

基于聚类和支持向量机的入侵检测研究

倪霖¹, 郑洪英²

(1. 重庆大学 机械工程学院, 重庆 400030; 2. 重庆大学 计算机科学与工程学院, 重庆 400030)
(nilin71@163.com)

摘要:提出了一种基于聚类和支持向量机的入侵检测算法,该算法可以有效地减小计算复杂性,提高检测性能。算法首先使用 K-MEANS 聚类算法对数据做一个初步的划分得到多个类;然后考察每个类中数据的标识,只有当类中的数据标识不止一个时才进行支持向量的查找。最后使用 KDD CUP 1999 进行了计算机仿真实验,实验结果说明了该算法的有效性。

关键词:聚类;支持向量机;入侵检测;KDD CUP 1999

中图分类号: TP309.2 **文献标志码:** A

Research of intrusion detection based on clustering and support vector machine

NI Lin¹, ZHENG Hong-ying²

(1. Department of Mechanic Engineering, Chongqing University, Chongqing 40030, China;
2. Department of Computer Science and Engineering, Chongqing University, Chongqing 40030, China)

Abstract: A new approach based on clustering and Support Vector Machine (SVM) was proposed. It could reduce the computational complexity and get better performance in the intrusion detection. Firstly, clustering was imposed on data set using K-MEANS partition algorithm and clusters were established. Secondly, the support vector was searched only when the labels in a cluster were different. Finally, the experimental results using KDD CUP 1999 demonstrate the efficiency of our approach.

Key words: Clustering; Support Vector Machine (SVM); intrusion detection; KDD CUP 1999

0 引言

随着计算机技术和通信技术的发展,由入侵而造成的损失以及和计算机相关的犯罪也急剧增加。因此,网络安全即确保系统按照预期目标正常、稳定的运行,成为人们关注的焦点。入侵检测系统(Intrusion Detection System, IDS)是从计算机或网络中抽取信息,用以检测来自于系统外部的入侵者和内部人员对系统的误用。它能在不影响网络性能的情况下对网络进行监测,为系统提供对内部攻击、外部攻击和误操作的有效保护,并弥补防火墙的不足,起到主动防御的作用。入侵检测系统需要把入侵数据从正常数据中划分出来,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]从数据挖掘技术的角度探讨了入侵检测的实现问题。文献[5]使用支持向量机进行入侵检测来克服入侵检测系统存在着在先验知识较少的情况下推广能力差的问题。

实际应用中数据的分类是非常复杂的,常常体现为高维、小样本和不可分性。所谓高维,指的是入侵检测中的数据集中在正常情况下有很多属性,例如在 KDD CUP1999^[6]中数据的

维数是 41。所谓小样本,指的是数据集所能体现的信息是有限的。小样本的统计学习问题是在样本数目有限的条件下尽可能多地利用样本集合所提供的信息的机器学习问题。在入侵检测中,对于一个特定的攻击类型来说,样本又是很缺乏的,有些攻击类型在 5 000 000 左右的样本中只有几个样本。在这样的样本结构下,要想准确地区分每一种攻击,对于任何一种方法都几乎是不可能的,更不要说去发现产生这些样本的概率等统计规律了。所谓不可分性,指的是要进行分类的样本可能是高度重叠,不可以用线性分类器进行简单的分类。而支持向量机(Support Vector Machine, SVM)是在小样本学习的基础上发展起来的分类器设计方法,专门用于小样本数据,对数据的维数不敏感。因此 SVM 方法适合于入侵检测领域高维异构数据集中的分类器设计,以及对于正常状态数据和异常状态数据的有指导学习和划分。SVM 方法作为模式识别中的一种新的学习和分类方法,在很多的分类问题上已经显示出很好的性能。

1 支持向量机的分类算法

SVM^[4]是从线性可分情况下的最优分类面发展而来的,使分类间隔最大实际上就是对推广能力的控制,这是 SVM 的核心思想之一。设样本为 n 维向量,某区域的 k 个样本及其所属类别表示为:

