基于改进 *K* 均值算法的入侵检测系统设计

刘华春,候向宁,杨 忠
(成都理工大学 工程技术学院,四川 乐山 614007)

摘 要:传统的入侵检测系统是将规则库与网络数据包逐一匹配,进行检测,当网络数据量巨增时,检测效率显著降低,甚至面临不能即时检测的巨大挑战。数据挖掘是从海量的数据中挖掘发现需要的各种有价值信息的技术,入侵检测系统中植入数据挖掘技术,将极大提高入侵检测系统的检测效率和智能性。研究了数据挖掘中 *K* - means 聚类算法应用于入侵检测领域中的难点问题。*K* - means 算法具有易受初始 *K* 值和孤立点影响,难以确定 *K* 值,对初始质心依赖程度高等不足问题。针对上述缺点,提出了改进的 *K* - means 聚类算法。设计了基于改进 *K* - means 的入侵检测系统并进行了实验。结果表明,将改进的聚类算法应用于入侵检测可显著提高异常检测效率;可自适应地建立入侵检测异常模式库;对未知的入侵攻击能有效防范;能进一步降低误检率。

关键词:数据挖掘;入侵检测;聚类算法;异常检测
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2016)01-0101-05
doi:10.3969/j.issn.1673-629X.2016.01.021

Design of Intrusion Detection System Based on Improved *K*-means Algorithm

LIU Hua-chun, HOU Xiang-ning, YANG Zhong
(Engineering & Technical College of Chengdu University of Technology, Leshan 614007, China)

Abstract: Traditional intrusion detection system is matched to the rule base and network packet one by one. When the network is the huge increase in the amount of data, detection efficiency significantly reduces, even in the face of enormous challenges not immediately detected. Data mining is a technology finds a variety of valuable information from the mass of data, data mining technology into the intrusion detection system will greatly improve efficiency and intelligence of this IDS. Focus on researching the *K*-means clustering algorithm in data mining for application to intrusion detection system. The *K*-means algorithm has some shortcomings, such as to be affected by the initial *K* value and outlier, difficulty of determining *K* value, highly depending on the initial center point. To overcome these disadvantages, an improved *K*-means clustering algorithm is proposed. And an intrusion detection system based on this is designed. The results show that the improved clustering algorithm is applied to intrusion detection, it can significantly improve the abnormality detection efficiency, and adaptively establish the abnormal pattern database of intrusion detection, and effectively prevent the unknown intrusion and greatly reduce the false detection rate.

Key words: data mining; intrusion detection; clustering algorithm; anomaly detection

0 引 言

传统的入侵检测系统(Intrusion Detection System, IDS)是采取分析和提取入侵模式和攻击特点,建立检测规则库及模式库,所以传统 IDS 在检测效率和智能性上存在明显不足。在网络带宽快速提高,入侵和攻击模式不断变化的新形势下,传统 IDS 的检测方式、检测效率面临巨大挑战,甚至不能即时响应和检测。数据挖掘(Data Mining, DM)能够从海量数据中根据不同的挖掘算法,挖掘出具有不同用途的知识和信息。因此,可以将数据挖掘技术植入到 IDS 中,应用适当的挖掘算法,就可解决前文提出的 IDS 效率和自适应问题。目前,DM+IDS 已成为入侵检测领域的一个重要研究方向。数据挖掘应用于入侵检测系统的研究,在国内外都有很多的研究机构及大学在进行,已取得了一定的研究成果,但总体仍处于初始阶段。

文中将数据挖掘技术应用于入侵检测系统,对于

收稿日期:2015-03-21 **修回日期**:2015-06-23 **网络出版时间**:2015-11-19
基金项目:四川省自然科学重点项目(A22012003)
作者简介:刘华春(1966-),男,硕士,副教授,研究方向为信息安全、机器学习。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20151119.1111.052.html

入侵检测系统具有较大的实际应用价值。

1 入侵检测与数据挖掘

1.1 入侵检测技术

入侵检测的原理是通过从网络中特定点收集和分析网络数据,以判别该网络中是否存在被攻击或违反安全策略的行为。入侵检测系统对网络进行实时监测和控制,所以能够提供对各种错误配置和来自网络内部、外部攻击的防范[1-2]。入侵检测系统能及时发现入侵行为,并产生警报信号,因此极大地提高了网络系统的安全性。

