2014 年全国大学生信息安全竞赛

作品简介

作品名称：基于支持向量机的网络异常检测系统

电子邮箱：mingyuqi.java@qq.com

提交日期：2014 年 7 月 17 日

摘要

随着信息化技术的深入和互联网的迅速发展，整个世界正在迅速地融为一体，计算机在经济和生活的各个领域正在迅速普及，计算机网络已经成为国家的经济基础和命脉，其地位越来越重要，整个社会对网络的依赖程度越来越大。而网络发展的同时，也产生了各种各样的问题，其中安全问题尤为突出。现在，网络中蠕虫、病毒及垃圾邮件肆意泛滥，木马无孔不入，DOS 攻击越来越常见，网络资源滥用，黑客攻击行为几乎每时每刻都在发生，所有这些极大地困扰着包括企业、组织、政府部门与机构等在内的各种网络用户。

能否及时发现网络黑客的入侵，有效地检测出网络中的异常流量，成为所有网络用户面临的一个重要问题。而网络异常检测是保护网络安全的重要途径，本作品——网络异常检测系统便是以此为背景，将所需要处理的网络信息流按一定的规则分类，分成正常类别与异常类别，并输出相应的判断结果，进而作出报警响应，从而实现对病毒、入侵的识别与防御，达到维护网络安全的目的。

检测器是检测系统的核心部件，目前检测器的核心算法有：贝叶斯分类算法、BP 神经网络算法、遗传变异算法等。其中本检测系统中的检测器采用支持向量机(Support Vector Machine，简称 SVM）分类算法，具有更好的数学理论依据对分类结果做支撑。

本系统的开发主要分为两大部分，一是检测器的研发，二是可视化系统的开发。本系统的核心模块——检测器具有很强的理论基础做支撑，本作品的检测器采取 KDD99 数据集进行实验，运用主成分分析(Principal Component Analysis，简称 PCA)方法对数据进行降维，运用支持向量机分类算法进行分类，而支持向量机的核心参数又运用试探训练法来确定，缩短了检测时间，提高了检测率，大大提高了本系统核心部件检测器的性能。此外，本系统可视化部分主要包括查看检测历史、统计一周内的检测情况、查看并修改异常类信息等功能，并且界面友好，操作简易，具有良好的用户体验。

一. 相关工作

KDD99是1998年美国在林肯实验室模拟真实的网络环境，仿真各种用户类型、各种不同的网络流量和攻击手段，收集到的网络连接和系统审计数据，经过特征分析和预处理之后，形成的一个标准数据集。该数据集现已成为网络入侵领域的网络流量标准数据，成为目前网络入侵检测器实验用标准数据。本检测器正是以该数据集作为标准输入，在Matlab中进行仿真实验，Matlab从数据库读取原始数据并对数据进行预处理，利用主成分分析方法对数据进行有效的降维缩短检测时间，并在前人基础上对检测器分类算法进行研究改进以及检测器的具体实现。

