# 基于孤立点挖掘的异常检测研究

李 睿,肖维民
(安徽工业大学 计算机学院,安徽 马鞍山 243002)

**摘 要**:随着网络技术的迅速发展,新类型的入侵行为层出不穷,人们迫切需要能检测出新类型入侵行为的技术。将数据挖掘与入侵检测相结合,能够增强入侵检测系统对海量数据的处理能力,使得入侵检测系统具有可扩展性和自学习能力,增强入侵检测系统的检测功能。从数据的观点来看,入侵检测本身是一个数据分析过程,在数量上远少于正常行为的入侵行为可看作孤立点。于是将数据挖掘中的孤立点挖掘技术作为一种网络安全检测手段,用来识别变种或未知入侵行为,对于改善入侵检测系统的性能有着重大的研究意义。文中着重通过对LOCL孤立点算法进行介绍,并提出改进算法,从而有效减少计算量,快速挖掘数据更新后的新孤立点,具有较高的实用价值。

**关键词**:入侵检测;异常检测;数据挖掘;孤立点挖掘;孤立点
**中图分类号**:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2009)06-0168-03

# Research on Anomaly Intrusion Detection Based on Outlier Mining

LI Rui, XIAO Wei-min
(School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

**Abstract**: With the fast development of the Internet, people urgently on technology with the ability of discovering new types of intrusion coming out endlessly. The combination of data mining and intrusion detection can make the intrusion detection system deal with a vast amount of data and have the ability of extension and self-study as well as enhance the detecting ability. According to the point of data, intrusion detection is a process of data analysis. The invasion which is obviously less than normal action in amount can be seen as the outlier. Therefore, outlier mining is a fundamental and important field in data mining, and it can be used to identify varieties or unknown invasions as one of tools for intrusion detection, which is of great value in promoting intrusion detection system. Focuses on introducing the LOCL algorithm of outlier mining, and to bring up the improved algorithm, which can reduce the computation effectively and mining the new outliers quickly when the dates updated. This thesis has high values.

**Key words**: intrusion detection; anomaly detection; data mining; outlier mining; outlier

## 0 引 言

保障网络系统的安全,仅仅依靠传统的被动防护是远远不够的。入侵检测(Intrusion Detection)作为一门新兴的安全技术,以其对网络系统的实时监测特性,逐渐发展成为保障网络系统安全的关键部件。入侵检测是从计算机网络的计算机系统中的若干关键点搜集信息并对其进行分析,从中发现网络或系统中是否有违反安全策略的行为和遭到袭击的迹象的一种机制。它采用预先主动的方式,对客户端和网络各层进行全面有效的自动检测,以发现和避免被保护系统可能遭受的入侵和破坏,它不仅检测来自外部的入侵行为,同时也检测内部用户的未授权(unauthorized)活动。

目前的商用IDS产品大多是由安全领域的专家根据经验总结已知入侵行为的入侵特征,然后利用某种特定结构表示入侵特征和检测策略,手工定制编码,最后对每个数据包模式匹配。这种机制的缺点:规则库是硬编码,漏报率高,无法检测变种或新型入侵等,大大限制了系统本身的可扩展性和自适应性。基于此,为了保障网络安全,众多实验室和研究机构针对识别新入侵类型做了大量研究工作,提出各种方法来检测变种或新型入侵行为,许多方法已得到了实现,有的还在进一步改进和完善之中。

利用孤立点挖掘技术应用于入侵检测系统的基本原理和方法尚未得到充分研究,因此本课题结合入侵检测的特点进行孤立点算法的研究,具有一定的理论价值和实际意义。

**收稿日期**:2008-09-18;**修回日期**:2008-12-11
**基金项目**:安徽省自然科学基金(2004kj062,2005kj070,2005kj071);安徽省教育厅青年教师资助计划(2004j9128)
**作者简介**:李 睿(1983-),女,安徽巢湖人,硕士研究生,研究方向为数据挖掘及数据库;肖维民,硕士,副教授,研究方向为网络技术及数据库。

