文章编号：2096-1472(2016)-04-11-02

# 一种基于聚类分析的入侵检测模型

付明柏
（昭通学院信息科学与技术学院，云南 昭通 657000）

**摘 要**：针对传统的入侵检测技术在大容量网络数据时存在检测性能不足的缺点，研究了一种基于聚类分析算法的新型入侵检测模型，通过聚类分析算法对多维数据进行分析，当不满足聚类要求时，归并邻近数据再次聚类。最后，设计了与K−means算法的对比仿真实验，实验结果表明，基于聚类分析的模型能够有效检测出异常序列，能够抵抗异常攻击。

**关键词**：入侵检测；聚类分析；网络安全
**中图分类号**：TP393.08 **文献标识码**：A

## A Intrusion Detection System Based on Cluster Analysis

FU Mingbai
( *School of Information Science and Technology*,*Zhaotong University*,*Zhaotong* 657000,*China*)

**Abstract:**Since conventional intrusion detection systems can't meet high demands of the network security,a new intrusion detection method based on clustering algorithm for intrusion detection system is designed in order to cluster analysis high dimensional data,and merge data nearly if cluster condition is not qualified.After stimulate experiment compared with K-means algorithm,the result shows this detection model can detect abnormal attack effectively.

**Keywords:**intrusion detection;cluster analysis;network security

## 1 引言(Introduction)

随着计算机与现代通信技术的高速发展，互联网已经成为人们生活中不可或缺的部分，以及随着3G、4G通信技术的发展，移动互联网在近几年呈现爆炸式增长。由于互联网的快速性、便利性和及时性，各大企业以及政府单位也纷纷成立其官方网站，从事相关的电子政务、电子商务以及网上办事系统等工作[1]。

然而，也正是由于互联网的开放性导致了其非常容易受到攻击，并造成重大损失。由于互联网的开发性和及时性，一些敏感文件，尤其那些具有商业价值或和数据安全紧密相关的数据，非常容易被误操作泄露在网络上，进而造成大规模数据外泄。即便是没有外泄，有价值的敏感数据也经常被黑客团体进行入侵攻击。近年来，网络黑客的入侵攻击事件，呈现出急剧增加的趋势。不仅企业和政府网站被攻击次数增多，针对中小企业和个人的网络攻击也呈现出增加势头。根据世界著名计算机安全厂商迈克菲《安全悖论》报告分析，超过83%的企业表示担心或非常担心自己的企业成为恶意攻击的目标。实际上，51%的企业已经遭受过各种攻击，他们当中的16%要耗费超过一周的时间，相应网络系统才可以恢复正常运营。其中，数据的丢失是安全攻击所造成的最严重后果[2]。目前，网络安全已经成为研究的热点也是各界瞩目的焦点。

一般情况下，企业采用防火墙作为其网络安全的第一道防线，但是随着网络攻击手段的逐渐成熟与多样化，尤其近几年发展尤为迅速，传统的防火墙机制已经无法保障大多数企业网络环境的安全。因为防火墙系统是被动的、静态的网络攻击和防御体系，而移动通信和互联网技术发展日新月异，相应的新的服务和协议也不断地出现，传统防火墙已经不能很好对其进行很好的动态扩展，而且防火墙技术也无法检查传输层以上的数据内容，所以许多网络攻击程序可以轻易地越过企业或者单位设置的防火墙，实施攻击行为。

## 2 入侵检测模型(Intrusion detection system)

### 2.1 入侵检测原理

入侵检测技术是指从计算机网络的不同环节中收集数据并对其分析和处理，找出其中是否存在恶意入侵的企图或者违背安全策略的行为，这个安全策略可以是事先根据特定网络环境设定好的。它可以针对计算机网络系统的运行情况依照安全策略进行监控，确保网络系统中的资源不被恶意攻击。

入侵检测系统的工作流程一般包括三大步骤[3]：(1)信息收集，从系统的不同网络环节中收集数据、用户连接的行为以及连接状态等信息，为下一步做准备。(2)信息分析，利用第一步收集到的信息，由设计好的算法进行分析和提取出存在的入侵特征。(3)根据第二步的分析结果，判断是否存在相应的入侵行为，记录日志并发出警告，网络系统管理人员可以做出相应的处理。

### 2.2 典型的检测模型

入侵检测系统有多种检测模型，目前常用主要有以下几种：统计方法、模式匹配、神经网络、数据挖掘、协议分析和免疫系统等[4,5]。

模式匹配的检测方法是通过将入侵的行为转化为计算机可以解析和识别的模式和特征向量，然后根据模式匹配搜索入侵特征数据库，如果遇到新的入侵特征，就将其加入特征数据库中。

基于概率统计的入侵检测系统通过对网络系统日常行为

不断进行自主学习，构建一个正常活动集的集合，如出现偏离正常活动比较大的行为，就将其作为异常活动的内容。

基于数据挖掘的入侵检测方法，就是从存放大量数据信息的数据库等信息库中挖掘用户的历史行为特征，并根据历史行为特征判定是否为异常活动的过程。

### 2.3 现有模型的缺点

入侵检测系统是网络安全防护的重要手段，能够有效地保护网络安全。但是目前入侵检测系统在以下方面也存在着很多问题，尤其当网络数据流量较大时更加明显，主要包括：误报/漏报率较高，产品适应能力差，检测性能不足，同时检测实时性较差，缺少主动防御功能等等[6]。