入侵检测工作过程主要由数据采集、数据分析和响应三个步骤组成。美国互联网工程任务组(IETF)为入侵检测系统制定了标准,并发起制订了系列的建议草案[3],提出了入侵检测系统框架模型。此模型把一个入侵检测系统分解为事件产生器、事件分析器、事件数据库和响应单元四个部分[4]。事件产生器进行网络数据的抓取和预处理,事件分析器进行规则的分析匹配,事件数据库存放规则模式,响应单元产生动作执行操作。根据采用的检测方法,入侵检测技术可分为异常检测和误用检测。

1.2 数据挖掘

数据挖掘又称数据库中的知识发现(Knowledge Discover in Database,KDD),能够从大量的、海量的数据中提取出未知的、并具有用户期望价值的信息。数据挖掘技术已广泛应用于机器学习、模式识别、人工智能、统计学等领域,是一个决策支持的过程。数据挖掘高度自动地分析海量数据,进行推理、归纳,挖掘出潜在的模式和规则,用户根据挖掘结果调整策略,进行决策,可有效降低风险,提高决策的正确率[5]。数据挖掘的过程,根据其工作内容,可分为数据准备、数据挖掘、挖掘结果的解释与评价三个阶段,也是针对具体应用项目的数据分析和处理过程[6]。应用于不同领域的数据挖掘,其数据内容、数据格式、挖掘算法,应根据具体的挖掘目标而进行设计。数据挖掘技术可分为以下几种类型:关联规则、序列模式、分类、聚类等[7]。

在传统的入侵检测系统中植入数据挖掘技术,研究探索适当的数据挖掘算法,通过从海量网络数据中,过滤掉正常数据模式,只提取异常入侵模式,智能地构建入侵检测模型,就可以极大地提高传统入侵检测系统的检测效率,并拓展其自适应性,从而降低传统 IDS 的误检率[8]。

2 聚类算法 *K*-means 研究

2.1 原始 *K*-means 算法

K - means 算法的主要思想是将输入数据按照一定的方法划分到不同的类中,在同一个类中的数据,数据特征具有最大的相似性,在不同类中的数据,其数据特征具有最大的相异性[9]。

若有数据集 D,其中有 N 个数据,每一个数据 X_i 有 q 维特征,由 q 维特征属性描述一个数据,即 $X_i = (X_{i1}, X_{i2}, \cdots, X_{iq})$,$X_i \in D$,$1 < i < N$。若聚类个数为 K,需满足条件 $K < N$。算法流程为:

(1)从数据集 D 中随机选择一个数 K(需满足 $K < N$),K 即为聚类后的个数,即有聚类中心 $r_1, r_2, \cdots, r_k$。设迭代计数器为 s,其初值 $s = 1$。

(2)第一个聚类中心为第一个数 x_i。计算每一个数据 x_i 到各个聚类中心的距离 $d(x_i, r_j)$,$1 < j < k$,比较这些距离 d 的大小,将 x_i 划分到距中心最近的类中,得到了聚合的 k 个类 $C_1, C_2, \cdots, C_k$。

(3)当增加一个数据到类中后,计算聚类中所有数据的属性均值,重新得到了新的聚类中心。

(4)计算准则函数。

(5)用准则函数是否收敛判别是否要继续,如果收敛,转到结束;如果不收敛,返回到第(2)步,进入新的一轮迭代过程,迭代计数器 $s = s + 1$。

(6)结束,显示聚类结果。

K - means 算法具有很多优点:算法简单;容易理解,易实现;能快速处理较大量的数据;当各个类相差明显时,能快速识别;算法的复杂度低,为线性的;具有良好的扩展性[10-11]。

K - means 算法存在如下缺陷:

(1)初始化聚类中心 K 值,对聚类结果的影响较大,选取不同的 K 值,得到的聚类结果有较大差异。而 K 值通常需要进行实验确定,也可根据经验来确定,没有一个通用的方法来确定。当 K 值取法不当时,会导致聚类结果的质量下降[12]。