## 1　基本概念

### 1.1　孤立点的定义

对于什么是"孤立点",目前有许多不同形式的定义,但却没有被人们普遍接受的统一方式。迄今为止,最有代表性的就是 Hawkins(1980)给出了孤立点的本质性的定义[1]:孤立点是在数据集中与众不同的数据,使人怀疑这些数据并非随机偏差,而是产生于完全不同的机制。孤立点挖掘的目的是发现数据集中明显不同于其他数据的数据对象。其基本思想是给定一个 $N$ 个数据对象或对象的集合,及预期的孤立点的数目 $n$,发现与剩余的数据相比是显著相异的、异常的或不一致的前 $n$ 个对象。如图 1[2] 中标记为 $V, W, X, Y, Z$ 的五个点相对于正常数据点有明显的不一致,类似这样的点就是孤立点。

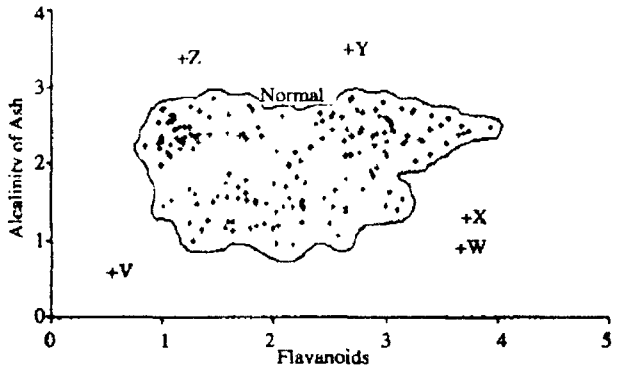



图 1　带有孤立点的数据分布图

孤立点挖掘问题可以被看作两个子问题[3]:

(1)在给定的数据集合中定义什么样的数据可以被认为不一致的;

(2)找到一个有效的方法来挖掘这样的孤立点。

### 1.2　异常检测技术综合分析

异常入侵检测是基于行为的检测技术,它根据使用者的行为或资源使用状况或用户的活动轮廓来判断是否入侵,而不依赖于具体行为是否出现来检测[4]。

异常检测从方法自身的原理讲具有检测系统中未知攻击的能力,由于新攻击方法总是不断出现,因此异常检测技术一直较受重视,产生了大量的异常检测技术,常用的技术有:统计分析方法、基于模式预测、神经网络研究、基于规则的方法,人工免疫法等。虽然人们已经提出了这么多的异常检测技术,异常检测还有很多缺陷需要投入更多的研究力量去克服。

## 2　孤立点检测算法

### 2.1　孤立点检测算法综合评述

目前,较为成熟的孤立点检测方法主要有:基于统计的方法、基于偏离的方法、基于聚类的方法、基于距离的方法以及基于密度的方法。这几种孤立点检测及相关算法的主要特性比较如表 1。

表 1　几种孤立点检测算法比较[5]

| 检测方法 | 典型算法 | 特点<br>(优缺点) | 效率<br>(时间复杂度) | 适用性 |
|---|---|---|---|---|
| 基于统计的方法 | 单样本多个离群检测算法(ESD) | 易于理解,只对单维数据集较为有效,需要知道数据集的先验知识 | 与数据集大小及采用的概率分布模型有关,一般为 $O(n\log n)$ | 不适合高维数据 |
| 基于偏离的方法 | 序列异常技术<br>OLAP 数据立方体技术 | 概念有缺陷,遗漏了不少孤立点 | 时间复杂度与数据集大小成线性关系 | 适用性不高 |
| | OLAP 数据立方体技术 | 搜索空间大,人工探测困难 | 不高 | 可适用于多维数据中 |
| 基于聚类的方法 | 聚类与孤立点检测算法结合 | 先聚类,后检测异常点 | 效率不高 | 适用于大规模的数据集 |
| 基于距离的方法 | 索引算法 | I/O代价较高,性能与索引结构有关 | 一般为 $O(kn^2)$ | 适用于低维空间的数据集 |
| | NL 算法 | 理论上可以处理任意维,减少了算法的 I/O 次数 | 一般为 $O(kn^2)$ | 适用于低维空间的数据集 |
| | 单元算法 | 在低维空间优于 NL 算法 | $O(ck+n)$ | 一般适用于较低维 |
| 基于密度的方法 | LOF 算法 | 可以识别局部异常 | $O(n\log n)$ | 不适用于高维空间 |
| | LSC 算法 | | $O(kn^2)$ | |