## 3 聚类分析检测模型研究(Research on detection model of cluster analysis)

### 3.1 聚类分析算法

由于现有的入侵检测技术在大容量数据流量时存在着数据的检测效率低以及检测性能差等缺点。如何从大量的网络数据中提取出有用的信息是如今研究的难点和热点。

聚类算法是一种由若干模式组成的分类算法，通常，模式是指一个度量的向量。聚类分析是以相似性为基础的，它在一个聚类中的模式之间的相似性比在不同聚类中的模式相似度之间要高，根据原始数据的分布特征进行分析，将初始的子空间特征向量通过模糊函数映射到高维矩阵中，然后再对该高维数据矩阵处理进行类别划分工作。聚类算法通常情况下适用于大数据量的处理，而且由于采用了指数迭代的思想，该算法有效性很高。本文采用矩阵空间加权的聚类算法就可以很好地处理高维数据，经过多次聚类分析，数据的精度也会大幅提高，因而其在处理大数据量的时候优势尤为明显，聚类算法的主要工作流程如图1所示。

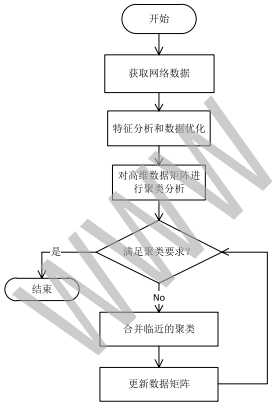


图1 聚类分析算法工作流程

Fig.1 The working follow-step of cluster analysis

### 3.2 聚类分析模型

在矩阵特征向量的表示中，规则的频率用支持度来表示。那么，$T$包含$K$的支持度就是可以表示为矩阵样本数据T中闭包K的事物数，用式(1)表示如下：

$$S(X,Y)=\frac{1}{n\times k}\left(\sum_{t_j\in T}\sum_{i_p\in(X\cup Y)}w[t_j][t_p]\right) \tag{1}$$

$K$表示词项$X$与词项$Y$关联的项目数，$n$表示矩阵中记录数。

在矩阵特征向量中，可信度可以表示为其关联规则度，定义为：

$$C(X,Y)=\frac{S(x\cup y)}{S(x)} \tag{2}$$

支持度和可信度要大于设定的最小阈值。包括相机的最大权值之和$w(I_i,k)$表示为：

$$w(I_i,k)=\left(\sum_{t_j\in T}\sum_{i_p\in(X\cup Y)}w[t_j][t_p]\right)+F(I_i)\sum_{l=1}^{k-q}w_{rl} \tag{3}$$

式中，$I_i$是$q$项集，$F(I_i)$为$I_i$项集出现的频率。

最后可以得出包含$q$项集的$I_i$项集权值阈值$K(I_i,k)$表示为：

$$K(I_i,k)=n\times k\times\min S-F(I_i)\sum_{l=1}^{k-q}w_{rl} \tag{4}$$

在特征空间中，函数$K(x,y)=(\phi(x),\phi(y))$的内积可以表示为特征向量$\phi(x)$和$\phi(y)$，所以目标函数可以表示为：

$$H(u,w)=\sum_{i=1}^{m}\sum_{j=1}^{n}u_{ij}^{m}D_{ij}^{2} \tag{5}$$

$w$可以表示为$(w_1,w_2,w_3\cdots,w_n)$，是特征聚类的中心。数据$x_j$映射到$w_i$的距离用$D_{ij}=\|\phi(x_j-w_i)\|$表示，引入极大范围函数结合极值条件可以得到式(6)。

$$w_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\bullet\phi(x_j)}{\sum_{j=1}^{n}u_{ij}^{m}},1\le i\le n \tag{6}$$

有高斯函数$G(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma})$代入式(3)得出目标函数

$$H(u,w)=\sum_{i=1}^{m}\sum_{j=1}^{n}u_{ij}^{m}\|\phi(x)-\phi(y)\|^{2} \tag{7}$$

再通过极值条件得出：

$$w_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}G(x_j,v_i)x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{8}$$

最后使用矩阵加权关联规则进行聚类。

### 3.3 聚类分析检测过程

本文设计的聚类算法的核心思想是：根据适当的阈值判断数据是否属于频繁数据来进行聚类分析，如果满足聚类条件则把频繁数据加入到新的项集中，并对非频繁的数据进行聚类为新的数据集重新进行判定。当矩阵加权集为空时迭代结束。该算法的主要步骤如下：

a.扫描初始矩阵，检索具有频繁数据项目的记录数、项目分类数以及矩阵样本关联总数。

b.根据关联算法计算相应词项的相似度，并记录下候选数据的记录次数。

c.根据记录的候选数据记录的权值、次数和阈值计算进行聚类分析。

d.若候选数据记录不为空时，回到第一步继续执行。否则，判定该候选数据为频繁数据，将其记录到新矩阵空间中。

e.根据上述算法来计算词项的关联关系并将其输出到新矩阵空间中。

对计算得到的空间矩阵进行分析，并可以通过再次数据聚类的方式来提高聚类算法的精度。

(下转第10页)