

基于改进 SVM 主动学习算法的入侵检测

黄双福, 陈贤富

(中国科学技术大学 电子科学与技术系, 安徽 合肥 230026)

摘 要: 入侵检测研究中, 采用基于支持向量机的主动学习算法, 有效地降低了学习的样本复杂度. 针对支持向量机主动学习算法中存在的随机构造的初始训练集样本质量不高和容易陷入次优等问题, 提出了一种结合核空间聚类的初始训练集构建方法, 并在距离准则的基础上引入了概率选择机制. 仿真实验表明, 在不降低检测效果的前提下, 该算法所需的学习样本更少, 并表现出较高的稳定性.

关键词: 入侵检测; 主动学习; 支持向量机; 样本复杂度

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2010)03-0075-03

Intrusion Detection Based on Improved SVM Active Learning

HUANG Shuang-fu, CHEN Xian-fu

(Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: In the study of the intrusion detection, the SVM based active learning algorithm can reduce the sample complexity. Aiming to improve the inferior quality of initial training set resulted from random construction and avoid the propensity to suboptimum in SVM based active learning algorithm, a modified algorithm was proposed. It both incorporates the algorithm of constructing initial training set with kernel-based clustering and the scheme of probabilistic query based on the distance criteria. The experiment result shows that the proposed algorithm needs less samples and is more stable under the same detection performance condition.

Key words: intrusion detection; active learning; support vector machine; sample complexity

1 引言

入侵检测技术是网络安全防御体系的关键技术, 将机器学习方法应用于入侵检测系统^[1], 能够使系统具有较强的适应性、自学习性和鲁棒性, 是目前入侵检测研究的一个重要方向. 然而该方法需要足够的训练数据, 以保证生成的检测模型具有良好泛化能力^[2].

为了在训练样本稀缺的现实网络环境下保证入侵检测的效率, 文献[3]在检测模型中引入了 SVM 主动学习算法. 该算法继承了学习器原来的非线性、推广能力强等特点, 并通过主动选择最优的样本来训练, 降低了学习的样本复杂度. 受随机构造的初始训练集样本质量不高和容易陷入次优等问题的影响, 检测结果不甚理想, 文中提出了一种结合核空间

聚类的初始训练集构建方法, 并在距离准则的基础上引入了概率选择机制. 入侵检测实验表明, 与已有算法相比, 文中算法具有较低的样本复杂度和较高的稳定性.

2 支持向量机主动学习算法

2.1 支持向量机原理

支持向量机(SVM)是一种基于核的学习方法^[3], 对于输入空间线性不可分的训练集, 利用满足 Mercer 条件的核函数将其非线性映射到一个高维特征空间(也称核空间), 使其在该核空间中线性可分, 并构造一个最大间隔超平面作为分类面, 从而获得最强的泛化能力. 设训练集

$$D_l = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$$

式中, $x_l \in R^n$, $y_l \in \{-1, 1\}$, 则在核函数 $K(x, z)$

隐式映射的核空间中,最大间隔超平面可通过求解下面的优化问题获得

$$\begin{aligned} \max \quad & \sum_{i=1}^l a_i - 1/2 \sum_{i,j=1}^l y_i y_j a_i a_j K(x_i, x_j) \\ \text{s. t.} \quad & \sum_{i=1}^l y_i a_i = 0, a_i \geq 0 \end{aligned} \quad (1)$$

求解所得分类超平面方程为

$$\sum_{i=1}^l a_i^* y_i K(x_i, x) + b^* = 0 \quad (2)$$

式中, a_i 为拉格朗日乘因子. 根据优化理论中的 Karush-Kuhn-Tucker 互补条件可知, 只有最靠近超平面的样本点所对应的 a_i^* 非零, 这些点被称为支持向量, 而其他样本点所对应的因子 a_i^* 为零. 这就意味着支持向量包括了重构超平面的所有必要信息, 不是支持向量的点对超平面没有影响^[4-5].

2.2 SVM 主动学习算法

按照训练样本处理方法, 可以将支持向量机的学习方式分为被动和主动两种. 被动学习中, 训练样本是基于独立同分布原则随机选择的^[6], 支持向量机被动地接受这些样本的信息进行学习. 而在主动学习方式下, 训练样本是根据学习进程主动选择的, 其学习过程分为: 初始化阶段和循环学习阶段, 具体描述如算法 1 所示.