收稿日期:2007-04-04;修回日期:2007-06-18。

基金项目:国家 863 计划项目(2006BAH02A09);重庆市科技计划重点项目(2006AB2025)。

作者简介:倪霖(1971-),男,重庆人,副教授,主要研究方向:信息系统、信息安全、项目管理; 郑洪英(1975-),女,重庆人,博士研究生,主要研究方向:信息安全。

$$(X_1, Y_2), (X_2, Y_2), \dots, (X_k, Y_k) \in R^n \times \{\pm 1\} \quad (1)$$

超平面为:

$$\omega \cdot x + b = 0 \quad (2)$$

将样本划分成两类,其中“ \cdot ”表示向量的点积。最佳的超平面应使两类样本到超平面最小的距离为最大。显然,式(2)中 ω 和 b 乘以相同系数后仍满足方程。不失一般性,设对所有样本 X_i ,式 $|\omega \cdot x + b|$ 的最小值为1,则样本与此最佳超平面的最小距离应为 $|\omega \cdot x_i + b| / \|\omega\| = 1 / \|\omega\|$ 。最佳超平面应满足约束:

$$Y_i[\omega \cdot X_i + b] \geq 1; i = 1, 2, \dots, k \quad (3)$$

ω 和 b 的优化条件应是使两类样本到超平面最小的距离之和 $2 / \|\omega\|$ 为最大。另外,考虑到可能存在一些样本不能被超平面正确分类,因此引入松弛变量 $\xi_i; i = 1, 2, \dots, k$ 。超平面的约束变成:

$$Y_i[\omega \cdot X_i + b] \geq 1 - \xi_i; i = 1, 2, \dots, k \quad (4)$$

对 $\|\omega\|$ 稍加变形后,问题变成最小化:

$$\Phi(w, b, \xi) = \frac{1}{2}w \cdot w + C \sum_{i=1}^k \xi_i \quad (5)$$

其中 C 为一正常数。式(5)中的第一项使样本到超平面的距离尽量大,从而提高泛化能力;第二项则使误差尽量小。

为求解这个优化问题,引入拉格朗日函数,其中 $a_i \geq 0, y_i \geq 0; i = 1, \dots, k$ 。

$$L(w, b, \xi, a, \gamma) = \frac{1}{2}w \cdot w + C \sum_{i=1}^k \xi_i - \sum_{i=1}^k a_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^k \gamma_i \xi_i \quad (6)$$

函数 L 应对 a_i 和 y_i 最大化,且对 ω, b 和 ξ_i 最小化。函数 L 的极值满足条件:

$$\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \xi_i} = 0 \quad (7)$$

从而得到:

$$\sum_{i=1}^k a_i y_i = 0 \quad (8)$$

$$w = \sum_{i=1}^k a_i y_i x_i \quad (9)$$

$$C - a_i - y_i = 0; i = 1, 2, \dots, k \quad (10)$$

将式(8)~(10)三个式子代入式(6),可以得到优化问题的对偶形式,最大化函数:

$$W(a) = -\frac{1}{2} \sum_{i,j=1}^k a_i a_j y_i y_j K(x_i, x_j) + \sum_{i=1}^k a_i \quad (11)$$

其约束为:

$$\sum_{i=1}^k a_i y_i = 0 \quad (12)$$

$$0 \leq a_i \leq C; i = 1, 2, \dots, k \quad (13)$$

判别函数为:

$$f(x) = \text{sign}(\omega \cdot x + b) \quad (14)$$

对于非线性分类,首先使用一个非线性映射 Φ 把数据从原空间 R^n 映射到一个高维特征空间 Ω ,再在高维特征空间 Ω 建立优化超平面。高维特征空间 Ω 的维数可能是非常高的,支持向量机理论巧妙地解决了这个问题。在非线性空间也只考虑在高维特征空间 Ω 的点积运算 $\Omega(x) \cdot \Omega(y) = K(x, y)$,不必明确知道 $\Omega(x)$ 是什么。这样,在非线性情况下支持向量机对分类问题成为最大化函数:

$$W(a) = -\frac{1}{2} \sum_{i,j=1}^k a_i a_j y_i y_j K(x_i, x_j) + \sum_{i=1}^k a_i \quad (15)$$

