用于入侵检测及取证的冗余数据删减技术研究

钱 勤[1,2] 张 琎[2,3] 张 坤[1] 伏 晓[2,3] 茅 兵[3]
(江苏省高级人民法院技术处 南京 210024)[1] (南京大学软件学院 南京 210093)[2]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[3]

摘 要 近年来计算机犯罪逐年增多,并已成为影响国家政治、经济、文化等各个领域正常发展的重要因素之一。入侵检测技术与入侵取证技术对于打击计算机犯罪、追踪入侵、修补安全漏洞、完善计算机网络安全体系具有重要意义。但是,随着网络的普及以及计算机存储能力的提升,入侵检测及取证技术目前需要分析的往往是GB乃至TB级的海量数据,而且有用信息往往湮没在大量由正常系统行为触发的冗余事件之中。这无疑给分析过程带来了巨大的挑战,也使分析结果的准确性不高。因此,如何设计出一种自动冗余数据删减技术来提高入侵检测及取证方法的准确率及效率,是当前入侵检测和取证领域的关键问题之一。文中即对这方面已有的研究工作进行了综述,首先介绍了冗余数据删减技术的发展历程及其在医学数据分析等传统领域的应用,然后重点介绍了针对入侵检测和入侵取证的现有各种冗余数据删减方法,最后通过对当前冗余数据删除技术的比较,指出了该领域当前存在的问题及未来的研究方向。

关键词 入侵检测,入侵取证,冗余数据删减
中图法分类号 TP393.08 **文献标识码** A

Technical Study of Reducing Redundant Data for Intrusion Detection and Intrusion Forensics

QIAN Qin[1,2] ZHANG Jian[2,3] ZHANG Kun[1] FU Xiao[2,3] MAO Bing[3]
(Technology Department, Jiangsu Provincial Higher People's Court, Nanjing 210024, China)[1]
(School of Software, Nanjing University, Nanjing 210093, China)[2]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[3]

Abstract For the past few years, the amount of computer crime has been increasing year by year, and it is threatening various aspects of human society such as national politics, economy, and culture, etc. In modern society, the research on intrusion forensics and intrusion detection plays a significant role for fighting against computer crime, tracing intrusion, patching vulnerability and improving security system of computer network. However, with the popularity of Internet and the improving capacity of computers' storage, we often need to handle mass data about GB size, even up to TB size for intrusion forensics and intrusion detection. It inevitably makes much useful information submerge in redundant events, which brings about a huge challenge and low accuracy of analysis result. So it will be a topmost breakthrough to design a kind of technology for reducing redundant data and improving its accuracy and efficiency. This paper summarized several methods on intrusion forensics and intrusion detection. Firstly, this paper discoursed the development course of redundancy-reducing techniques and the application in traditional field such as medical domain. Then it systematically introduced all kinds of redundancy-reducing methods in intrusion forensics and intrusion detection. Finally, it figured out the existing problems and research direction in the future. It also gave some conclusions through the comparison on current situation of redundant data reducing techniques.

Keywords Intrusion detection, Intrusion forensics, Redundant data reduction

1 引言

随着计算机、网络的普及以及社会信息化程度的加深,利用网络及计算机漏洞进行黑客入侵的案例也逐渐增多。据报道,在2011年4月份发生的"索尼被黑"事件导致黑客从索尼在线PlayStation网络中窃取了7700万客户的信息,包括信用卡账号。这一黑客攻击事件导致索尼被迫关闭了该服务。索尼在5月份表示,攻击导致其损失了1.7亿美元。由此可见,计算机网络发展越迅猛,计算机犯罪的危害也就越大,而且计算机犯罪不仅会造成巨大的经济损失,也可能危害国家的公共安全。

为了遏制计算机犯罪日益增长的趋势,入侵检测技术和

本文受国家自然科学基金项目(61100198/F0207),国家973项目(2010CB327903)资助。
钱 勤(1967-),女,硕士生,主要研究方向为信息安全,E-mail:qianqin@software.nju.edu.cn;张 琎(1990-),男,硕士生,主要研究方向为信息安全;张 坤(1975-),男,主要研究方向为信息安全;伏 晓(1979-),女,博士,讲师,主要研究方向为计算机取证、网络安全,E-mail:fx@software.nju.edu.cn(通信作者);茅 兵(1967-),男,博士,教授,博士生导师,主要研究方向为系统安全、分布式系统。

