大的优点是它将低维线性不可分问题经过函数映射转化,使得问题在高维空间线性可分,并且采用核函数,不需要寻求映射函数,从而使得计算不依赖于样本维数。SVM 克服了传统机器学习方法的一些不足,具有泛化性好、小样本、全局最优等优点,目前已经成功地应用到人脸识别、语音处理、入侵检测等领域。

入侵检测本质上是一个分类问题,采用 SVM 将网络数据区分为正常数据与异常数据,或者分为多类,从而能达到有效检测入侵之目的。LS-SVM 是 SVM 的一种改进,它构造了新的二次损失函数,并将原支持向量机二次规划问题变为求解线性方程,提高了 SVM 求解速度,但是也丧失了 SVM 的稀疏性优点。本文讨论了 SVM 和 LS-SVM 二类分类技术,将 SVM 与 LS-SVM 应用到入侵检测之中,对它们的检测性能从推广能力和花费时间上进行了比较。

2 分类支持向量机 C-SVM[1,2]

对于训练数据$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中,$x_i\in\mathcal{R}^d$,$y_i\in\{-1,+1\}$,为了求得一个最优分割超平面 $y=W\cdot X+b$,求解如下的凸二次规划问题:

$$\underset{w,b,\xi_i}{\text{minimize}}\ \frac{1}{2}w^Tw+C\sum_{i=1}^{N}\xi_i$$

$$st\quad y_i(w^TX_i-b)+\xi_i-1\geqslant 0,\xi_i\geqslant 0,1\leqslant i\leqslant N \tag{1}$$

其中 w 为超平面的法向量,b 为偏移,而 C 为不完全可分时引入的惩罚参数,ξ_i 为放松约束条件时引入的松弛变量。引入 Lagrange 乘子,对于线性不可分问题引入核函数,原问题转化为对偶问题:

$$\text{maximize}\ \sum_i a_i-\frac{1}{2}\sum_{i,j}a_ia_jy_iy_jK(X_i,X_j)$$

$$s\,t\quad 0\leqslant a_i\leqslant C,\sum_{i=1}^{n}a_iy_i=0,1\leqslant i\leqslant N \tag{2}$$

其中,a_i 为 Lagrange 乘子,$\varphi(X_i)$,$\varphi(X_j)$分别为 X_i,X_j 的映射,$K(X_i,X_j)$为核函数。当 C 为$+\infty$时,没有错误惩罚,相当于原始问题的训练数据完全分开情况。将$\sum_{i=1}^{n}a_iy_i=0$ 带入式(1),得到:$w=\sum_{i=1}^{n}a_iX_iY_i$,根据二次规划优化问题的解满足 KKT 条件,可以求得 $b=y_j-\sum_{i=1}^{n}y_ia_i^*K(X_i,X_j)$,$a_i^*$ 为大于

*)基金项目:国家高技术研究发展计划("863"计划)基金资助项目(2005AA775050),江苏省高技术研究计划基金资助项目(BG2004004、BG2005037、BG2005038)。**任勋益** 讲师,博士,研究方向为信息安全技术、计算机网络和网格计算;**王汝传** 教授,博士生导师,主要研究方向为计算机软件、计算机网络和网格、信息安全、移动代理和虚拟现实技术等;**谢永娟** 硕士生,研究方向为计算机软件、信息安全。

零的系数，因为只有当 $a_i>0$ 时，才对 Q_D 大小有影响，把对应于 $a_i>0$ 的支持向量 X_i 称为支撑向量，从而得到分类函数 $f(X)=\mathrm{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i,X)+b)$，该函数称为支持向量机。

3 最小二乘法支持向量机 LS-SVM[3,4]

对于训练数据$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中，$x_i\in\mathcal{R}^d$，$y_i\in\{-1,+1\}$，为了求得一个最优分割超平面 $y=W\cdot X+b$，将样本分开。LS-SVM 将问题转化为：

$$\underset{w,b,e}{\text{minimize}}\ \frac{1}{2}w^Tw+\gamma\sum_{i=1}^{N}\xi_i^2$$
$$st\quad y_i(w^Tx_i)+b)+\xi_i-1=0,i=1,\cdots,N \tag{3}$$