其约束为(12)~(13)两个式子。此时决策面为:

$$\sum_{\text{支持向量}} a_i y_i K(x, x_i) + b = 0 \quad (16)$$

判别函数为:

$$f(x) = \text{sign} \left[\sum_{\text{支持向量}} a_i y_i K(x, x_i) + b \right] \quad (17)$$

这里 $K(x, y)$ 称为核函数,核函数的选取应使其为特征空间的一个点积,即存在函数 Φ ,使 $\Phi(x) \cdot \Phi(y) = K(x, y)$ 。常用的核函数有三种:

1) 多项式核函数

$$K(x, y) = (x \cdot y + 1)^d; d = 1, 2, \dots \quad (18)$$

2) RBF(Radial Basis Function)核函数

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (19)$$

3) Sigmoid核函数

$$K(x, y) = \tanh[b(x \cdot y) - c] \quad (20)$$

一般而言,核函数的选择实际上是选择一组具有几个参数的核,比如RBF核中参数 σ^2 需要确定。通常这个选择可以使用启发式选择方式,而且这个选择必须适用于具体数据。当缺乏可靠的条件时,在应用中需要使用验证集或者交叉验证来设置这些参数。

2 基于支持向量机的入侵检测模型

基于SVM的入侵检测模型使用训练数据集对支持向量分类器进行训练,训练出的SVM分类器用于实时的入侵行为检测。SVM的训练算法采用一种有指导的学习算法,即SVM训练算法是采用已经标出分类结果的数据集进行训练。另一方面,在对支持向量机进行训练之前,我们先对训练数据使用K-MEANS聚类算法方法做一个预处理。预处理一方面可以减少构造支持向量的时间;另外,聚类的目的是把数据划分成多个族,分析各个族中数据的特征。因为支持向量只存在于异类数据的边界,而属于同一类的数据不需要建立支持向量。这一思想可以用图1来描述。在图1中,类 C_1, C_3 中样本数据的类标识都是统一的,而在 C_2 中,具有不同类标识的样本数据由于距离较近分到了同一类中。因此,支持向量只可能存在于类 C_2 中。

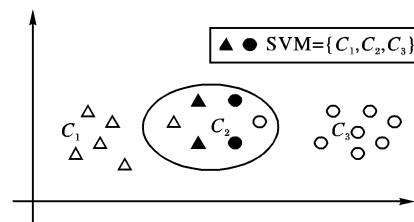


图1 聚类族中的支持向量

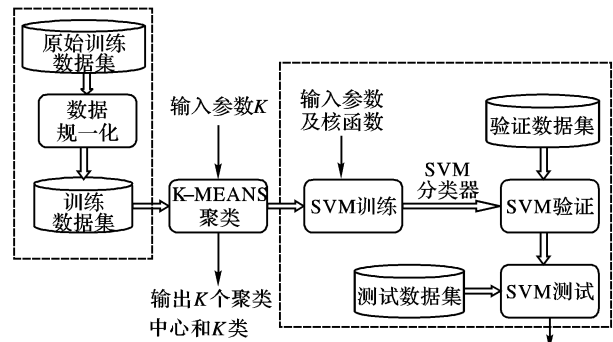


图2 基于聚类和支持向量机的入侵检测模型

因此,基于聚类和支持向量机的入侵检测模型主要由数据规一化处理、K-MEANS聚类和支持向量机分类器三部分构

成,如图 2 所示。

3 实验结果

3.1 数据归一化

在 KDD CUP1999 所有 41 维特征中,每个特征的取值范围有很大差异,有的特征最大值达到 10^9 ,而有的特征最大值只有 1。而 SVM 实际上是一种基于“距离”的分类算法,根据样本距离分类超平面的远近来确定样本的属性(正常、异常)。这必然会导致在得到的 SVM 中,个别特征对整体性能的影响较大。因此就很有必要使用某些归一化方法来减小支持向量性能对个别特征的过度依赖。规范化采用下述公式, $X = \{x_{ij} | i = 1, \dots, N, j = 1, \dots, D\}$ 是输入数据, N 是样本数据的数目, D 是样本数据的特征维数, μ 是均值, σ 是样本的标准差。另外,我们只选用了 11 维特征,它们分别是: duration、src_bytes、dst_bytes、logged_in、count、srv_count、same_srv_rate、dst_host_count、dst_host_srv_count、dst_host_same_src_port_rate、dst_host_srv_diff_host_rate。