算法 1: SVM 主动学习算法.

候选未标注样本集 D_u , 训练样本集 D_l .

(1) 从 D_u 中随机选择少量样本进行类别标注, 由其构成初始训练集 D_l ;

(2) SVM 在 D_l 上训练, 得到一个初始分类超平面;

(3) 根据样本选择准则从 D_u 中选择最能提升分类性能的样本, 标注类别加入 D_l ;

(4) SVM 在更新后的 D_l 上重新训练;

(5) 若算法结束条件满足, 算法终止; 否则返回步骤(3).

样本选择准则是支持向量机实现主动学习的关键, 它关系到分类性能的提高以及标注工作的减少. 现有的做法大都是基于等分版本空间策略^[5], 即支持向量机选择距当前超平面最近的样本进行类别标注, 设 x 为选中的样本, 则满足:

$$\begin{aligned} x = \arg \min_{x \in D_u} \quad & \left| \sum_{i=1}^l a_i^* y_i K(x_i, x) + b^* \right|, \\ x_i \in D_l \end{aligned} \quad (3)$$

由支持向量机的特点可知, 未标注样本离分类面越近就越可能成为新的支持向量, 从而重构和优

化当前分类超平面, 提升分类精度. 这种策略通过选择最有利于分类性能提升的样本来进一步学习, 从而减少训练所需标注样本个数.

2.3 算法的分析与改进

如前文所述, 支持向量机的分类函数面为核空间的一个超平面 H . 设在初始训练集上训练得到的超平面为 H_s , 实际最优的超平面为 H_r , 则 SVM 主动学习过程就是 H 从 H_s 向 H_r 优化调整的过程. 如果 H_s 与 H_r 之间差异较小, 则可以减少后继学习过程中所需的样本数. 在初始训练集规模相同的情况下, 如果初始样本在核空间分布得比较紧密, 则学习得到 H_s 一般只能反映出该局部区域的分类信息, 难以较高的概率保证 H_s 和 H_r 全局相近性. 相反, 如果初始样本在核空间分布比较分散并具有一定的代表性, 则学习所得 H_s 就综合了较多的分类信息, 使得 H_s 与 H_r 的相近性有了进一步的提高. 于是文中提出一种结合核空间聚类^[6]的初始训练集构建方法: 将待选的未标注样本集 D_u 在核空间聚类, 将 D_u 划分为聚类簇 $D_{u1}, D_{u2}, \dots, D_{uN}$, 选取每簇中的代表性样本(一般为聚类中心样本)构建初始训练集, H_s^* 为在其上训练所得分类超平面. 相对于随机性方法, 该方法保证了初始训练集中样本的多样性, 使得整个算法获得了更好的基点.

另外在主动学习进程中, 当前超平面 H 与最优超平面 H_r 总是存在着一定差异, 在算法早期表现的更为突出. 而式(3)代表的距离准则仅以当前超平面 H 为参考, 可能导致部分本该成为支持向量的样本被忽略, 使得学习所得分类面难以正确地收敛到 H_r , 只能获得一个次优的分类面. 为此, 文中以距离准则为基础, 提出了一种概率选择机制.

若 n 为当前训练集样本个数:

(1) 首先随机地从未标注样本集 D_u 预选 $T(n)$ 个样本, 记为 $D^\# = \{x_1^\#, x_2^\#, \dots, x_{T(n)}^\#\}$;

(2) 然后从这个 $T(n)$ 候选样本中选择待标注样本, 准则为

$$\begin{aligned} x = \arg \min_{x \in D^\#} \quad & \left| \sum_{i=1}^l a_i^* y_i K(x_i, x) + b^* \right|, \\ x_i \in D_l \end{aligned} \quad (4)$$

显然, 若 $T(n)$ 恒为 1, 则该选择策略等同于被动学习算法中的随机选择方法; 若 $T(n)$ 为待选未标注样本集 D_u 中的样本总数时, 则该策略等同于原主动学习策略. 而本文算法中 $T(n)$ 满足: 对任意正整数 n_1, n_2 , 若 $n_1 < n_2$, 则 $T(n_1) \leq T(n_2)$. 一般地, $T(n)$ 可为指数大于 1 的幂函数, 即 $T(n) =$

$n^p, p > 1$. 可见,在算法前期, $T(n)$ 较小,该选择机制具有一定的随机性,容纳进部分因为 H 与 H_r 的误差而导致的离当前分类面较远的有效样本. 而在算法后期, $T(n)$ 较大,该选择机制的随机性减弱,加快了向 H_r 的收敛.