二. 本作品的研究内容

2.1 作品整体框架

本作品的整体框架如下图 2-1 所示。

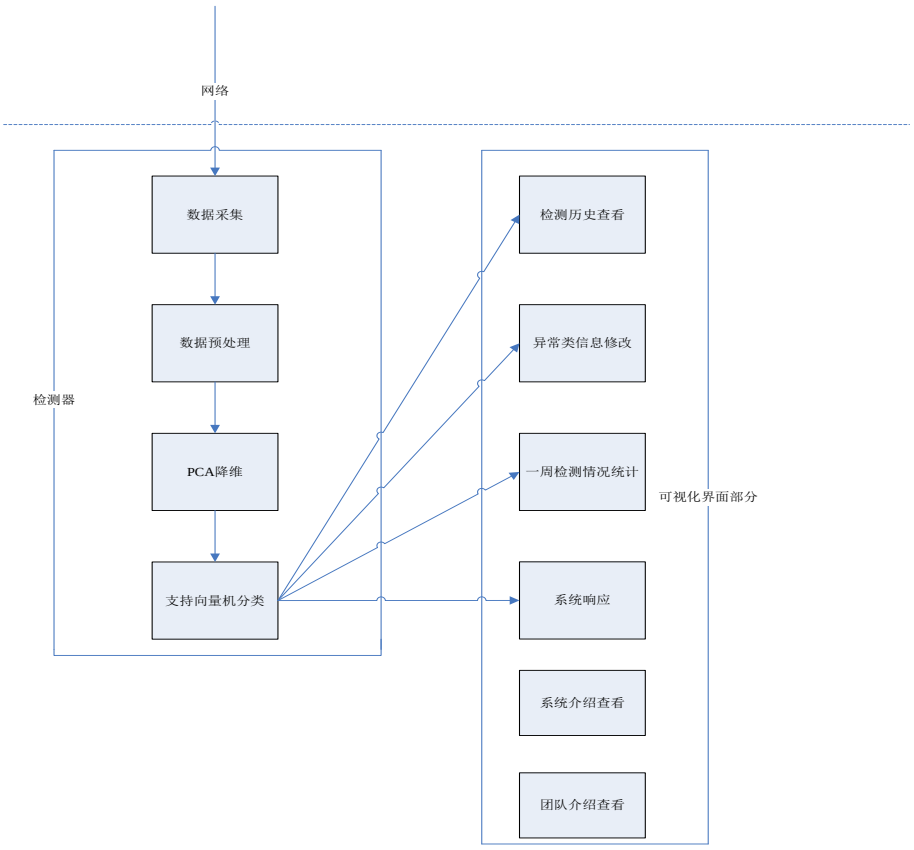

图2-1 整体框架

2.1.1 检测器部分

由图2-1可知，本系统检测器的形成由数据采集模块、数据预处理模块、PCA模块和支持向量机分类模块组成，以下是对这四个模块的简单介绍：

（1）数据采集模块

数据采集模块是用来捕获网络中的数据流，而本作品系统中采用从数据库自动读取，是其他所有模块实现的基础。

（2）数据预处理模块

数据预处理模块是通过相应程序对从数据库中提取的数据进行数据的标准化，主要其包括三个部分的处理：文本型属性的数值化、连续型属性的离散化和属性数据的归一化。

（3）PCA模块

PCA模块主要是对已预处理过的数据进行主成分分析，得到数据集的主要成分，为使用支持向量机分类作准备。

（4）支持向量机分类模块

支持向量机分类模块是本系统的核心模块，它的形成主要分为两大部分，一是训练，二是预测。本系统在对支持向量机的训练过程中，采用了Boosting分步算法确定了支持向量机的参数，得到了较高的效率，然后在此基础上用来对入侵检测数据进行分类。

这四个模块构成了检测器，形成了本系统的核心部件。

2.1.2 可视化界面部分

可视化界面部分的实现全都得益于检测器的实现，可视化界面部分主要是对检测器的分类结果进行处理，主要分为系统响应、检测历史查看、异常类信息修改以及一周检测情况统计四大模块。以下是这四大模块的简单介绍：