(2)孤立点对聚类结果有较强影响,而且,在聚类算法处理时,数据的输入顺序会影响聚类结果[13]。

(3) *K* - means 算法中数据对象之间的距离是用欧氏距离来表示。这样只能处理连续型数值而不能处理离散型数据对象[14]。网络数据中一些数据特征值是连续型数值,而一些是离散型的,如数据帧标志、类型等。*K* - means 算法无法直接处理这些离散特征数据。

2.2 IDS *K*-means 算法

由于 *K* - means 算法的缺陷,不能直接应用于入侵检测,文中将对其进行改进,将改进的 *K* - means 算法称为 IDS *K* - means 算法。

2.2.1 IDS *K* - means 算法设计

对于聚类个数 K 值确定困难的缺陷,提出一种预定距离的聚类算法。该算法的思路为,预先确定一个

聚类半径 r,第一个聚类中心以第一个数据为中心。第二个数据获取后,计算与前聚类中心的距离,若小于 r,则将第二个数归到这个类中,重新计算该类的中心数值。若大于等于 r,以第二个数据作为一个新类的中心。依次类推,后面到达的数据,计算其与已有各个类中心的距离,小于 r 则归入该类,大于 r 则为一个新类中心。

对于提高检测效率,确定聚类结果是正常数据模式还是入侵数据模式的问题。将正常聚类模式和异常聚类模式分别放在正常行为表和异常行为表中。预先设定一个阈值参数 β,当某一类成员的数目与所有成员比例大于或等于 β 时,表明该类是一个正常数据类,反之为入侵数据的聚类。由于在网络中,正常数据的数量远大于入侵数据的数量;将正常数据过滤掉,只保留异常的疑似入侵的数据进行下一阶段的检测,可以极大提高检测效率。

对于传统 K – means 聚类算法只能处理数字量,而无法处理离散量的问题,将离散属性转化为 0 和 1 的数值属性,采用离散属性值出现的频率进行量化,把最高的值作为聚类中心的值,再利用 K – means 算法进行聚类分析。

2.2.2　IDS K – means 算法流程

Step1:

input:训练数据和半径参数 r;

output:训练数据的聚类 $C_1, C_2, \cdots, C_k$。

算法流程:

(1)将输入的训练数据集 T 归一化预处理,减少特定较大数据对聚类结果的影响。

(2)读入数据集 T 中的第一个数据 X_1,以 X_1 为中心值,构造聚类 C_1。

(3)重复(2),读入下一个数。

(4)读入数据集 T 后续的数据 $X_i, i = 1, 2, \cdots, n$。计算每一个数 X_i 与已有的类 C_j 中心值的距离 $d(X_i, C_j)$。

(5)若 $d(X_i, C_j) \leqslant r$,将 X_i 归入到 C_j 类,即 X_i 属于 C_j 类中;再新计算 C_j 类的中心值,将 C_j 类的成员加 1。

(6)若 $d(X_i, C_j) > r$,将 X_i 作为中心值,创建一个新的类。

(7)重复输入的数据,直到全部数据结束。

Step2:

input: $C_1, C_2, \cdots, C_k$,阈值 β。

output:正常数据的聚类和异常数据的聚类。

算法流程:

(1)若某一个聚类中,其成员数目与全部数据之比大于或等于参数值 β,则该类为正常行为数据的聚类,将其移入正常聚类表,构造正常行为模式库。

(2)若某一类中,其成员数目与全部成员之比小于参数值 β,则该类为异常行为数据的聚类,把其放入异常聚类表,构建异常行为模式库。

Step3:

孤立点的处理,文中采用基于统计的方法,对聚类算法运行后,生成的每一个类 i,计算类 i 中数据成员所占的比率 $q(i)$ 值,根据 $q(i)$ 值进行排序,$q(i)$ 值越小,表明 i 中的成员数据越不适合这个类,可能是个孤立点,取 $q(i)$ 值最小的类 i 作为孤立点,从该类中删除。然后将孤立点重新进行聚类,直到所有孤立点数据全部放到合适的类中为止。这样能有效减少因输入数据的顺序而形成孤立点后,对聚类结果的影响。