$$\chi'_{ij} = \frac{\chi_{ij} - \mu}{\sigma} \quad (21)$$

表 2 训练集 1 使用 K-MEANS 聚类算法产生的聚类中心

类	1	2	3	4	5	6	7	8	9	10	11
1	0.385	-0.057	0.8535	-1.349	-1.3245	0.0224	-1.7662	-2.4617	-0.1817	-1.1724	0.0183
2	-0.04	10.3	1.53	-1.35	-1.33	0.17	-0.6	-3.11	-3.56	-1.37	-0.31
3	-0.213	0.088	1.53	-1.3273	-1.2904	0.17	-2.1922	0.37	0.31	-1.2224	2.0037
4	-0.195	0.066	1.2878	-1.1673	-1.1404	0.17	0.4731	0.3263	0.2973	-1.2235	-0.2987
5	30.955	-0.07	0.44	-1.33	-1.31	0.17	-1.72	-3.125	-3.05	-0.94	-0.31
6	-0.342	-0.010	-0.5392	-1.1834	-1.1615	0.17	0.03341	-2.978	-3.4941	-1.1859	-0.17305
7	-0.228	1.83	1.53	-1.33	-1.3067	0.17	-1.9767	0.37	0.31	-1.3017	1.67
8	0.24	47.59	1.53	-1.35	-1.33	0.17	-0.36	-3.04	-3.48	-1.37	-0.31
9	-0.249	0.061	1.53	-1.3236	-1.2896	0.17	-0.78	0.37	0.31	-1.37	1.2636
10	-0.456	-0.064	-0.65	-1.35	-1.33	0.17	-1.6257	-3.1657	-2.0386	-0.46	13.206
11	3.701	-0.066	-0.4078	-1.3422	-1.3189	0.17	-1.6222	-2.5867	-2.9333	-1.0389	2.81
12	0.888	-0.056	1.53	-1.35	-1.3249	0.17	-1.0538	-0.7227	-0.6512	-1.3566	0.095
13	0.07	0.078	1.53	-1.35	-1.3267	0.17	-2.515	0.1867	0.31	0.225	1.43
14	-0.098	0.113	1.53	-1.3319	-1.2933	0.17	2.3007	0.19	0.2741	-1.1441	0.612
15	-0.461	-0.07	-0.65	-1.0467	-1.3048	-4.7633	-1.3629	-2.8562	-1.9252	-1.3357	-0.31

表 3 各类数据标识统计表 ($K=15$)

标识	类
全“0”	1, 3, 6, 8, 15
全“1”	2, 7
“0”“1”混合	4, 5, 9, 10, 11, 12, 13, 14

3.3 检测结果

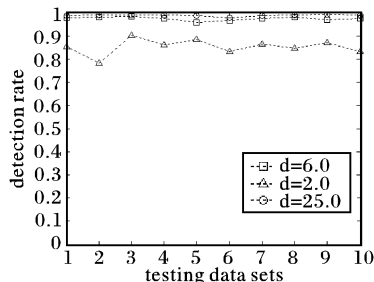


图 3 使用多项式核函数的检测率

支持向量机中的参数确定是非常重要的,实验中使用试验来选择参数,从较好的参数中选出最好的一个。图 3,4 给出了训练集 1 在不同核函数以及不同参数取值情况下的检测结果。

$$\mu = \frac{1}{N} \sum_{i=1}^N X_{ij} \quad (22)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_{ij} - \mu)^2} \quad (23)$$