### 2.2　孤立点算法在入侵检测中的应用分析

孤立点算法能够应用在入侵检测中是建立在两个基本假设之上:

(1) 正常数据的数量远远大于入侵数据量;

(2) 入侵数据在某些属性的取值偏离正常取值范围。

在入侵检测中利用聚类算法挖掘入侵行为,孤立点不再是算法的副产品,而是作为主要的分析对象。同时由于算法的本质和网络数据本身就具有较大的噪声,所以聚类前的数据预处理将是至关重要的,否则会严重影响性能。

文中将基于密度的算法作为研究重点,加以改进。

## 3　LOCL 孤立点算法及改进算法

### 3.1　动态 LOCI 算法简介

Spiros Papadimitriou 等人提出的基于密度的局部关联识别算法(Local Correlation Integral, LOCI)[6]。LOCI 算法只适应于静态环境。如果数据集中的某个对象发生变化,则需要重新计算集合中所有对象的 LCIF 值,由于计算 LCIF 的时间复杂度太高,将 LOCI 算法直接应用于动态的数据库环境是不现实的。通过对 LCIF 特性的研究可知,每个数据对象的 LCIF 值,仅与该对象所处的局部环境有关,数据更新一般也仅对某些相关对象的异常程度造成影响。因此,针对动态数据环境,特别是网络入侵检测的实际需要,提出一

种动态的局部异常增量更新挖掘算法(Dynamic Local Correlation Integral, DLOCI),在原有计算结果的基础上,只对受影响的部分数据重新计算,可以避免重新计算所有对象的 LCIF 值,从而在动态环境下大大提高孤立点的挖掘速度。

### 3.2 基本思想

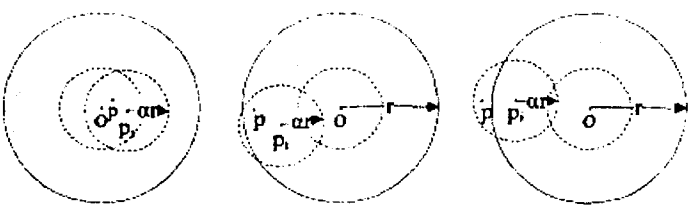
当添加或删除一个对象,某些相关对象的 LCIF 值之所以会发生变化,是因为添加或删除动作使数据集中某些对象的邻域发生了变化。通过采用动态的挖掘算法,使之能够在数据集更新的情况下,用较小的计算代价,达到和对所有数据对象重新计算同样的效果。

假设 $D$ 为一数据集合,$o$ 为其中任意对象,添加或删除对象 $p$,有下列几种情况需要分别进行处理[7]:

1) 如果新添加或删除对象 $p$ 在对象 $o$ 的 $\alpha\gamma$ - 邻域中,那么对象 $o$ 的 $n(o, \gamma)$ 和 $n(o, \alpha\gamma)$ 值肯定发生变化。除此之外,当对象 $o$ 的 $r$ - 邻域中某个对象 $pi$ 的 $\alpha\gamma$ - 邻域也包含对象 $p$ 时,那么对象 $pi$ 的 $n(pi, \alpha\gamma)$ 值同样发生变化(见图 2(a))。

2) 如果新添加或删除对象 $p$ 在对象 $o$ 的 $\gamma$ - 邻域中,但不在 $\alpha$ 邻域中,那么仅仅对象 $o$ 的 $n(o, r)$ 值和对象 $o$ 的 $\gamma$ - 邻域中 $\alpha\gamma$ - 邻域包含对象 $p$ 的对象 $pi$ 的 $n(pi, \alpha\gamma)$ 值同样发生变化(见图 2(b))。