入侵取证技术应运而生。入侵检测技术的目的是迅速地发现入侵行为，为防范入侵提供良好的措施。而入侵取证技术则能够获取入侵轨迹，为司法部门将犯罪人员绳之以法提供了充分、可靠和强有力的电子证据。

对于入侵检测和入侵取证而言，其关键在于通过分析各类候选数据找出其中的入侵行为。但是，随着网络的普及以及计算机存储能力的提升，入侵检测及取证技术目前需要分析的往往是GB乃至TB级的海量数据，而且有用信息通常湮没在大量由正常系统行为触发的冗余事件之中。这无疑给分析过程带来了巨大的挑战：这些冗余、无关的日志数据不仅会导致分析效率降低，而且会导致分析算法的误报率增加、分析结果可信度降低。另外，巨大的数据集也使实时入侵检测和实时取证分析难以实现。因此，设计出一种实用的针对入侵检测及取证的自动冗余数据删减技术显得尤为重要。本文即对这方面已有的研究工作进行了综述。

本文第2节对冗余数据删减技术及其在传统领域的应用进行了简单的介绍；第3节和第4节分别讨论用于入侵检测领域和取证领域的各类现有冗余数据删减方法；第5节针对现有的冗余数据删减技术进行了分析及比较；最后，对该领域的研究现状进行了总结并展望了未来研究方向。

2 冗余数据删减技术简介

在许多科学领域中，研究者都需要对包含大量冗余或无关数据的大规模数据集进行分析，因此如何对冗余数据进行有效删减是一个具有普遍性的重要问题，也是研究者长期关注的问题。例如在传统的医学数据处理领域，如何处理来自不同位置的数据是一个充满挑战的问题：此时，需要处理的数据规模之大、复杂度之高常常让分析的难度变得很大。但是，一个大数据集中往往包含着丰富的有用数据。因此，为提高处理效率，获得高质量的处理结果，研究者需要采用一些策略提前对大数据集进行删减。

目前较有代表性的冗余数据删减方法主要有[1]：基于结构化索引[2]的技术、基于频率(Frequencies)的技术[3]、基于同现(Co-occurrence)的技术[4]、基于图论(graph-theoretic)[5-7]的技术等。

基于结构化索引[2]技术的主要思想就是为数据建立一系列的结构化代码，这些代码作为一个索引，允许研究者快速访问与分析相关的数据。这种方法使用了结构化或者半结构化的“焦点群体”指南，这些指南指的是那些在同一个数据集中而出现在不同文件中毫无关联的问题。例如，在一次临床的抗HIV的药物试验中，有一个面试过程，各个试验的参与者会被问及一些问题。然而，这些问题会被分成不同的领域：人口统计信息(Demographic Information)、审讯经历(Trial Experience)、药理研究及其可接受性(Acceptability and Knowledge of the Study Drug)、性交以及避孕措施(Sexual and Contraceptive Practices)。针对每个领域，每一个问题和参与者的回答被赋予一个代码名。这样，一些相关问题的数据就很容易从整个数据集中提取出来。基于结构化索引技术具有两个优势：(1)使大量数据的分析变得更为高效；(2)对建立基于数据处理的主题代码而言，把相关的数据整合到一起也是相当有用的。但是，该项技术不能用于数据查询，因此该技术的性能会大打折扣。另外，这些代码在本质上既非基于数据处理的代码也非基于主题的代码。

基于频率方法[3]的思想是计算短语或者词语在一个数据集中出现的频率。该方法对于鉴别大规模数据集中反复出现的冗余或有用数据很有用。数据集中简单的关键字搜索或者文字计数能让同一分析中不同子集之间的文字进行快速比较。基于频率技术具有以下特征：(1)关键字的量化过程可以决定什么样的关键字包括在内，什么样的关键词排除在外，从而形成一个基于主题的代码集合；(2)这个代码集合对于一个数据集的任何位置来说都是共享的，换而言之，各个位置可以利用同一个代码集合进行同样的有效的分析。但是，这种方法也存在一定程度上的缺陷：1)词语计数或者是关键字检索往往需要研究者知道应该去检索什么样的词语，具有一定的主观性，而语言的准确性也是一个比较棘手的问题；2)不能确定一个特定的关键字能否作为一个抽象的标记，尤其是缺少上下文帮助的时候。