综合以上,改进的 SVM 主动学习算法具体描述如算法 2 所示.

算法 2:改进的 SVM 主动学习算法.

候选未标注样本集 D_u , 训练样本集 D_l , 初始训练集样本数 N_0 , 当前训练样本数 n .

(1) 将 D_u 在核空间聚类,划分为 N_0 个聚类簇,选取每个聚类簇中的代表性样本构建 D_l ;

(2) SVM 在 D_l 上训练得到 H_s^* ;

(3) $n = N_0 + 1$;

(4) 从 D_u 中随机预取 $T(n)$ 个样本, $x_1^\#, x_2^\#, \dots, x_{T(n)}^\#$, 再从中选择到当前超平面最近的样本 x 标注类别后加入 D_l 中;

(5) SVM 在更新后 D_l 的上重新训练;

(6) $n := N_0 + 1$;

如果检测精度达到某一设定值,算法结束;否则,转至步骤(4).

3 仿真实验及结果分析

3.1 实验数据

实验数据采用入侵检测研究领域共同认可并广泛使用的 KDD cup 数据集. 该数据集包括大约 500 万条数据记录,每条都是从军方网络环境中模拟攻击所得的原始网络数据中根据设定的 41 个特征提取出来的,它们都是描述网络连接统计信息的特征向量. 这些数据记录包含五类数据: DoS, Probe, R2L, U2R 四类异常数据以及正常数据(Normal).

实验中,从 KDD 训练样本集中选择 5 000 条数据作为候选样本. 考虑到实际网络中,正常数据流量远远大于异常数据流量,因此数据选取时,正常样本数远多于异常样本数,并且异常样本包含但不区分所有的四种攻击类型. 同时从 KDD 测试集中选择 10 000 个数据作为测试样本集,且测试集的样本分布与候选样本集相同.

3.2 实验结果与分析

为了定量地评价入侵检测系统的检测性能,先定义了如下的评估指标:

检测率 = 正确分类的样本数 / 总的样本数;

误检率 = 正常样本被错分为异常的样本数 /

已知的正常样本数;

漏检率 = 异常样本被错分为正常的样本数 / 已知的异常样本数;

为了验证文中所提算法的有效性,我们特别关注了在达到同等入侵检测率条件下, SVM 被动学习算法,主动学习算法及文中提出的改进算法(p 取 1.5,为经验值)三者所需样本数目的比较结果. 重复实验十次,取平均值对三种算法进行对比,结果如表 1 所示^[7].

表 1 达到同等检测效果三种算法所需样本数

方法	检测率 %	误检率 %	漏检率 %	样本数
被动学习	97.37	0.91	2.12	1 000
主动学习	97.37	0.78	2.37	276
本文方法	97.37	0.82	2.26	211

十次实验后检测率的平均值和标准偏差,与 SVM 被动学习相比, SVM 主动学习算法在小规模标注样本集上即可获得较好的检测效果,从而减少了手工标注样本的工作,并且改进的 SVM 主动学习算法可进一步减少训练样本数:为获得 97.37% 的高检测率,文中提出的改进算法所需训练样本约为原主动学习算法的 3/4. 由实验可知,三种算法所对应检测率的标准偏差都随训练样本数增加呈现总体趋势上地减小,体现出学习算法对大样本的渐进稳定性. 而本文算法的标准偏差整体上低于原主动学习算法,在训练集规模相同的情况下,表现出较高的稳定性.

4 结束语

为解决入侵检测中训练样本获取代价过高的问题,提出了一种改进的 SVM 主动学习算法. 初始化阶段,将待选样本在核空间聚类,通过选择聚类中心样本构建初始训练集,以提高初始样本的质量;在循环学习阶段,以距离准则为基础,引入了概率选择机制,使得学习进程中样本选择策略更为合理. 实验结果表明,改进的算法可以减少训练样本数,有效地降低标注代价,提高了检测系统的实用性.