3　基于数据挖掘的入侵检测系统

系统设计遵循通用入侵检测系统模型(CIDF),文中在 CIDF 模型的基础上,引入数据挖掘技术,将改进的聚类算法 IDS K – means 应用于入侵检测,增加了聚类分析模块。

3.1　系统结构

该入侵检测系统的结构设计,包括通用 IDS 结构的部分,即事件产生器、事件分析器、事件数据库、响应单元部分外,还包括数据挖掘模块部分,即聚类分析、关联规则分析共计六大模块,如图 1 所示。

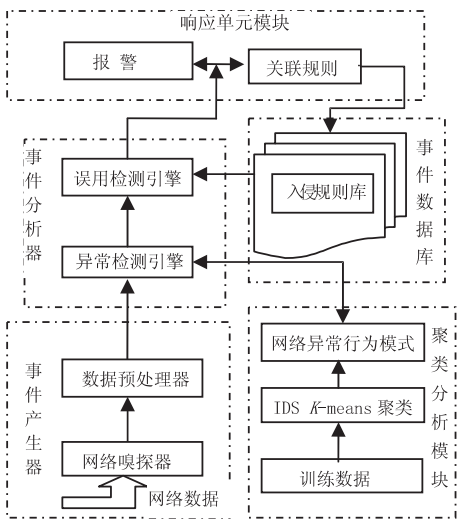

图 1　基于改进聚类算法的入侵检测结构

各模块详细功能如下:

