基于RST和SVM的入侵检测系统

刘碧森，钟守铭

(电子科技大学应用数学学院 成都 610054)

【摘要】结合粗糙集理论和支持向量机的特点，构建了基于粗糙集理论和支持向量机的知识简化系统，针对入侵检测系统进行了研究分析，建立了基于粗糙集理论和支持向量机的入侵检测系统，最后通过仿真实验比较，其结果表明该系统有效改进了以往的入侵检测系统的。

关 键 词 支持向量机； 粗糙集理论； 入侵检测； 知识简化

中图分类号 TP 273^{+}.22； O234 文献标识码 A

IDS Based on RST and SVM

LIU Bi-sen，ZHONG Shou-ming

(School of Applied Mathematics, UEST of China Chengdu 610054)

Abstract In this paper, RST is integrated with SVM, and knowledge reduction system is constructed based on RST and SVM, and integrated Rough Set Theory into SVM. Then Intrusion Detection System (IDS) based on RST and SVM is built by analyzing IDS. Finally, comparison of detection ability between the above detection method and others is given. So the IDS based on RST and SVM is a better project.

Key words rough set theory; support vector machine; intrusion detection; knowledge reduction

入侵检测作为网络安全研究的重要内容，近年来，引起了国内外学者的广泛关注。入侵检测通过检查有关的审计数据，以判断系统中是否有违背安全策略或计算机系统安全的行为。

把入侵检测看作为一个分类问题，也就是对给定的审计数据进行分类：什么样的数据是正常的,什么样的数据是异常的。文献[1]把入侵检测看作是区分“自我”(也就是“正常”)和“非自我”(也就是“异常”)的过程,提出了基于免疫模型的入侵检测技术。文献[2]利用神经网络来提取特征和分类。文献[3, 4]从数据挖掘技术的角度探讨了入侵检测的实现问题。以上方法都需要大量或者是完备的审计数据集才能达到比较理想的检测性能，并且训练时间较长。但在大多数情况下，数据都是不完备或不确定的，那么，如何在小样本且有不精确、不确定数据的情况下，提取审计数据特征，实现入侵检测呢？本文运用RST和SVM的各自优点建立了基于粗糙集和支持向量机的入侵检测系统，解决了此问题。

1 基于粗糙集和支持向量机的入侵检测系统的建立

粗糙集理论(Rough Set Theory, RST)是一种处理不精确、不确定与不完全数据的新的数学理论，为研究不完整数据进行分析、推理、发现数据间的关系，提取有用属性，简化信息处理，研究不精确、不确定知

收稿日期：2004 – 06 – 09

作者简介：刘碧森(1976 –)，男，硕士，主要从事稳定性理论、学习机、神经网络等方面的研究.

识的表达、学习、归纳方法等提供了一个有力的工具[5]。支持向量机(Support Vector Machine, SVM)是一种建立在统计学习理论基础之上的机器学习方法，是一种通用性极强，推广能力很好的学习机器，是解决分类问题强有力的工具和方法[6]。其最大的特点是根据Vapnik结构风险最小化原则，尽量提高学习机的泛化能力，即由有限的训练集样本得到小的误差仍然能够保证对独立的测试集保持小的误差。入侵检测系统结构如图1所示。

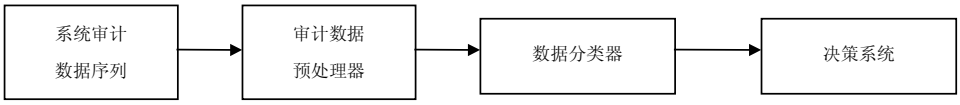

图1 入侵检测系统

下面利用RST和SVM各自的优点把它们结合在一起，构成基于RST和SVM的知识简化系统如图2所示。方法是通过交换修改决策模型，产生最小决策规则网络。该方案可按图2的思路进行。