基于代码同现[4]的思想是从唯一的数据对象实体中抽取一个独立的片段，其中片段中含有几段代码，即这个片段的代表，借此观察代码同时出现的次数。例如，一段文本中记录了从面试记录文本中提取的抗HIV方法，其中可能提到“family”，“faith”和“doctors”。这些参照词语代表了不同的意思，需要不同的代码。这段文本被赋予3个不同的代码，被理解为在这个文本片段中同时出现。基于代码同现技术具有以下几个长处：(1) 代码同现往往会提供一些有用的信息，让我们知道一个数据集中的主题域或者概念是怎么分布的；(2)通过检测代码同现(代码项，主题等)之间的关联关系来识别数据集中的模式；(3)与基于频率技术一样，观察两个代码在单独的数据文件中或者整个数据集中同时出现的次数是比较有益的。但是，怎样去搜集代码同现的报告是一个值得考虑的问题。

基于图论[5-7]的技术又称语义网络分析，可用于鉴别文本体中复杂的语义关系。最常见的基于图论的删减技术包括层级聚类(Hierarchical cluster)分析技术和多维尺度(Multidimensional scaling-MDS)分析技术。层级聚类分析技术[8]是一种采用凝聚的思想而得出的方法。层级聚类分析把数据分组，相同组内的成员相似度高，组间成员的相似度低。在聚类分析之前，分析者通常用相似矩阵来显示数据。这个矩阵既可以是二进制的(用0或者1表示)，也可以用一组值来表示相似程度。层级聚类分析技术具有以下几个优点：(1)使用相似矩阵作为输入参数，研究者能够利用聚类分析技术看到大数据集中的数据模型；(2)如果在相似矩阵中选取恰当的代码，就能形成一个较为精简的上下文；(3)聚类分析也能用于基于文本的内容分析，使用健壮的聚类分析技术能够产生三维数据的输出布局。但是，就其本身性质而言，聚类分析方法虽然较为简单，但是得到目标群体的综合特征却并不容易。多维尺度分析技术是另一种常见语义网分析方法。Schiffman[9]等人将MDS定义为一个强大的数学过程，通过在图表空间中展示物体之间的相似之处使数据系统化。对于N对数据之间可视化到的相似点，MDS让研究者们用空间中尽可能少的维度展示出所有项。MDS技术具有以下几个长处：(1)可以使研究者在空间尽可能少的维度中找到数据项；(2)MDS的主要技术核心是采用数据点之间的相对位置形成的图来说明问题，点与点之间距离近代表相似度高，能够进一步

地分析与解读源文本数据；(3)配比度(介于0—1)是MDS图中一个重要的衡量标准，它用于评估原始输入距离与结果图表之间的一个匹配程度，越接近0代表MDS图表的配比度越高。但是，当维度越来越多时，则必然会降低匹配等级，因为两个以上的维度在纸上就很难绘画出来了，并且会越来越难理解。如Borgatti[10]所言，一旦超过了4个维度或者更多，MDS便不能制作出人类可以接受的复杂数据。

上文这些数据删减方法主要应用于文本数据源的删减，例如医学数据分析领域，其分析的数据来源主要是临床的医学报告、医学调查报告和医学面试记录等文本数据，所以上述方法比较适合。对于入侵检测和取证领域来说，其数据来源更加复杂，包括诸如IDS报警之类的各种日志文件、程序、脚本、内存镜像、磁盘备份、交换区文件、临时文件、硬盘未分配的空间，系统缓冲区和打印机及其他设备的内存等等，均是其重要数据来源。因此，本节介绍的方法无法直接使用。但是上述方法的建立索引、计算频率、聚类以及利用空间维度的数据删减的思想仍值得入侵检测及取证研究者借鉴。

3 用于入侵检测的冗余数据删减技术

随着网络的迅猛发展，越来越多的系统遭受入侵攻击的威胁。作为一种有效的安全防护措施，入侵检测系统(IDS)在近几年已经得到了广泛应用。但高误报率是IDS至今未解决的难题之一。据统计，IDS每周的报警中99%都是误报[18]。为解决IDS误报问题，研究者提出了针对IDS报警数据的冗余数据自动删减技术，即误报删减(alert reduction)技术。此类技术主要包括以下几种方法：基于分类技术的方法、基于聚类技术的方法、基于孤立点检测技术的方法、基于统计学的方法。下文将分别对其加以介绍。

3.1 基于分类技术的数据删减方法

分类技术[11]是数据挖掘领域的基本技术，其过程大致分为两步：第一步，建立一个模型，描述预定的数据类集或概念集。该模型可通过分析属性描述的数据库元组来构造。第二步，使用模型进行分类并且评估模型(分类法)的预测准确率。如果认为模型的准确率可以接受，就可以用它对类标号未知的数据元组或对象进行分类。