3.2 数据集的构造

实验中使用了 5 个训练集,然后对 10 个测试子集做测试。5 个训练集中数据分布如表 1 所示。

表 1 5 个训练数据集结构

训练集	实例总数	正常实例	入侵实例
1	390	329	61
2	220	150	70
3	300	200	100
4	320	200	120
5	350	250	100

表 2 给出了当分类个数 $K=15$ 时,训练集 1 使用 K-MEANS 聚类算法得到的 15 个类的聚类中心,表 3 给出了 15 个类中样本标识的统计情况,其中“1”代表入侵数据,“0”代表正常数据。

从图 3 和图 4 的对比实验可以发现,不同核函数得到的检测率不同,在这里对 10 个测试集进行测试的结果是多项式核函数更适合数据集的数据分布情况。

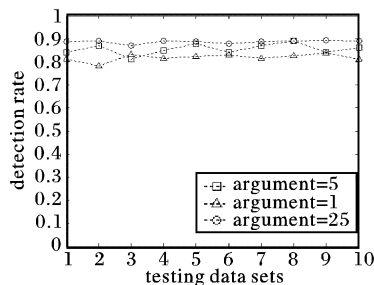


图 4 使用 RBF 核函数的检测率

4 结语

本文探讨了支持向量机在网络入侵检测方面的应用,提出了使用 K-MEANS 聚类算法进行初始划分来降低计算复杂度、提高检测效率,并给出了基于聚类和支持向量机的入侵检测框架和工作原理,利用 KDD CUP 1999 进行了仿真实验,所得到的性能比较令人满意。(下转第 2452 页)

d) 该方案中,秘密密钥是经过满足门限的参与者集合经过协商产生,不只依赖于其中一个人,和整个参与者集合的每一个人都有关系,即我们给出的是秘密密钥协商方案。

4 多重秘密密钥协商方案安全性分析

上述基于线性码理论构造的秘密密钥协商方案,要求协商各方拥有不对外公开的线性码和一个多项式。该方案简单易行,容易操作和实现。对于安全性讨论如下:

1) 方案能够防止第三方攻击。如果有一个第三方截获了参与者 P_{i_j} 发送给其他参与者的子秘密 $(K_1^j, K_2^j, \dots, K_n^j)$, 他需要向其他参与者发送一个自己的子秘密,该子秘密和其他子秘密模加后应是一个码字,相当于攻击者要在线性空间 $GF(q)^n$ 中找到一个码字。由于攻击者不知道线性码的生成矩阵,那么其成功的概率是 $\frac{q^k}{q^{n+1-k}} = \frac{1}{q^{n+1-k}}$ 。显然攻击者成功的可能性非常小。

2) 通过线性码校验矩阵的计算实现了多个参与者之间的相互认证。经过成功认证,满足门限个数的参与者集合就有了协商的秘密密钥。

3) 上述方案中,满足门限个数的参与者协商的秘密密钥是通过求解线性方程组: $g_0 = \sum_{j=1}^t x_j g_{i_j}$ 求得 $x_j; j = 1, 2, \dots, t$, 然后通过计算 $D = \sum_{j=1}^t d^j = \sum_{j=1}^t s^j g_0 = \sum_{j=1}^t x_j T_{i_j}$ 获得的。生成矩阵是参与者私有的,攻击者也不知晓。攻击者对求解线性方程组: $g_0 = \sum_{j=1}^t x_j g_{i_j}$ 一无所知,所以攻击者就是知道 $T_k = \sum_{j=1}^t K_k^j; k = 1, 2, \dots, n$ 也不能计算出 $D = \sum_{j=1}^t d^j = \sum_{j=1}^t x_j T_{i_j}$ 。故想获得真正的秘密密钥 k_1, k_2, \dots, k_r 的概率是非常低的。

4) 线性码 C 的生成矩阵若以 $G = (I_k, P), H = (-P^T, I_{n-k})$ 为例。容易看出,整个线性空间 $GF(q)^n$ 中有 $q^{k(n+1-k)}$ 个相互不等价的标准型生成矩阵,即有 $q^{k(n+1-k)}$ 个相互不等价的线性码 $[n+1, k; q]$ 。所以攻击者找到正确的生成矩阵 G 的概率是 $\frac{1}{q^{(n+1-k)}}$, 当然,攻击者找到正确的校验矩阵 H 的概率也是 $\frac{1}{q^{(n+1-k)}}$ 。