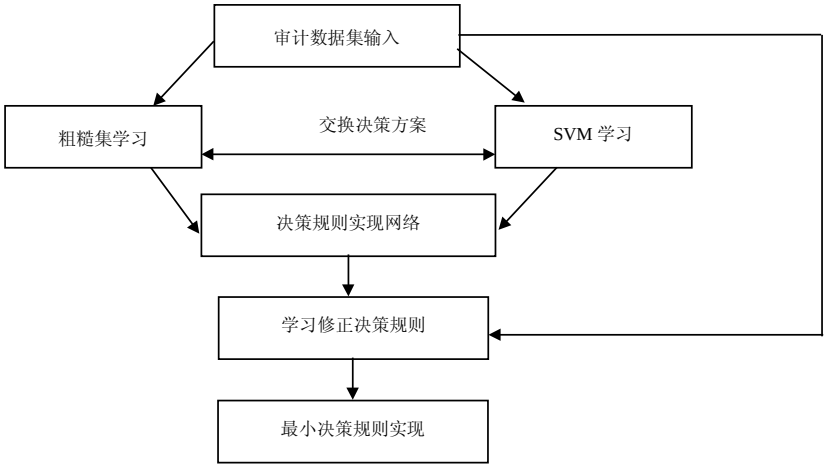

图2 基于RST和SVM的知识简化系统

首先把审计数据集输入RST系统(前置系统)，粗糙集学习按“先简化规则、后简化属性”的方法进行；同时，SVM系统(后置系统)对输入的每一个同样的审计数据集对象进行学习，学习采用Sigmoid型核函数的SVM，用来调整决策规则的依赖因素。决策方案的修正是通过SVM学习和粗糙集学习之间的交换改进的，直到粗糙集学习选出最少属性构成的决策规则能全部正确划分所有的审计数据集的样本为止。建立基于RST和SVM的入侵检测系统，其系统框图如图3所示。

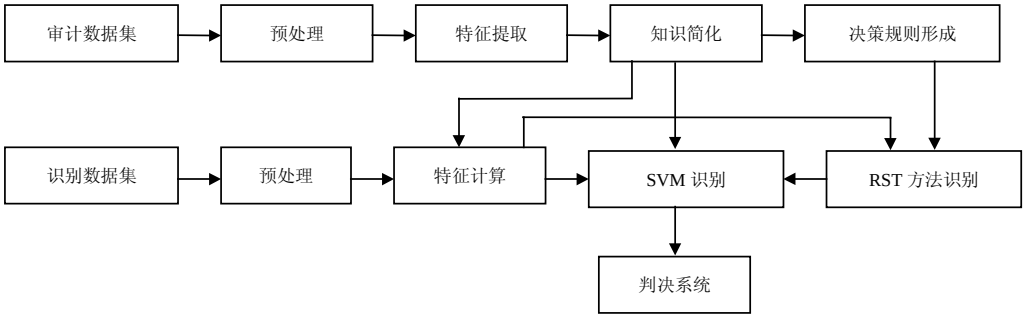

图3 基于RST和SVM的入侵检测系统

审计数据预处理器用来对大量的系统审计纪录进行处理或变换。由于支持向量机的分类器只能对维数相同的数字向量进行分类，但系统审计数据中的数据不但长度不尽相同，而且很有可能是不精确、不确定数据，所以必须利用RST系统将原始数据转换成支持向量机能够识别的数字向量。支持向量机分类器对这些

数字向量进行分类，产生判决结果。当然，这些判决结果可以直接作为整个入侵检测系统的输出，但为了进一步提高整个系统的正确率，可以设定一些判决准则，例如发生数目、百分比等来进行最终的判定。这个过程是由决策系统完成的。

2 实验仿真

在下面的计算机仿真中，人工智能实验室(AI Lab.)公开提供的lpr数据进行仿真，以此为例来讨论基于RST和SVM的入侵检测系统的工作过程。

2.1 获得训练样本

使用长度为 k 的滑动窗口对已知正常的系统调用执行迹进行扫描,可以得到正常的系统调用短序列样本。由于入侵的非法活动只占程序执行的一小部分,所以异常短序列只占异常执行迹的很小一部分。当用长度为 k 的滑动窗口对于异常的执行迹进行扫描时,会得到一组既有正常短序列又有异常短序列的系统调用短序列列表。将这组短序列列表与已获得的正常短序列样本进行比较,不同于正常短序列的那些系统调用短序列就构成了异常短序样本。设得到了 n 个样本，可将所有的短序列样本记为 $(x_1, y_1),\cdots,(x_n, y_n)\in R_k\times\{\pm1\}$ 的形式，其中正常样本的标签为+1；异常样本为–1，且 $k=6$ 。所得到的正常和异常的系统调用短序列训练数据样本如表1所示。

表1　正常和异常的短序列训练样本

$k=6$ 的短序列训练样本						分类标签
5	3	67	67	5	139	正常(+1)
3	67	67	5	139	67	正常(+1)
			⋮			⋮
19	4	6	9	10	6	正常(–1)
3	6	9	10	6	6	正常(–1)
			⋮			⋮