若利用分类技术对IDS进行误报删减，一般来讲，分析者需要首先用他们的知识去区分真实报警和误报，从而获得一个准确的训练数据集来生成分类器。然后基于这样的分类器，就可以删除无用报警，报告安全事件，对入侵事件进行调查或者去鉴别网络和配置的问题。

目前，有不少研究者利用分类技术对IDS进行误报删减。例如，Pietraszek[12]提出的ALAC(Adaptive Learner for Alert Classification)是基于分类置信度，通过把报警分成真实报警和误报来降低误报率。Law和Kwok[13]通过运用KNN(k-nearest-neighbor)分类器，提出了一种新型的方法来降低误报数量。他们首先建立一个模型描述报警序列，然后将偏离模型的报警视为异常报警。Alharbt和Imai[14]使用连续和非连续的序列模式检测偏离正常警报流模型的异常。Davenport等[15]则提出了一种分析方法来控制误报频率，通过使用基于Neyman-Pearson定律的支持向量分类器来最小化误报的缺失率。

分类技术主要包括以下几种算法：(1)基于决策树分类；(2)基于神经网络分类；(3)基于关联规则挖掘分类等。就决策树分类而言，决策树非常擅长处理非数值型数据，从决策树中可以轻松地提取到分类规则。其主要优点是，算法本身描述简单，分类的速度较快，适用于大规模数据集的数据处理。相反，其不足之处是，其算法偏向于选择不是最优的属性，逻辑表达能力较差。就基于神经网络分类而言，其优点噪能力强，对没有经过校准的数据有着较好的预测分析能力。其缺点就是，神经网络的表现形式让人很难理解，学习时间长。就关联规则挖掘概念的分类而言，其优点是其精确度比C4.5决策树高，其缺点是比较依赖于置信度和支持度。

这几种方法中，目前在IDS误报删减领域应用较多的是基于关联规则挖掘的分类方法，上节介绍的现有基于分类的IDS误报删减方法即ALAC和KNN属于此类。基于神经网络分类方法使用较少是因为其表现形式比较复杂，难以理解，仅适用于时间允许的应用场合；基于决策树分类方法使用较少是因为该方法趋向于选择较多而非较优的属性字段，给分析者的分析工作带来了一定的难度。所以未来研究者可针对现有几种方法的缺点进行一定程度的改进，以提高数据分类的效率。

3.2 基于聚类技术的数据删减方法

聚类技术[11]也是数据挖掘的基本技术，其主要思想是将物理或抽象对象的集合分组成为由类似的对象组成的多个类。在同一类中的对象之间具有较高的相似度，而不同类中的对象差别较大，其中相似度是根据描述对象的属性值来计算。

目前，也有一些研究者利用聚类技术对报警集合中的报警进行误报删减。例如：在文献[16-18]中，Klaus Julisch提出了一种基于概念聚类的模型。这个模型通过在每一个报警属性上建立泛化的数据结构对报警中数据对象进行聚类，从而支持报警原因分析。Manganaris等[19]提出入侵检测系统中的误报可以使用聚类的方法来进行处理，并借以识别及解决触发警报的根源。这个方法主要使用频繁项集和关联规则进行数据建模，从而达到报警聚类的目的。Joshua Ojo Nehinbe[20]提出，决定攻击模式的最简单的办法就是把最原始的数据集分成两类，这两类会给出由数据集属性值算出的最小平均信息量。这个过程被称为分离聚类，即首先把数据分成两类，每一个类中至少含有一个属性来区别与另外一个类的不同点，两类互相独立。重要的是，组成一个簇的警报中的各个数据对象非常相似，而与另外一个簇中的数据对象相异。

聚类技术已经被广泛地研究了许多年，聚类算法很多，较有代表性的主要算法[21]包括：(1)基于划分的K-means算法；(2)基于层次的BIRCH算法；(3)基于网格的STING算法，等。就基于划分的K-means算法而言，算法只适用于聚类均值有意义的情况，不适用于发现非凸型的聚类，同时对噪声和异常数据过于敏感。就基于层次的BIRCH算法而言，它的主要数据结构是聚类特征树，聚类特征树是根据不断插入的对象而动态建立的，该算法适用于大型数据库构造完成聚类特征树。算法表现出线性可扩展性，但只适用于呈凸形或球形的聚类。就基于网格的STING算法而言，由于采用多分辨率的方法进行聚类分析，所以聚类的质量取决于网格结构的最低层粒度。若粒度较细，那么处理的代价就会增加；但若粒度较粗，则会降低聚类的质量。