参考文献:

- [1] 田俊锋,刘仙越. 入侵检测数据分类模型 - PCANN[J]. 微电子学与计算机,2007,24(9):126-129.
- [2] 张家超,孔媛媛,刘飞. 基于最小二乘隐空间支持向量机的 IDS 检测算法的设计[J]. 微电子学与计算机,2008,25(11):171-174.

(下转第 82 页)

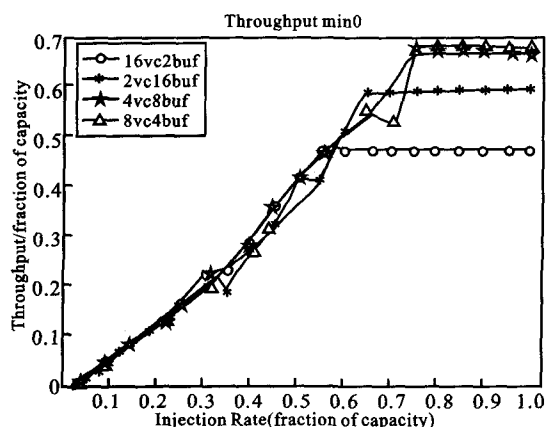


图 6 固定存储资源下不同 VC 数的最小吞吐率

表 1 不同路由方式在不同流量图下的吞吐率

归一化 吞吐率		流量图		
		Uniform	Neighbor	Transpose
路由 算法	DIM	0.8	1.9	0.2
	MAD	0.6	1.2	0.6
	ROMM	0.6	1.4	0.5

5 结束语

文中建立了一种基于 SystemC 的片上网络模型,主要采用的交换策略是基于虫孔交换,其特点是它具有可配置性,能对路由器参数、网络拓扑参数以及数据生成参数进行配置仿真。文中还提出一种较为精确的随机仿真流程,用来消除了启动阶段和收尾阶段的误差。仿真结果能够表明不同参数的片上

网络性能结果,为多核 DSP 等片上网络系统的仿真建立了基础。

参考文献:

- [1] Benini L, De Micheli G. Networks on chips: a new SoC paradigm [J]. Computer, 2002, 35(1): 70-781.
- [2] Coppola M. OCCN: a network-on-chip modeling and simulation framework[C]// Design, Automation and Test in Europe Conference and Exhibition. France: Grenoble, 2004(3): 174-179.
- [3] Jalabert A. Xpipes compiler: a tool for instantiating application specific networks on chip[C]// Design, Automation and Test in Europe Conference and Exhibition. France: Grenoble, 2004(2): 884-889.
- [4] 丁帅,吴宁,葛芬,等. 片上网络路由单元的系统级建模研究[J]. 微电子与计算机, 2009, 26(1): 93-96.
- [5] Caldari M. Transaction-level models for AMBA bus architecture using SystemC 2.0 [C]// Design, Automation and Test in Europe Conference and Exhibition. Italy: Ancona University, 2003: 26-31.
- [6] 张宇昂,李丽,杨盛光,等. 基于 SystemC 的时钟周期精确交易级建模及通信细化[J]. 微电子与计算机, 2007, 24(12): 187-190.

作者简介:

李烨挺 男, (1983-), 硕士研究生. 研究方向为基于片上网络结构的多核 DSP 设计.

梁利平 男, (1969-), 博士, 研究员. 研究方向高性能数字信号处理器和高速混合信号集成电路设计技术.

(上接第 77 页)

- [3] Almgren M, Jonsson E. Using active learning in intrusion detection [C]// Proceedings of the 17th IEEE Symposium on Security Foundations Workshop. IEEE Computer Society Press. Sweden: Goteborg, 2004: 88-98.
- [4] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 北京: 电子工业出版社, 2004.
- [5] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-verlag, 1999.
- [6] Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Machine learn-

ing research, 2001(2): 45-66.

- [7] Girolami M. Mercer kernel based clustering in feature space [J]. IEEE Transactions on Neural Networks, 2002, 13(3): 780-784.

作者简介:

黄双福 男, (1983-), 硕士研究生. 研究方向为网络安全、机器学习.

陈贤富 男, (1963-), 博士, 副教授. 研究方向为智能信息处理、信息安全.