（1）系统响应

系统响应是根据支持向量机的分类结果实行的，若是异常类便发出警报声音和弹出相应的提示窗口，以提醒用户对网络进行防护。

（2）检测历史查看

检测历史查看是对检测系统的检测历史进行查看，主要分为今日检测历史的查看、

正常类型检测历史的查看和异常类型检测历史的查看。

（3）异常类信息修改

异常类信息的修改主要从用户的角度出发，用户可对异常类信息作出自己的专业判断，然后可对异常类信息增加信任变为正常类型，系统也会相应的对此操作进行记录

（4）一周检测情况统计

一周检测情况统计故名思意是对系统使用当日起往后的一周的检测情况进行统计，用折线图反映出最近一周检测到的正常类型和异常类型的个数。

2.2 作品功能结构

本作品的功能结构如图 2-2 所示。

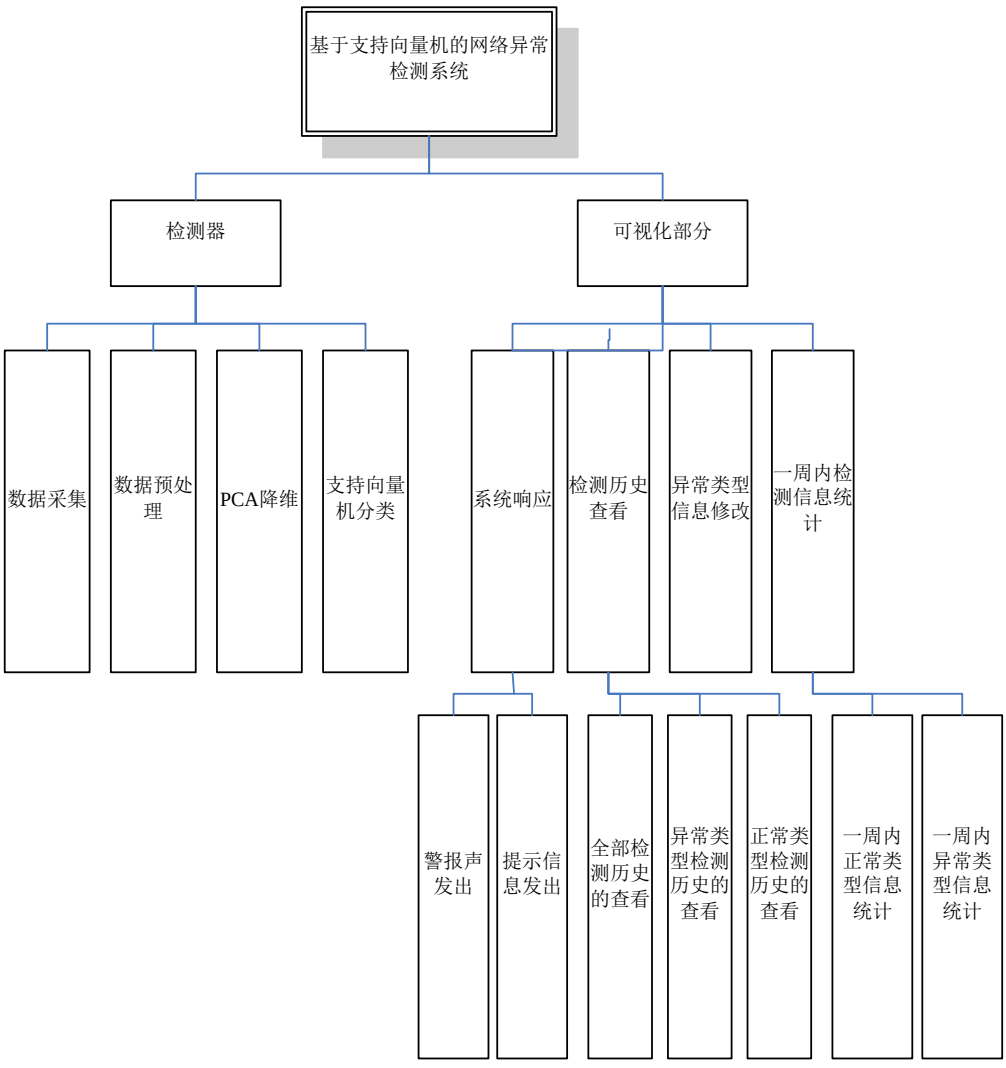

图2-2 功能结构图

2.3 检测器的实现

本系统的核心部件——检测器的分类算法采用的是支持向量机分类算法，支持向量机训练中采取支持向量库中的SVM训练数据，使用SVM主动学习算法训练SVM分类器；在本作品中是通过额外的大量数据（30万条）分步进行训练得到的成熟的检测器。

经过训练的SVM分类器用于对经过数据预处理的网络数据包进行检测，并将检测结果传送给响应模块，同时存入支持向量库，若发现入侵，相应模块则采用相应的响应策略；在本作品中是通过界面显示返回的判别结果实现该响应功能。