其中 w,b,ξ_i 与 SVM 含义相同，γ 表示惩罚参数。在此，约束变为了一组等式，目标优化函数中使用了二次项 ξ_i^2，因此 LS-SVM 因此称为最小二乘法支持向量机。引入 Lagrange 乘子，对于线性不可分问题引入核函数，并使用 KKT 条件，原问题转化为求解线性方程组：

$$\begin{bmatrix}0 & 1^T\\ 1 & \Omega+I/\gamma\end{bmatrix}\begin{bmatrix}b\\ a\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{4}$$

其中 $y=y=[y_1;\cdots;y_n]^T$，$a=[a_1;\cdots;a_n]^T$，$1^T=[1;\cdots;1]^T$，Ω 为核矩阵，$\Omega_{ij}=K(x_i,x_j)$。

通过式(2)求得分类函数 $f(X)=\mathrm{sgn}(\sum_{i=1}^{n} y_i a_i K(X_i,X)+b)$。由于 LS-SVM 求解的是线性方程组，因此计算速度快，效率高。

4 基于 SVM，LS-SVM 的入侵检测方法

使用 SVM，LS-SVM 进行入侵检测与使用其他机器学习方法基本上是一样的，它们都是通过对训练数据的学习得到分类模型，利用分类模型对待分类数据进行分类。不同的是，有些分类器输出的是一个概率，如朴素贝叶斯(NB)，神经网络(NN)，而有些分类器输出的就是一个代表类别的数值，比如 C4.5，SVM，LS-SVM 等。无论哪一种分类器都可以将数据分为两类和多类，在本文中我们将利用支持向量机对入侵数据进行二类分类，即输出值是{+1，-1}。基于 SVM 或者 LS-SVM 的入侵检测方法如图 1 所示。

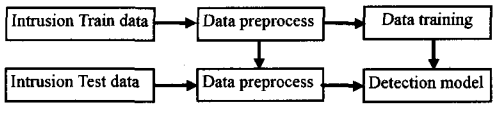

图 1 基于 SVM 的入侵检测方法

在此方法中，对训练数据进行预处理，包括将数据特征数值化，数据归一化，特征选择等。目前 SVM 处理的数据都是数值，因此必须将入侵数据中的文本特征，比如 TCP，Flag 等数值化。数据归一化是为了削除不同数据特征之间的差别，使其在同一个范围类可比较。数据特征选择的目的是为了降低维数，提高学习效率。经过预处理后，利用 SVM 对数据进行训练，训练可以采用不同的参数和使用交叉验证方法。训练最主要的结果是一组拉格朗日乘子$\{a_i^*\}$，利用该乘子可以得到模型：$f(X)=\mathrm{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i,X)+b)$。预测时，预测数据的预处理可以借鉴训练预处理的结果，本文通过主元分析对数据特征降维，得到的结果是一个线性变换矩阵，那么把该矩阵存储起来供测试数据降维。对测试数据进行预处理后，利用得到的模型进行计算，最终得到判断值，如果是+1，则认为正常，否则认为该数据是入侵数据。

5 入侵数据及预处理

本文采用 KDD CUP'99[5]作为实验数据，该数据总体上可以分为两类：正常数据和入侵数据(拒绝服务攻击 DoS，探测攻击 Prob，用户提升权限 U2L，远程非法授权攻击 R2U)。数据的每一条记录包含了从 TCP/IP 连接中抽取的 41 个特征，有 3 个特征(protocal_type，service，flag)是字符变量。我们首先对字符变量处理，比如对 protocal_type 特征，TCP 标识为 1，UDP 标识为 2 等，将所有字符变量处理为数字。其次为了使数据平等处理，需要对任一个特征变量归一化。设 n 为选择的记录总条数，对第 p 条记录第 i 个特征变量 x_{pi}，采用下式归一化到[-1，+1]范围：

$$\tilde{x}_{pi}=2*\frac{x-\min(x_i)}{\max(x_i)-\min(x_i)}-1,i=1,\cdots,41,p=1,\cdots,n \tag{5}$$

其中 $\min(x_i)$为第 i 个特征最小值，$\max(x_i)$为第 i 个特征最大值。训练数据采用 KDD CUP'99 提供的子集 Kddcup.data_10_percent，该数据集每一条记录最后加上了分类标签{±1}。测试数据采用 Corrected 子集。