事件产生器:包括数据包嗅探器和预处理器两个子模块。从网络中捕获数据包,并将获取的数据包进行分析解码处理后,供后面的模块使用。

聚类分析器:采用 IDS K – means 算法构建网络正常行为模式库和异常行为模式库。

事件数据库:存放异常入侵规则模式数据,并维护异常入侵规则数据,供误用检测和关联规则进行模式

检测。

事件分析器:分析和处理网络数据,包括异常检测和误用检测两个模块。实现过滤和检测双重功能。

(1)过滤功能:异常检测模块通过网络正常行为模式和异常模式对输入的网络数据进行模式识别,把正常的网络数据过滤,保留异常网络数据送误用检测。

(2)检测功能:误用检测将异常检测过滤后通过的疑似入侵数据与异常事件数据库中的入侵规则进行检测,判断该数据是哪一类入侵数据。

响应单元:当误用检测为异常数据时,产生入侵行为触发,让 IDS 产生动作,阻止入侵行为继续发生,通过报警,记录到日志文件,通知防火墙切断该连接,通知管理员等。

关联规则分析:将入侵的网络数据进行关联挖掘,挖掘出入侵行为与异常数据之间的关联关系,并将其转化为入侵规则,添加到入侵规则库中。

3.2 工作流程

该入侵检测系统的工作流程设计为两个阶段,分别为训练和检测阶段,如图 2 所示。

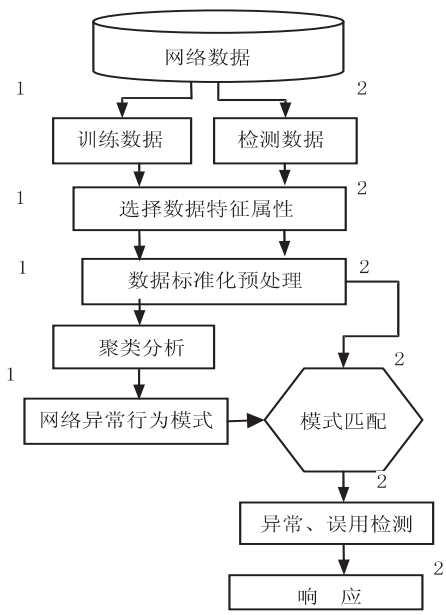

图 2 入侵检测系统工作流程

1)训练阶段:如图 2 中 1 流程所示,系统在训练阶段要将大量的网络数据作为训练数据存入数据库。

(1)根据数据取出关键特征进行预处理。

(2)采用 IDS K - means 聚类算法对数据进行聚类分析。

(3)提取网络数据正常模式和异常入侵数据模式。

(4)过滤正常网络数据。

2)检测阶段:如图 2 中 2 流程所示。

(1)输入网络数据。

(2)对数据进行预处理。

(3)正常网络数据过滤,将网络数据与模式库中的数据进行匹配,如果为正常数据,过滤掉,提高系统的检测效率。

(4)将入侵数据送误用检测,判断该入侵数据为哪一类攻击。

(5)触发响应模块,报警。

(6)如没有与入侵规则库匹配成功,该数据为未知攻击类型,则由关联规则挖掘出攻击行为与数据的关系,将其添加到入侵规则库中,使系统具备了发现未知攻击的能力。

3.3 仿真实验

3.3.1 实验设计

实验数据采用 KDD CUP99 数据集[15],通常采用该数据来对设计的 IDS 进行各种性能测试。其中的所有数据都是在实际运行的互联网环境下,模拟真实攻击的情景得到的数据,数据格式如下:

0,tcp,http,SF,51,8127,0,0,0,2,0,1,0,1,0,0,0,0,1,0,0,0,1,2,0.00,0.00,0.00,0.50,1.00,0.00,1.00,255,255,1.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,phf

该数据集的每一个数据由 42 个属性值构成,其中有数值型属性,也有离散型属性,第 42 个属性标识该数据记录是正常行为产生的数据,还是入侵行为产生的数据。

该实验的目的是验证改进的 IDS K - means 算法的有效性及性能分析。实验中,采用数据集的 10% 来测试设计的 IDS 系统的各种性能,把实验数据随机分割成 S_1 和 S_2 两个子集。将 S_1 作为训练数据集,用于训练 IDS 构建正常网络数据模式和入侵网络数据模式,S_2 作为测试数据集,其中包含有 S_1 中没有出现的网络攻击数据,用来检测该 IDS 的检测能力。

将实验数据导入到 SQL Server 数据库中,建立数据库,并新建训练数据表和测试数据表。

3.3.2 实验结果分析

采用误检率来评价该 IDS 检测效果的度量指标,公式如式(1)所示:

$$\text{误检率} = \frac{\text{丢弃的攻击数据包的数量}}{\text{攻击数据包数量}} \tag{1}$$

在实验中,经过多次选取不同参数 r 和 β,测试该 IDS 的检测性能。r 为聚类半径参数指标,β 为孤立点阈值参数指标。

(1) r 对实验结果的影响($\beta = 0.4$)。

设定孤立点阈值 $\beta = 0.4$ 时,通过改变不同 r 的值,其结果如表 1 所示。

从表 1 可以看出,当聚类半径 r 越大时,误检率越高,表明 IDS 将入侵数据当成正常数据,会把攻击数据

过滤,不做检测,即检测不出入侵数据,容易造成漏检。

表 1　不同 r 值对结果的影响

聚类半径 r	误检率/%
1	0
5	0.2
10	8.58
20	29.9
30	71.51
40	94.67

分析:聚类半径 r 直接影响后的结果。当 r 越大时,聚类后的类数量就越少,这样入侵数据被当成正常数据的几率越大。当 r 越小时,聚类后类的数量就越多,数据匹配就更细化,入侵数据被当成正常数据的几率就越小。所以,聚类半径 r 对误检率有非常显著的影响,r 越小,误检率越低。

(2)不同 β 值的影响(设定聚类半径 $r=5$)。

对数据进行多次聚类分析,设定聚类半径 $r=5$,不断变化阈值 β,如表 2 所示。

表 2　孤立点阈值 β 的影响

阈值 β	误检率/%
0.1	35.89
0.2	25.05
0.3	4.85
0.35	0.21
0.4	0.08

从表 2 可以看出,孤立点阈值 β 越小,误检率越高,β 越大,误检率越低。

分析:孤立点阈值 β 对误检率也有很大的影响,β 值是某一类成员数与全部数据的比率。当 β 越小时,表明更多的入侵数据被当成正常数据类,误检率就高。反之,孤立点阈值 β 越大,更多的数据要进行再次聚类,入侵数据被当成正常数据的可能性越低,这样,误检率就越低。