若检测有误，则进行误差分析，重新根据新数据的错误分类特征进行重新训练SVM分类器过程；在本作品中是通过人工修改核心算法完成。

由于本作品的研究目的是构建一个分类更快，准确率更高的网络异常分析判断检测器，所以研究的重点更多的是放在对网络访问特征数据的预处理和将这些数据通过分类、训练，形成一个最佳分类效果的分类上。

在以下几点上有研究与改进：

（1）对大量网络数据属性特征的提取与筛选，提出基于PCA的SVM网络入侵方法，利用主成分分析做到对数据的有效降维。

（2）针对支持向量机分类原理，采用Boosting分步训练法，得到了较好的结果。

（3）对于训练过程中相关参数的确定，提出了试探法算法流程，能较快得到较优参数。

三. 测试结果与分析

3.1 检测器测试

实验过程中定义检测率和误报率两个性能指标，这两个指标能有效反应检测器的效果与性能。

$$\text{正常样本检测率（True Negative rate，TN）}=\frac{\text{被正确判断的正常样本数}}{\text{正常样本的总数}}$$

$$\text{误报率（Fales Positive rate，FP）}=\frac{\text{被误判的正确样本数}}{\text{正确样本的总数}}$$

将 corrected.gz 里面的 30 万条数据随机分为 10 份，按 corrected1~corrected10 编号，比较经过该检测器预测的数据类型与实际数据类型，得出个测试数据集相应性能指标如表 3-1 所示。

表 3-1 各个测试集相应指标

测试子集	TN（检测率）	FP（误报率）
Corrected1	79.7166% (10521/13198)	20.28%
Corrected2	36.5037% (3650/9999)	63.50%
Corrected3	100% (9999/9999)	0.00%
Corrected4	99.6599% (9964/9998)	0.34%
Corrected5	99.2293% (10815/10899)	0.77%
Corrected6	97.8491% (9781/9996)	2.15%
Corrected7	90.6917% (10883/12000)	9.31%
Corrected8	96.8725% (9881/10200)	3.13%
Corrected9	94.4562% (9371/9921)	5.54%
Corrected10	79.668% (7966/9999)	20.33%