3) 如果新添加或删除对象 $p$ 不在对象 $o$ 的 $\gamma$ - 邻域中,仅当对象 $o$ 的 $\gamma$ - 邻域中某个对象 $pi$ 的 $\alpha\gamma$ - 邻域包含对象 $p$ 时,对象 $pi$ 的 $n(pi, \alpha\gamma)$ 值才发生变化(见图 2(c))。




(a) p 在对象 o 的 α γ-邻域中　(b) p 在对象 o 的 γ-邻域中,但不在 α 邻域中　(c) p 不在对象 o 的 γ-邻域中

图 2 数据更新时,点 $o$ 邻域变化示意图

如果从对象 $p$ 的角度来讨论受影响集合,如果某个对象 $o$ 存在于对象 $p$ 的 $\alpha\gamma$ - 邻域中,那么对象 $p$ 同样也在对象 $o$ 的 $\alpha\gamma$ - 邻域中,满足图 3(a) 的条件,那么对对象 $p$ 进行的更新动作可能改变对象 $o$ 的 LCIF 值,则可推理出对象 $p$ 的 $\alpha\gamma$ - 邻域中所有的对象都应包含在受影响集合。

同理,对象 $p$ 的 $r$ - 邻域中所有的对象满足图 3(b) 的条件,于是它们也都包含在受影响集合中。而如果某个对象 $o$ 在对象 $p$ 的 $(1+\alpha)\gamma$ - 邻域中,当且仅当至少存在一个对象同时属于对象 $o$ 的 $r$ - 邻域和对象 $p$ 的 $\alpha\gamma$ - 邻域时,才满足图 3(c) 的条件,即如果[8]:

$$N(o, r) \cap N(p, ar) \neq \emptyset$$

那么对象 $o$ 才包含在受影响集合中。

### 3.3 算法实现

数据集 $D$ 中数据对象 $o$ 的数据结构为:

```
Struct NodeType
{Data type NodeID; //每个对象的唯一标识符
int count_num; //统计邻域中的对象总数
int sample_num; //取样邻域中的对象总数
Bool Lcif // Lcif 值,用来标识是否为孤立点
Datatype * cNode; //该链表储存统计邻域中所有对象的 NodeID 号
Datatype * sNode; //该链表储存取样邻域中所有对象的 NodeID 号
}
```

动态环境下 DLOCI 算法的运行时间远远小于 LOCI 算法的运行时间,并且数据量越大,效果越明显。

## 4 结束语

基于密度的孤立观点比其他观点更贴近 Hawkins 的孤立定义,因此能够检测出局部孤立数据。局部孤立点观点摈弃了以前所有孤立点定义中非此即彼的绝对孤立观点,根加符合现实生活中的含义。

考虑孤立点挖掘在网络环境中的实际应用,文中在详细分析基于密度的局部关联识别算法后,提出一种改进算法——DLOCI 算法。该算法利用原有挖掘的中间结果,有效减少计算量,快速挖掘数据更新后的新孤立点,具有很高的实用价值。

**参考文献:**

[1] Hawkins D. Identification of Outliers[J]. London: Chapman and Hall, 1980.

[2] Hodge V J, Austin J. A survey of Outlier Detection Methodologies[J]. Artificial Intelligence Review, 2004, 22(2): 83 - 171.

[3] Han J W, Kamber M. Data mining: concepts and techniques[M]. New York: Morgan Kaufmann Publishers, 2001.

[4] 唐正军. 网络入侵检测系统的设计与实现[M]. 北京: 电子工业出版社, 2002.

[5] 孙 云, 李舟军, 陈火旺. 孤立点检测算法及其在数据流挖掘中的可用性[J]. 计算机科学, 2007, 34(10): 7 - 10.

[6] Papadimitriou S, Kitagawa H, Gibbons P B, et al. LOCI: Fast Outlier Detection Using the Local correlation Integral [M]. USA: [s. n.], 2007.

[7] 蒋良孝, 蔡之华. 异常挖掘方法研究[J]. 计算机工程与应用, 2003, 39(19): 23 - 27.

[8] White G M. Nuggets and data mining[M]. USA: Data Mining Technologies Ins, 1998.