第 16 卷 第 2 期 莆田学院学报 Vol.16 No.2
2009 年 4 月 Journal of Putian University Apr. 2009
文章编号:1672-4143(2009)02-0068-03 中图分类号:TP391 文献标识码:A

一种基于聚类和关联规则修正的入侵检测技术

黄 斌[1], 史 亮[2], 陈德礼[1]
(1. 莆田学院 电子信息工程系, 福建 莆田 351100;2. 厦门大学 软件学院, 福建 厦门 361005)

摘 要:针对目前基于 K-Means 算法的入侵检测技术所存在的符号类型数据处理能力欠缺、误报率较高的问题,提出了一种基于聚类和关联规则修正的入侵检测技术。将关联规则挖掘技术引入到聚类分析机制中,利用针对符号型属性的关联规则挖掘结果对聚类结果进行修正,从而有效降低由于在入侵检测单纯使用聚类分析所导致的误报。详细阐述了改进的具体实现方案,并通过实验验证了该技术的可行性。

关键词:入侵检测;聚类算法;关联规则

An Intrusion Detection Method Based on Clustering and Association Correction

HUANG Bin[1], SHI Liang[2], CHEN De-li[1]
(1. Electronic & Information Engineering Department, Putian University, Putian Fujian 351100, China;
2. Software School, Xiamen University, Xiamen Fujian 361005, China)

Abstract: This paper analyses the existing problems of the current intrusion detection techniques base on K-Means Algorithm: failing to analyse the attribute composed by character, higher false-detection rate, etc, and brings forward some improvement: We use Association Rule into clustering analysis to reduce the false-detection rate in our algorithm. In this paper, we introduce the improved method concretely, and shows the feasibility and effect through an experiment.

Key words: intrusion detection; clustering algorithm; Association Rule

0 引言

基于数据挖掘的入侵检测技术的研究已经成为入侵检测领域的一个重要研究方向,其中基于聚类的入侵检测技术引起了研究人员的广泛兴趣[1-2]。

K-Means 算法作为一种常用的聚类算法,在大数据集处理上具有较好的可伸缩性、高效性和良好的扩张性,因此被广泛应用于包括入侵检测等诸多领域。但 K-Means 对噪声和异常点很敏感,即使是少数这样的数据对平均值的影响也很大,如果直接将聚类结果用于入侵检测,会导致较高的误检率。此外,该算法所处理的数据对象仅限于数值类型,对符号类型无能为力,而在实际应用中,特别是对于入侵检测领域,需要处理的数据往往既包括数值类型,也包括符号类型。当利用K-Means 算法处理这种混合型数据时,通常的解决方法是转换符号类型数据为数字值,但由于符号类型数据问题域的顺序遭到破坏,得出的结果意义不大。

针对基于聚类的入侵检测技术中所存在的不足问题,本文提出一种基于聚类结合关联分析的入侵检测技术,其基本思想是在聚类分析过程中引入

收稿日期:2008-10-30
基金项目:福建省自然科学基金项目(2008F50602) 福建省自然科学基金 - 青年人才项目(2008F3101)
作者简介:黄斌(1981-),男,福建莆田人,助教,硕士。

关联分析算法,利用关联规则挖掘技术分析网络连接记录中的符号属性,得出各符号属性间的关联关系,并用该结果对聚类分析的结果进行修正,从而克服单纯使用聚类分析的不足,提高检测效果。

1　基于聚类的异常检测

聚类算法是一个将数据集划分成若干个聚类的过程,使得同一聚类内的数据具有较高的相似性,而不同聚类中的数据不具有相似性。相似或者不相似根据描述数据的属性值来度量,通常使用基于距离的方法。通过聚类,可以发现数据的密集和稀疏的区域,从而发现数据整体的分布模式,以及数据属性间有意义的关联。