2.2 入侵检测仿真

基于RST和SVM的检测系统进行入侵检测的过程分为两个阶段:训练阶段和检测阶段。1) 在训练阶段，先利用RST系统将不完备数据、不精确数据进行处理，然后把得到的训练样本根据支持向量机原理得到支持向量和相关参数。2) 在检测阶段，可以对未知状态的系统进程执行迹进行判定了。首先，审计数据预处理器先用长度为 k =6的滑动窗口对该执行迹进行扫描，得到了一组系统调用短序列。将这些系统调用的短序列输入RST和SVM分类器，可以得到判决向量。然后根据判决规则对该判决向量进行判决，即可判定该执行迹的状态。表2给出了本文方法训练数据和检测数据的分配情况。

表2　仿真中训练数据和检测数据的分配情况

训练数据集		检测数据集	
正常执行迹	异常执行迹	正常执行迹	异常执行迹
20	20	2 700	1 000

表3给出了本文的仿真结果与文献[7]得出的研究结果的比较。文献[7]主要研究了4种算法。Stide方法是根据训练数据集预先列出长度为k的所有独特、连续的系统调用序列作为正常行为的特征。然后统计所要检测的执行迹的这些特征，并将它们与已建立的正常行为的特征比较，如果不同特征的数目超过阈值就判定有异常发生。t-Stide与Stide类似，只是在检测时利用的是不同特征出现的频率作为判决规则。RIPPER是William Cohen提出的一种规则学习算法。这种算法主要应用于分类问题。HMM方法是用隐马尔可夫模型(HMM)对正常的执行迹进行建模。在检测阶段，对于状态转移概率和输出概率给出阈值，如果低于阈值，则判定有异常发生。其提供的数据是在检测率为100%的情况下最低虚警率的数据。由表3可以看出，HMM的检测性能最好，但花费的时间较长。对于相同的检测结果，本文的RST和SVM所需的训练时间要远远小

于HMM方法。

表3 几种入侵检测方法的性能比较

	Stide	t-Stide	RIPPER	HMM	RST&SVM
训练所花时间	10min	10 min	1 min	5 d	6 min
最低虚警率	0.0	0.000 75	0.001 6	0.000 3	0.000 2

3 结 论

通过实验可以看出，基于RST和SVM的入侵检测系统具有以下优点：1) 它不需要所有的正常和异常的信息，在给出较少的正常和异常执行迹的情况下就能得到比较理想的检测效果；2) 该方法所需要的训练时间和检测时间比其他方法短；3) 该方法能够随时升级，并进行高效的实时检测。所以，此系统是原来入侵检测系统的有效改进。

参 考 文 献

[1]Forrest S, Perrelason A S, Allen L, et al. Self_Nonself discrimination in a computer[A]. In: Rushby J, Meadows C, eds. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy[A]. Oakland, CA: IEEE Computer Society Press, 1994. 202-212

[2] Ghosh AK, Michael C, Schatz M. A real-time intrusion detection system based on learning program behavior[A]. In: Debar H, Wu SF, eds. Recent Advances in Intrusion Detection (RAID 2000)[C]. Toulouse: Spinger-Verlag, 2000. 93-109

[3] Lee W, Dong X. Information-theoretic measures for anomaly detection[A]. In: Needham R, Abadi M, eds. Proceedings of the 2001 IEEE Symposium on Security and Privacy[C]. Oakland, CA: IEEE Computer Society Press, 2001. 130-143

[4] Lee W, Stolfo S J. A data mining framework for building intrusion detection model[A]. In: Gong L, Reiter MK, eds. Proceedings of the 1999 IEEE Symposium on Security and Privacy[C]. Oakland, CA: IEEE Computer Society Press, 1999. 120-132

[5] 张文修, 吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊数学与数学, 2000, 14(4): 1-12

[6] Vapnik V N. The nature of statistical learning theory[M]. New York: Spring-Verlag, 1995

[7] Warrender C, Forrest S, Pearlmutter B. Detecting intrusions using system calls: Alternative data models[A]. In: Gong L, Reiter MK, eds. Proceedings of the 1999 IEEE Symposium on Security and Privacy[C]. Oakland, CA: IEEE Computer Society Press, 1999. 133-145

编 辑 刘文珍