这几种方法中，目前在IDS误报删减领域应用较多的是基于划分的方法，上节介绍的K-means分类的IDS误报删减方法即属此类。BIRCH算法由于只适用于呈凸形或球形的聚类，因此只有遇到特定类型时才能使用；STING算法比较依赖于网格的底层结构，故该种方法使用较少。未来研究者可通过对各种算法的分析比较，根据各种聚类分析方法的优缺点和适用的领域，选择出适合特定问题的聚类方法。对现有算法进行改进和融合，也是未来研究者值得研究的方向。

3.3 基于孤立点检测技术的数据删减方法

在数据分析过程中经常发现一些数据对象，它们不符合数据的一般模型，这样的数据对象被分析者称为“孤立点”。孤立点检测技术是一种新型的数据挖掘技术，近年来得到了越来越多的关注。这种技术能够识别大规模数据集中的异常数据。在网络安全领域，有人已经成功地利用孤立点检测技术实现了IDS[22,23]。实际上，该技术也可以用来进行误报删减，因为误报占据了IDS报警的绝大部分，因此与误报相比，真实的报警完全可以被视为“孤立点”。

目前，已有一些研究者进行了此项工作。例如：在文献[24]中，Xiao使用了一种基于频繁模式的孤立点检测技术，此方法的基本思想是将那些频繁出现的报警属性值集视为“普遍特征”，然后根据报警包含这类特征的多少来识别及过滤误报。在文献[25]中，Syarif Iwan设计了一个误报删减系统，它能够识别真实报警，有效地降低误报，而且还能分析出误报产生的根本原因。这种方法是一种基于非监督的数据挖掘技术，即孤立点检测技术。这个系统的第一个功能是检测出代表真实攻击或者入侵的孤立点，第二个功能是分析误报发生的根本原因。值得一提的是，这些方法都是基于无标记数据并且在最大程度上降低了人工干预。这个系统通过Apriori算法建立频繁项集，从而产生关联规则来展示出所有的误报。

孤立点检测技术的典型算法包括：(1)基于统计的方法；(2)基于密度的方法；(3)基于距离的方法；(4)基于聚类的方法；(5)基于人工神经网络的算法；(6)基于频繁项集的算法。基于统计的方法视偏离分布模型的点为孤立点，对识别出的结果比较好解释。但是，这种方法只适用于单变量的数据模型，并且需要较多的先验知识。基于密度的方法提出了孤立强度的概念，量化了异常程度，避免了数据集的疏密程度不一对挖掘结果的影响。但是，对孤立的稀疏和异常的意义难以理解。基于距离的方法不需要知道数据的分布情况，算法的概念直观。同时，I/O较大，效率较低，运行效率的复杂度呈指数关系变化，不适合多维数据集。基于聚类的方法，聚类和孤立点监测一次完成，在数据集大小上的伸缩性比较好。当然，这个方法却存在一定的缺陷：计算量大，计算结果受指定参数k值的影响较大，如果修改k值，则需要重新构造数学模型和重新计算。基于人工神经网络的算法无论数据集的规模是大还是小，其检测结果都令人满意。不过，当孤立点中包含放射状的孤立点时，性能就会明显下降。基于频繁项集的方法基于无标记数据并且在最大程度上降低了人工干预，无需任何背景知识。与现有方法相比，它的优点在于能够实时过滤误报，而且能通过对新报警的不断学习来自动适应新的误报类型。但是，其准确度以及自动化程度还有待进一步提高。

这几种方法中，目前在IDS误报删减领域应用较多的是基于频繁项集的方法，上节介绍的现有基于分类的IDS误报删减方法即属此类。第(1)种方法使用较少是因为方法本身需要较多的先验知识，第(2)种方法使用较少是因为孤立点之间的稀疏程度以及异常的意义让人较难理解，第(3)种方法使用较少是因为不适用于多维数据集，第(4)种方法使用较少是因为计算结果受指定参数k值的影响较大，第(5)种方法使用较少是因为当孤立点呈放射状时效果不佳。因此，未来研究者应该尽可能地利用各种方法的优点，避免缺点，设计出高效的算法。

3.4 基于统计学的数据删减方法

除上述方法外，也有研究者提出了基于统计学的数据删减方法，比较有代表性的是朴素贝叶斯Naive Bayesian(NB)方法。该方法的主要思想是对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。这个方法对所有的警报设置优先级，对不重要的警报设置的优先级比较低，相反，关键的警报设置的优先级比较高。