运用聚类算法进行网络入侵检测的过程可分 3 个阶段,分别是收集分析数据、对分析数据进行标准化和用聚类算法对数据分类。过程如图 1 所示。

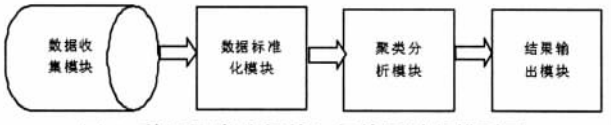

图 1　基于聚类分析的入侵检测过程模型图

利用聚类进行异常检测是基于以下两个事实:1) 入侵行为和正常行为存在的差异;2) 现实应用中异常行为的数量要远低于正常行为的数量。

首先我们用没有标记而含有少量入侵攻击的训练数据进行聚类,对数值型属性进行标准化[3],然后利用聚类算法 (如 K-Means) 得到一个聚类集。根据上述的两个事实我们知道,正常数据的数量应远大于入侵数据且相互间存在明显差别,由此可以对聚类集中的类别进行标记,选出其中数量较多的聚类,认为其中的数据是正常的并将其标记为正常集。在实际检测过程中,对于一条新的数据,测量其与聚类集中的各个聚类之间的相似度,找出相似度最大的聚类,查看该聚类的标记,如果聚类的标记是正常集,那么该数据就认为是正常数据,反之则认为其是入侵数据。

2　聚类分析结合关联规则

针对 K-Means 算法处理符号类型数据能力不足的问题,我们提出一种利用关联规则挖掘技术来对聚类结果进行修正的方案。具体到入侵检测领域来说,我们使用关联分析算法来分析网络连接记录中符号属性,把聚类分析与关联分析结合起来对网络连接数据进行检测,这样可以克服单纯使用聚类分析的不足,提高检测效果。图 2 为具体的实现流程。下面我们详细阐述该方案的设计思想。

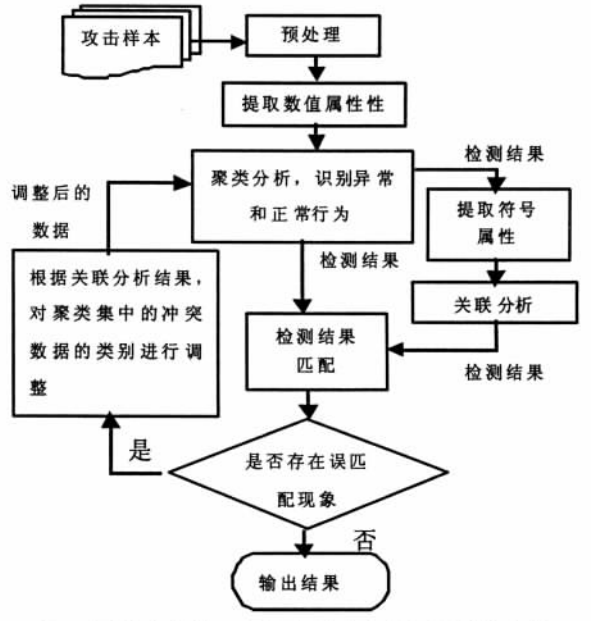

图 2　聚类分析结合关联分析方法的入侵检测流程

2.1　关联规则生成

首先我们将网络数据包预处理成适合数据挖掘的网络连接记录。利用聚类分析方法直接对预处理后的网络连接记录集进行聚类:提取网络连接记录中的数值属性,利用 K-Means 聚类算法初步将数据划分成正常聚类和异常聚类。

对这两个集合提取主要符号属性,包括:协议类型(protocol type),服务类型(service),TCP 连接标记(flag)和记录状态(state)等,使用关联分析 Apriori 算法[4]得出主要符号属性之间的关联特征。得出的关联规则如:smtp∩SF→normal [support, confidence],表示服务类型为 smtp 且 TCP 连接标记为 SF 时,记录状态为 normal 的概率是 confidence,这条规则发生的概率是 support。然后将支持度 support 和置信度 confidence 分别大于最小支持度和最小置信度的规则列为强关联规则,分别建立正常行为模型和入侵模型。

2.2　关联规则对聚类结果的重新划分

关联规则对聚类结果重新划分的流程如下:

(1) 利用聚类分析方法直接对预处理后的网络连接记录集进行聚类,首先提取网络连接记录中的数值属性,利用 K-Means 聚类算法初步将数据划分成正常聚类和异常聚类;

(2) 利用聚类得到的分类结果,在此基础上进行提取网络连接记录中的符号字段,并进行关联规则的分析;

(3) 通过关联规则的分析得到有关正常类和异常类的关联规则,运用这些规则检验原来聚类的结果,将正常聚类中特征规则与入侵模型匹配的连接记录以及异常聚类中特征规则与正常模型匹配的连接记录提取出来组成新的记录集;