数据特征选择可以有效提高数据学习和测试的效率，KDDCUP 入侵数据是一个典型的高维异构数据。在此，采用主元分析(Principal Component Analysis，PCA[6])技术对特征进行约简，PCA 是一种线形变换方法，能够在损失较小的情况下对数据维数进行压缩。我们对 Kddcup.data_10_percent 子集共 494021 个样本集 X(其大小为 40×494021，去掉第 21 维，因为其方差为 0，不能用 PCA 处理)进行 PCA 处理，得到 40 个特征值对角阵 $\Lambda=$ diag (9.638，4.719，3.729，2.745，1.995，1.853，1.541，1.194，1.141，1.137，1.016，1.000，0.991，0.958，0.846，0.865，0.848，0.762，0.729，0.481，0.380，0.366，0.339，0.157，0.150，0.135，0.052，0.029，0.021，0.017，0.015，0.015，0.009，0.006，0.006，0.005，0.005，0.001，0.001，0.000)及相应的 40 个特征向量 $V=[V_1,V_2,\cdots,V_{40}]$。选择主元方差贡献率 θ 为 85%，15 个主元方差贡献率已经达到 86.509%，因此，我们选择 15 个特征值。对测试数据 X，作线性变换 $P=V_{15}^T*X$，就可以得到具有 15 个特征的新的训练样本集 $P15\times494021$。

6 实验结果及分析

经过约简后，从约简的数据集中随机抽取数据 4 次，70%用于训练，30%用于测试，进行了 4 组比较实验，实验数据如表 1 所示。

表 1 实验数据

Test	Total data	Normal data	Abnormal data
1	2156	404	1852
2	4230	821	3049
3	5340	1113	4227
4	7156	1391	5765

实验采用 RBF 核、5 倍交叉验证，使用 LIBSVM[7]通过网格搜索得到最优参数，比如对于实验 4，搜索到的最优参数为 $C=32.0,\sigma^2=0.5$，利用最优参数对数据进行训练，得到模型，然后用训练好的模型对测试数据进行测试，如实验 1 得到的测试结果 Accuracy=99.3012%(2842/2862)(classification)，其他的实验结果见表 2。为了更客观地判断测试结果，

我们还采用 ROC[8](Receiver Operating Characteristic Analysis)曲线下的面积 AUC(Area Under the ROC)来衡量分类器的结果,ROC 克服了使用正确率忽略代价的不足,它使用正确分类的比率 TPR (True Positive rate)与错误分类的比率 FPR(False Positive rate)的比值衡量分类器的性能,AUC 是 ROC 曲线下的面积,AUC 越大分类器性能越好,我们对实验 4 使用 LibSVM 的 Plotroc 工具得到 AUC=0.9993,其他的实验结果见表 2。

表 2 实验结果

Test	C-SVM			LS-SVM		
	Test correct	AUC	Test time	Test correct	AUC	Test time
1	99.65%	1.0000	0.5s	84.45%	0.9998	0.2s
2	99.58%	0.9999	0.7s	85.567%	0.9899	0.4s
3	99.53%	0.9998	1.1s	85.652%	0.9890	0.6s
4	99.30%	0.9993	1.4s	86.213%	0.9989	1.0s

从表 2 可以看出,SVM 比 LS-SVM 的检测准确率高,AUC 大,说明 SVM 分类能力强,而 LS-SVM 的检测时间比 SVM 少,说明其检测速度更快。因此,在使用两种方法进行入侵检测时,应该根据实际检测要求选用不同的方法,对于实时性要求高的选用 LS-SVM,准确性要求高的选用 SVM。

结束语 SVM 和 LS-SVM 都是基于结构化风险,克服了传统学习方法的过拟合、局部最小点的缺点,本文将二者用于入侵检测之中,采用 KDDCUP'99 数据集对二者的性能进行了比较。比较发现,SVM 的 AUC 比 LS-SVM 大,但是 LS-SVM 的检测速度更快。这主要是 LS-SVM 采用等式约束条件,克服了 SVM 求解 QP 问题耗时多的缺点,但是另一方面它失去了 SVM 稀疏性的优点。如何克服 LS-SVM 的稀疏性,提高其准确率,以及对 SVM 进行改进,提高其检测效率,是我们下一步要研究的工作。