从上述实验结果及分析可知,采用该 IDS K – means 聚类算法的入侵检测系统,聚类半径 r 和孤立点阈值 β 直接影响聚类结果,从而对 IDS 检测结果产生重大影响,合理选择聚类半径 r 和阈值 β 直接关系到 IDS 系统的检测性能。在实际应用的入侵检测系统中,需要根据具体情况调整合适的 r 和 β 参数值,以达到满意的检测效果,即提高检测效率,降低误检率。

4　结束语

文中研究了在入侵检测系统中植入数据挖掘的聚类技术,达到提高检测效率、降低误检率的目标。详细研究了聚类算法 K – means 的流程、优点及不足,创新性地提出了根据聚类半径 r 和阈值 β 进行聚类的改进 IDS K – means 算法。设计了基于 IDS K – means 算法的智能入侵检测系统结构,并采用模拟网络攻击数据包 KDD CUP 99 对系统进行了实验测试。研究结果表明数据挖掘技术应用于入侵检测系统可有效地提高异常检测效率;能够自适应建立入侵检测异常模式库,对未知的入侵攻击能有效防范;调整合适的聚类半径 r 和阈值 β,能达到较好的检测效果。

参考文献:

[1] 郭红艳,谷保平. 改进 k 均值算法在网络入侵检测中的应用研究[J]. 计算机安全,2008(5):24-26.

[2] 刘　静. 基于聚类的网络入侵检测的研究[D]. 太原:太原理工大学,2008.

[3] 秦子燕. 基于聚类分析的入侵检测方法研究[D]. 无锡:江南大学,2008.

[4] Sabahi F, Movaghar A. Intrusion detection: a survey[C]// Proc of the third international conference on systems and networks communications. [s. l.]:[s. n.],2008:23-26.

[5] 李　洋. K-means 聚类算法在入侵检测中的应用[J]. 计算机工程,2007,33(14):154-156.

[6] 张建萍,刘希玉. 基于聚类分析的 K-means 算法研究及应用[J]. 计算机应用研究,2007,24(5):166-168.

[7] 陈小辉. 基于数据挖掘算法的入侵检测方法[J]. 计算机工程,2010,36(17):72-73.

[8] Gaddam S R, Phoha V V, Balagani K S. K-Means+ID3: a novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods[J]. IEEE Transactions on Knowledge and Data Engineering, 2007,19(3):345-354.

[9] 李文华. 基于聚类分析的网络入侵检测模型[J]. 计算机工程,2011,37(17):96-98.

[10] Ensafi R, Dehghanzadeh S, Mohammad R, et al. Optimizing fuzzy K-means for network anomaly detection using PSO [C]//Proc of ACS/IEEE international conference on computer systems and applications. Doha, Qatar: IEEE, 2008: 686-693.

[11] 杜　强,孙　敏. 基于改进聚类分析算法的入侵检测系统研究[J]. 计算机工程与应用,2011,47(11):106-108.

[12] 吴庆涛,邵志清. 入侵检测研究综述[J]. 计算机应用研究,2005,22(12):11-14.

[13] 宋宇翔,刘　琰. 特征和分类器联合优化的网络入侵检测算法[J]. 计算机工程与应用,2012,48(19):77-81.

[14] 朱广彬. 基于数据挖掘的入侵检测技术研究[D]. 北京:北京交通大学,2010.

[15] University of California, Irvine. KDD cup 1999 data[EB/OL]. (1999-10-28)[2012-03-20]. http://kdd. ics. uci. edu/databases/kddcup99/kddcup99. html.

基于改进K均值算法的入侵检测系统设计

作者： 刘华春， 候向宁， 杨忠， LIU Hua-chun， HOU Xiang-ning， YANG Zhong
作者单位： 成都理工大学 工程技术学院,四川 乐山,614007
刊名： 计算机技术与发展 ISTIC
英文刊名：
年，卷(期)： 2016(1)