Axelsson[26]写了一篇关于入侵检测的著名文章，这篇文章采用了Bayesian条件概率指出入侵检测中基本率谬误(base-rate fallacy)的含意。他研究发现，即使是准确测试中的确实结果，也未必就能说明高概率的假定变量就是对的。在入侵检测领域，这个发现意味着，一个模型即使能够准确识别恶意事件，也会引发很多无用的报警，因为在输入流中攻击的先验概率通常来说比较低。虽然Axelsson的文章与我们所要研究的问题并非太相近，但是有一点是明确的，用统计学的思想去解决误报问题是可行的。目前，已有研究者运用Bayesian统计学的思想去建造基于异常情况的IDS模型[27,28]。

Bayesian方法具有以下几个优点：(1)描述了变量之间的因果关系，概率化使得Bayesian的学习允许样本的不完整和噪声数据的存在；(2)能挖掘出知识的隐含性；(3)具有良好的可理解性和逻辑性；(4)结合先验知识，便于进行预测分析。当然，Bayesian在一定程度上存在不足：(1)Bayesian基于先验知识的使用，如果先验知识不正确或存在误差，那么最后导致的结论是难以想象的；(2)需要大量的数据处理以及计算，故其空间与时间消耗也是比较大的。

4 用于入侵取证的冗余数据删减技术

随着黑客的攻击水平不断提高，使用入侵检测工具并不能从根本上保证网络不受攻击。要从根本上解决入侵问题，必须依靠法律的力量对黑客行为予以约束。因此，打击黑客行为的关键问题就是如何对入侵者的行为进行取证分析。但随着计算机网络的发展及存储能力的提高，入侵取证时获得的候选证据往往是海量数据，其中重要的证据隐藏在大量的无关或冗余数据当中。这给入侵取证及调查带来了不小的困难。因此，为提高取证效率、增强取证结果的准确性，在入侵取证领域同样需要用到冗余数据删减技术。

目前关于入侵取证冗余数据删减的研究还很少，人工智能技术(artificial intelligent techniques)[29]是用于解决这类问题的主要方法，例如：基于模糊决策树的方法[30]、基于模糊专家系统的方法[31]、人工神经网络(Artificial Neural Networks-ANNs)方法[36,37]和支持向量机(Support Vector Machines-

SVMs)方法[39,40]等。下文将分别对其加以介绍。

4.1 基于模糊决策树的数据删减方法

模糊决策树是传统决策树的一种推广,它将模糊理论应用于训练与匹配过程,结合了决策树的可理解性和模糊集合的表示能力,用来处理模糊性和不确定信息,使决策树具有更好的健壮性,提高了决策树的可理解性,并使决策树归纳算法的扩展能力增强。

Liu,Zai-Qiang[30]等人提出了一种用于网络入侵取证分析的模糊决策树推理方法,以协助网络取证人员在网络环境下对计算机犯罪事件进行取证分析及数据删减。决策树因ID3程序被Quinlan[32]推广开来,ID3是基于概念学习系统的算法。ID3的主要工作机制是通过搜索训练实例中的属性并且从一个属性集合中提取属性。这个算法采用了贪心搜索去选择最好的属性。虽然决策树技术容易理解、比较有效并且能够处理大规模数据,但是其方差还是较大。此时,引入模糊逻辑可以提升性能。

基于模糊决策树技术的优点是:(1)决策树易于理解和实现,在学习过程中不需要使用者了解很多的背景知识,很容易理解决策树所表达的意义;(2)对于决策树,数据的准备往往是简单或者是不必要的,而且能够同时处理数据型和常规型属性,在相对短的时间内能够对大型数据源得到良好的处理结果;(3)易于通过静态测试来对模型进行评测,可以测定模型可信度;如果给定一个观察的模型,那么根据所产生的决策树很容易推出相应的逻辑表达式。其缺点是:1)对连续性的字段比较难预测;2)对有时间顺序的数据,需要很多预处理的工作;3)当类别太多时,错误可能会增加的比较快;(4)一般的算法分类的时候,只是根据一个字段来分类。

对于入侵取证数据删减而言,基于模糊决策树方法的优点在于不需要任何的背景知识,表达形式易于理解,能够同时处理数据型和常规型属性;缺点在于对连续性的字段比较难预测,遇到有时间顺序的数据需要较多预处理工作,删减时可利用的字段过少。