参 考 文 献

[1] Vapnik V. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995

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

[3] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers [J]. Neural Processing Letters (S1370-4621), 1999, 9(3): 293-300

[4] Suykens J A K. LS-SVMlab Toolbox User's Guide [EB/OL]. http://www.esat.kuleuven.ac.be/sista/lssvmlab/

[5] http://kdd.ics.uci.edu/databases/kddcup99/task.htm

[6] 边肇祺,张学工,阎平凡,等. 模式识别[M]. 北京:清华大学出版社,2000

[7] Lin Chihjen. LIBSVM: a library for SVMs (Version 2.6) [DB/OL]. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf

[8] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 1997, 30 (7): 1145-1159

(上接第 75 页)

$f(90,95)=85.5$, $T_D=f(85.5,80)=68.4$。由于 $T_D<80$(拥有权限 R 的信任阈值),因此 D 不仅不能被授予权限 R,而且委托$(D,E,80)$是没有意义的,从而实现了委托的深度控制。

跟其他的深度控制方案相比,CTBAD 模型的深度控制既简单又灵活实用,很好地控制了委托传递的深度。

5.3 授权冲突问题

设想有这么一种情况:如果从某一权限到某一主体存在多条授权路径,即存在多条证书链,那么可能会存在授权冲突,即对某条路径而言,一致性验证得到通过,而对另外一条路径,一致性验证不通过,此时应该如何来解决。

可以有两种解决方法:一种是选择信任值传递计算结果为最大的那条授权路径进行一致性验证,即在多条授权路径的情形下,只要有一条授权路径通过了一致性验证,就可以通过一致性验证,这是基于对每条授权路径的合法性和合理性的认同而做出的决定。另一种方法是选择信任值传递计算结果为最小的那条授权路径进行一致性验证,即在多条授权路径的情形下,只要有一条授权路径没有通过一致性验证,就不能通过一致性验证,这是基于对高度敏感信息的保护需求而做出的决定。

在具体应用的过程中,可以对两种解决方法有选择地加以利用。如果是针对一般性的非敏感信息,可以采用前一种算法;而对于高度敏感的信息,基于安全原则,应该采用后一种算法。

结束语 在多域环境下,保护被访问资源的安全是一个很重要的问题,很多学者在这个领域进行了深入的研究,已经有了大量的研究成果。但信任的计算以及委托深度控制等问题还没有得到比较好的解决。本文对目前信任管理系统中存在的上述问题进行了一定的研究,提出了一种可计算的基于信任的授权委托模型——CTBAD 模型,重点探讨了 CTBAD 模型的信任计算方法以及信任传递机制。跟目前的信任管理系统相比,CTBAD 模型不仅具有很强的实用性和灵活性,而且实现比较简单。

参 考 文 献

[1] Blaze M, Feigenbaum J, Lacy J. Decentralized trust management // Proceedings of the 1996 IEEE Symposium on Security and Privacy. Washington, DC, USA, 1996: 164-173

[2] Chakraborty S, Ray I. TrustBAC - Integrating Trust Relationships into the RBAC Model for Access Control in Open Systems // SACMAT'06. Lake Tahoe, California, USA, 2006: 49-58

[3] Hong Fan, Zhu Xian, Wang Shaobin. Delegation Depth Control in Trust-management System // Proceedings of the 19th International Conference on Advanced Information Networking and Applications. 2005: 1-4

[4] Chu Yang-Hua, Feigenbaum J, et al. REFEREE: Trust management for Web applications. World Wide Web Journal, 1997, 2 (3): 127-139

[5] Li Ning-Hui, Mitchell J C, Winsborough W H. Design of a role-based trust-management framework // Proceedings of the 2002 IEEE Symposium on Security and Privacy. Oakland, CA, USA, 2002: 114-130

[6] Bertino E, Ferrara E, Squicciarini A C. Trust-X: A Peer to Peer framework for trust negotiations. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(7): 827-842

[7] Freudenthal E, Pesin T, Port L, et al. dRBAC: Distributed role-based access control for dynamic coalition environments // Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCS' 02). Vienna, Austria, 2002: 411-434

[8] 廖俊国,洪帆,朱更明,等. 基于信任度的授权委托模型[J]. 计算机学报,2006,29(8):1265-1270