(4) 对该记录集迭代(1)至(3)的步骤,直到用关联规则进行检查时得到的新的记录集为空;

(5) 输出正常类和异常类的结果。

重新划分聚类后使得异常聚类中的异常记录数增大而正常记录数大大减小,从而降低误检率。

3 实验结果

实验中我们所用的数据集是 KDDCup99 的 10%版本[5],我们从中随机选取了大约 49433 条记录。KDDCup99 的数据集中包含了 41 条属性,在本文中我们主要针对拒绝服务类型的攻击作为检测对象,这里我们参考文[6]中的结果,选取了如下数值属性:dst_host_count, dst_host_srv_count, same_srv_rate, dst_host_same_srv_rate, count, dst_host_same_src_port_rate, srv_count,使用 K-Means 算法进行聚类分析。选取 4 个关键符号属性值利用 Apriori 算法进行关联规则挖掘:protocol_type, service, flag, state。得到关联规则如表 1 所示。

表 1 生成关联规则表

关联规则	置信度/%	支持度/%
http ∧ REJ => normal	100.00	8.32
smtp ∧SF => normal	100.00	8.07
ftp ∧ SF => normal	100.00	4.15
domain_u ∧ SF => normal	100.00	1.62
finger ∧ SF => normal	100.00	0.72
other ∧ SF => normal	100.00	0.50
tcp ∧ SF => normal	100.00	72.03
telnet ∧ SF => normal	100.00	0.40
private ∧SO => neptune	100.00	0.55
http ∧ SF => normal	100.00	56.93
ecr_i ∧ SF => smurf	67.67	1.14
private ∧ SF => normal	37.56	0.15

入侵检测主要以检测率和误检测率为评估指标。在测试中使用聚类结合关联规则的技术同样对上面的网络连接记录进行检测,与单独采用聚类算法的检测结果进行比较分析,采用攻击检测率和误检率来衡量评价分析实验。通过实验得到单纯使用 K-Means 算法进行检测和使用聚类结合关联规则的检测方法结果,两者比较如表 2 所示。

表 2 检测结果比较

技术方法	检测率/%	误检率/%
单纯使用 K-Means 算法	85.03	3.01
聚类分析结合关联规则方法	85.03	0.12

可以看出使用聚类分析结合关联规则的方法将关联规则挖掘技术引入到聚类分析机制中,利用针对符号型属性的关联规则挖掘结果对聚类结果进行修正,从而有效减少由于在入侵检测单纯使用聚类分析所导致的误报,可以有效地检测率攻击的同时使得误检率大大降低。

4 结论

本文针对当前基于聚类分析方法的入侵检测技术的不足,采用聚类分析结合关联分析的方法应用到入侵检测中,并给出了该检测方法的流程,实现了该方法的模型。通过实验证明采用这种方法能够在保证检测率的前提下有效降低误检率。我们下一步工作是将该方案应用到其他类型攻击的检测中,比如非法访问及权限提升攻击等。

参考文献:

[1] Leonid Portnoy, Eleazar Eskin, Salvatore J Stolfo. Intrusiondetection with unlabeled data using clustering [C]// Philadelphia PA: Proceedings of ACM CSS workshop on data mining applied to security, 2001:5–8.

[2] W Lee, S J Stolfo, K W Mok, et al. Algorithms for mining system audit data[C]// New York:Proceedings of IEEE Symposium on Security and Privacy, 1999.

[3] 向继,高能,荆继武. 聚类算法在网络入侵检测中的应用[J]. 计算机工程,2003,29(9):48–50.

[4] [美] Jiawei Han, Micheline Kamber. 数据挖掘:概念与技术[M]. 北京:高等教育出版社,2001.

[5] KDD99Cupdataset[DB/OL]. [2008-09-16]. http://kdd.ics.uci.edu/ databases/kddcup99/kddcup99.html.1999.

[6] Wing W Y NG, Rocky K C Chang, Daniel S Young. Dimensionality reduction for denial of ServiceDetection problems using RBFNN output Sensiti-vity" [C]// Berlin: Proceedings of the second international conference on machine learning and cybernetics, 2003.

[责任编辑 林 锋]