从表中各性能指标可以看出实验得到了预期的效果，各个测试集得到了较高的检测率和较低的误报率，说明该检测器是有效的。

从表3-1中各性能指标可以看出实验得到了预期的效果，各个测试集得到了较高的检测率和较低的误报率。与传统检测器相比，得到如下图所示的折线图：

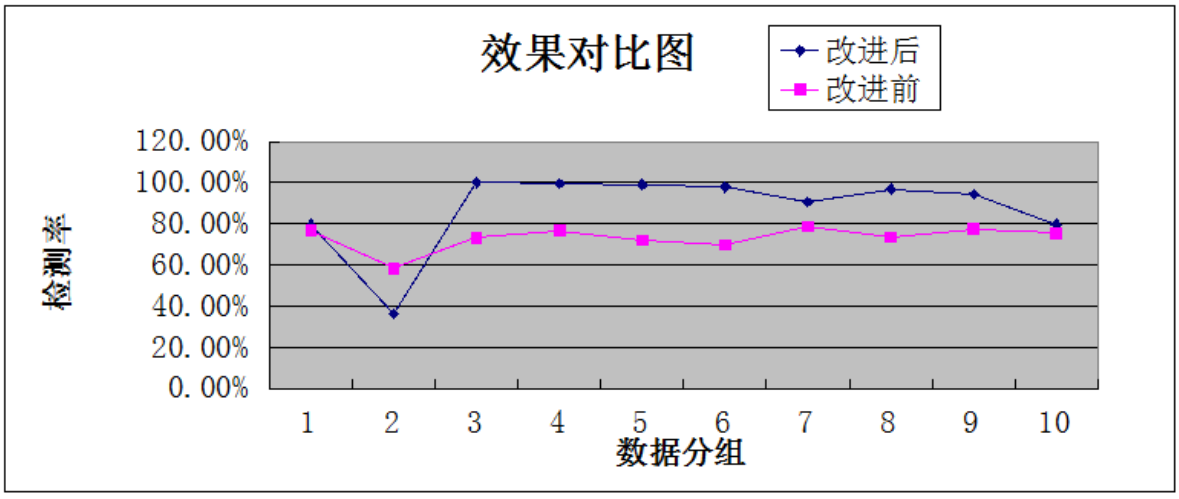

图 3-1 检测器效果对比图

由以上的图表可知，测试得到了较好的结果，说明该检测器是有效的。

3.2 其它功能测试

下表 3-2 是针对各个模块进行的测试记录。

表 3-2 测试用例记录表

测试模块	用例个数	测试结果
数据采集模块	8	全部通过
数据预处理模块	10	全部通过
PCA 模块	8	全部通过
支持向量机测试模块	15	全部通过
系统响应模块	15	全部通过
一周检测情况统计	7	全部通过
异常类信息模块	9	6 个通过，3 个没通过
检测历史模块	10	全部通过
系统介绍模块	3	全部通过
团队介绍模块	4	全部通过

对以上测试结果进行分析：

（1）按缺陷特性分析

表 3-3 按缺陷特性分析表

模块/特性	存在问题
异常类信息模块	文本框输入的信息没有限制为数字，当输入非数字类型的信息时，系统并没有提示错误信息，而是可以继续使用，只是没有实现更改异常类型信息的目的。

（2）按缺陷类型统计分析

表 3-4 按缺陷类型统计表

缺陷类型	缺陷个数	所占百分比
代码错误	10	22%
用户界面	4	8.9%
标准规范	13	28.9%
新增需求		
需求变动		
设计文档	8	18.2%
配置相关		
性能压力	10	22%
其他		

缺陷分析的目的是为了得出：缺陷原因、遗留缺陷以及规避措施等。 结合以上统计数据可以得出以下结论：

系统的错误处理能力不够，以后要提高设计能力，在架构设计上要增加一些异常处理机制。

提示/建议缺陷主要为提示信息文字描述，用户界面等缺陷。以后要在界面上和提示语方面制定规范，提高设计界面水平。

四. 应用前景分析

支持向量基是基于统计学习理论的，将其应用到入侵检测系统中，可保证在先验知识不足，小样本的情况下支持向量基分类器仍有较好的分类准确率，在此基础上对KDD99数据集进行主成分分析，有效的对数据降维，明显降低了检测时间。本系统的检测器达到了能够对系统异常情况准确预测的目的，该方法避免了基于传统机器学习的局限性,保证了较强的推广能力,从而使整个网络异常检测系统具有较好的检测性能。

五. 创新点总结

1. 目前大部分检测器核心算法多采用 BP 神经网络、贝叶斯分类、遗传变异等算法，而本文采用支持向量机分类算法，不仅有数学理论对分类依据做强有力的支撑，使其具有强大的说服力，而且在各项性能指标上均优于传统检测器。

2. 在属性约简上，提出了主成分分析与支持向量机的结合，以确定属性的选取，简化数据量的同时，保证了正确率，缩短了检测时间。

3. 在训练参数的确定方面，本文提出了相应的算法流程，依此流程大大减少了检测器训练和最优参数确定时间，该参数试探算法具有良好的普适性。

4. 我们在支持向量机基础上，又采用了 Boosting 分步训练的方法，以形成成熟的检测器，使其在检测率、误报率、漏报率三个指标上均高于一般的检测器。

5. 在大量数据处理方面，包括数据特征值的提取、数据筛选与标准化，对入侵检测和网络安全的研究有很好的借鉴意义。

6.开发出了具有可视化界面的实际系统，能够运行，实时判断。

六. 未来工作

1. 本系统检测器只能网络正常数据与异常数据，而不能针对具体是哪类异常或攻击做出判断。今后可以在SVM多分类算法上进行研究改进，如一对一，一对多，二叉树算法，使分类器实现可具体判断攻击的类型。

2.在检测器的训练过程中，参数的确定上依然有很大的提高空间。今后可以考虑采用与神经网络、遗传变异算法相结合确定最优参数。

3.可以做底层的网络抓包模块，使之成为一个真正的网络异常检测系统。