5 方案的优点

上述方案同其他类似方案比较具有如下优势:

1) 不论是 Diffie-Hellman 密钥交换协议还是后来的 STS 协议都要基于离散对数假设,而这样的假设至少到目前为止

是没有理论保证的。本文方案不需要这样的假设,只是基于线性码理论的特性而需要协商各方能够共同选取一个线性码。这样做带来的好处是秘密密钥协商各方如果认为有必要就可以经常更换协商的秘密密钥,而且可以防止第三方攻击。

2) 方案中用到的就是矩阵的相乘和加减运算,这些计算的计算量很小,且计算方便,容易操作。在计算复杂性上明显优于其他方案。

3) 方案有认证功能。每个参与者可以通过方案中的验证步骤对协商的秘密密钥的正确性作出判断。如果验证步骤中等式不成立,则说明有第三方攻击。

4) 上述方案具有门限性,即大于等于门限个数的参与者组成的集合才可以进行秘密密钥协商,这使得秘密密钥协商更具可操作性和广泛性。同时,该秘密密钥协商过程中,一次秘密密钥协商,可以产生多个秘密密钥。

6 结语

本文给出了一个有门限可认证的多重秘密密钥协商方案,该方案可以防止第三方攻击,也是计算安全的。该方案只有参与者集合满足门限的情况下可以进行秘密密钥协商,方案本身基于线性码理论,具有认证功能;同时,运用该方案,满足门限的参与者集合一次协商,可以产生多个通过协商的秘密密钥。

参考文献:

- [1] SHANNON C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27: 379-423; 623-656.
- [2] DIFFIE - HELLMAN R. RFC 2631, Key Agreement Method[S]. 1999.
- [3] DIFFIE W, VAN OORSCHOT P C, WIENER M J. Authentication and authenticated key exchanges[J]. Designs, Codes and Cryptography, Kluwer Academic Publishers, 1992, (2): 107-125.
- [4] 冯登国,裴定一. 密码学导引[M]. 北京: 科学出版社, 1999: 223-244.
- [5] VAN LINT J H. 编码理论导引[M]. 余敏安,陈冬生,译. 北京: 科学出版社, 1988: 37-40.
- [6] 王新梅,马文平,武传坤. 纠错密码理论[M]. 北京: 人民邮电出版社, 2001: 48-52.
- [7] MASSEY J L. Minimal codewords and secret sharing[C]// Proceedings of the 6th Joint Swedish-Russian Workshop on Information Theory. Sweden: Molle, 1993: 22-27.
- [8] MASSEY J L. Some applications of coding theory in cryptography [C]// Codes and Ciphers: Cryptography and Coding IV (ED. PG Farrell). England: IMA, 1995: 33-47.
- [9] 谭晓青. 无信任分配中心的动态秘密分享方案[J]. 湘潭大学自然科学学报, 2005, 27(4): 42-45.

(上接第 2442 页)

参考文献:

- [1] FORREST S, PERRELASON A S. Self-nonself Discrimination in a Computer[C]// Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. California: IEEE Computer Society, 1997: 202-212.
- [2] GHOSH A K, MICHAEL C, SCHATZ M. A Real-Time Intrusion Detection System Based on Learning Program Behavior[C]// Recent Advances in Intrusion Detection (RAID). Toulouse: Springer-Verlag, 2000: 120-132.
- [3] LEE W, STOLFO S J, MOK K W. A Data Mining Framework for

Building Intrusion Detection Models[C]// Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland: IEEE Press, 1999: 120-132.

- [4] VAPNIK V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.
- [5] 饶鲜,董春曦,杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798-803.
- [6] Lincoln Labs. KDD-cup data set[DB/OL]. [2004-12-02]. <http://kdd.ics.uci.edu/databases/kddcup99.html>.