4.2 基于模糊专家系统的数据删减方法

模糊逻辑是一种强大的技术,它旨在解决那些包含不准确、不确定信息的人类推理和决策过程。就模糊成员而言,运用模糊逻辑使得我们能够量化一个信息集合成不同的参数。Zadeh[33]介绍到,模糊逻辑的基础来源于模糊集合理论。在过去的几十年里,该项技术已经成功地应用于金融、交通管制、汽车速度控制以及核反应乃至地震监测当中。模糊专家系统提供了模式分类函数,其核心在于使用了基于Takagi-Sugeno(TS)模糊模型[34,35]的知识结构。

Kim JS[31]提出了一种基于模糊逻辑的专家系统。这个系统能够分析特定网络环境中的计算犯罪,也使得数字证据变得更加自动化。除此之外,Kim JS为取证还提出了数字证据的抽象层概念。这个系统为取证专家们提供了可分析的信息,在一定程度上降低了时间消耗以及取证分析的成本。同时,增加调查者的签名字段也大大提升了数字证据的精确度和完整性。但是,就实时取证以及添加更多一般化的模糊规则而言,它值得进一步完善。

对于入侵取证数据删减而言,基于模糊专家系统方法的优点在于能够自动地分析网络环境中的计算机犯罪从而提取出相应的数字证据,其他证据则属于无用的范围,可以被删减;缺点在于数据删减的实时性不高,应用于删减的模糊规则还不尽完善。

4.3 基于其他人工智能技术的数据删减方法

人工神经网络(ANNs)是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型。它是由大量处理单元互联组成的非线性、自适应信息处理系统,能够将输入转换成要求的输出[36,37]。转换的结果取决于单元的特征以及单元之间的连接权重。一个神经网络能够进行信息分析,并且能提供与经过验证的数据之间的概率评估。通过对需处理问题的输入输出进行一段时间的学习及校准,神经网络可能获得相应的经验。神经网络根据一定的规则集衡量一些性能指标的重要性,例如:输入特征、校准花费的时间、测试时间和分类准确率等等。所以,该方法就是特征分级方法的一个扩展[38]。一旦将输入特征的重要性分级了,ANNs就只会用包含那些重要特征的数据集去校准或者测试。

支持向量机(Support Vector Machine)作为一种可校准的机器学习方法,依靠小样本学习后的模型参数进行导航星提取,可以得到分布均匀且恒星数量大为减少的导航星表。SVMs通过确定支持向量集合来分类数据,其中支持向量集合就是概述特征空间中超平面的校准输入集的成员[39,40]。SVMs提供了一般机制,通过使用一个核函数将数据与超平面保持一致。用户在SVM校准阶段会提供一个函数给SVMs。与人工神经网络相似,一旦将输入特征的重要性分级了,SVMs就只会用包含那些重要特征的数据集去校准或者测试。

目前已有研究者将上述两种人工智能方法应用于取证分析领域,以便减少人工干预来自动化整个取证过程,删减需要处理的数据量。例如,Mukkamala[29]主要致力于SVMs人工智能技术在网络入侵取证分析中识别的重要特征,同时也利用ANNs进行处理。最后通过比较发现,SVM在3个方面优于ANNs:SVMs训练和运行速度更快;SVMs能够训练大量的模式,而ANNs遇到大量模式的时候需要耗费大量的时间去训练,甚至可能出现模式难以聚集在一起的情况;SVMs轻松地实现更高的检测准确率。因为识别重要输入和冗余输入的能力会直接删减冗余数据、达到更快的校准以及很可能得到更加准确的结果,所以为了达到最优的性能,找出取证数据中的重要输入特征显得尤为重要。

5 现有冗余数据删减技术的分析与比较

5.1 数据来源比较

传统的数据分析方法通常可分成两种,一种是内容分析,另一种是主题分析。而分析的目标对象通常就是文本文件数据。现有用于入侵检测的冗余数据删减技术的主要目的是删除IDS误报,因此其分析数据对象是IDS报警集。对于入侵取证而言,数据删减主要用于减少需要分析的候选证据数量。候选证据来源众多,一个是系统方面,另外一个是网络方面。其中,来自系统方面的证据包括:系统日志文件,备份介质,程序、脚本、进程、内存镜像,交换区文件,临时文件,硬盘未分配的空间,系统缓冲区和打印机及其他设备的内存等等;来自网络方面的证据有:防火墙日志,IDS日志以及其他网络工具所产生的记录和日志等。因此,用于入侵取证的冗余数据删减方法需要处理的数据类型最为复杂,数据量最大,对删减方法

的要求也最高。如表1—表3所列。

表1 冗余数据删减技术的数据源比较

研究领域	数据来源
医学数据分析	文本数据文件
入侵检测	入侵检测系统(IDS)报警
入侵取证	系统证据和网络证据

表2 各数据源中的数据类型对比

数据来源	数据类型
文本数据文件	医学报告,面试记录,文档
入侵检测系统(IDS)报警	审计日志,网络流量信息
系统证据	系统日志,备份,程序,脚本,进程,内存,缓冲
网络证据	防火墙日志,IDS日志,网络工具产生的日志

表3 各数据源中的数据量对比

数据来源	数据量
文本数据文件	每个kBs大小
入侵检测系统(IDS)报警	每日MBs大小或者GBs大小
系统证据	每日MBs大小或者GBs大小
网络证据	每日MBs大小或者GBs大小

5.2 分析方法比较

传统冗余数据删减技术通常是通过对大数据集进行编码、分析频率及逻辑分布,获得大数据集的代表性或"摘要"性数据,从而达到减少需要分析的数据量,提高分析效率的目标。其实现删减的思想在于通过对原数据集的"抽象"得到一个在某种程度上可与之等价的小数据集,其结果并不影响原数据集。现有针对入侵检测及取证的冗余数据删减技术主要是借助数据挖掘方法、统计学方法及人工智能方法。其实现删减的思想在于对数据集的分类甄别。其处理过程是在原数据集上完成,能够实现数据量的真正减少。数据挖掘方法、统计学方法及人工智能方法的比较如表4、表5所列。

表4 冗余数据删减技术的方法比较

方法		实时性	自动化程度	性能	复杂度
数据挖掘方法	分类	高	中	高	中
	聚类	低	低	高	低
	孤立点检测	高	高	高	中
统计学方法	中	低	高	低	
人工智能方法	模糊决策树	高	高	高	低
	模糊专家系统	低	高	低	高
	人工神经网络	高	高	低	高
	支持向量机	高	高	高	中

表5 冗余数据删减技术性能指标的比较

方法		准确性	完备性	数据删减率
数据挖掘方法	分类	中	高	中
	聚类	中	高	中
	孤立点检测	高	高	高
统计学方法	中	高	中	
人工智能方法	模糊决策树	高	高	中
	模糊专家系统	中	中	低
	人工神经网络	中	中	中
	支持向量机	高	中	高

数据挖掘方法中分类方法的实时性比较高,但由于分类器的训练样本常常需要人工加标签,因此自动化程度一般。这类方法在分类器训练完成后处理性能比较高,复杂度一般。聚类的方法需要足够的背景知识(例如网络拓扑),而且分析模型的创建过于依赖专家经验。另外,该方法只能进行非实时的离线处理,不利于对攻击的及时响应。其算法的复杂度较低,故性能较高。孤立点检测算法无需任何背景知识,需要的人工干预也很少,所以它能克服现有方法在这方面的缺陷,大大提升了自动化程度。它能够实时过滤误报,而且能通过对新报警的不断学习来自动适应新的误报类型。虽然其算法复杂度一般,但是性能却比较好。

统计学的方法实时性一般,由于比较依赖于先验知识,因此自动化程度较低;而且由于算法复杂度较低,所以性能较高。

人工智能方法中除了模糊专家系统的实时性较低之外,其他方法都能较好地实现实时性。因为人工智能方法是研究使计算机来模拟人的某些思维过程和智能行为,故在很大程度上降低了人工干预,大大提升了自动化程度。模糊决策树的方法的性能有很多指标用于评估,通常采用命中率和误报率,该方法的高命中率以及低误报率使得其有较好的性能。模糊专家系统的方法应用于删减的模糊规则还不是很完善,故其删减数据性能较差。ANNs的方法性能由于受特征数量的影响较大,而SVMs的方法性能受特征数量的影响较小,SVMs训练和运行速度快而ANNs遇到大量模式的时候需要耗费大量的时间去训练,所以ANNs的性能较低而SVMs的性能较高。

结束语 随着数字信息的爆炸式增长,冗余数据也逐渐给各种科学领域的研究带来了严峻的挑战。冗余数据的存在形式各种各样,不同领域对于冗余数据删减的要求也不尽相同。入侵检测及取证冗余数据删减的需求是在保持合理的算法空间和时间开销的基础上,快速准确地找到冗余数据,将其与有用数据分离开来从而减少分析者的工作量,使分析者准确地把握安全形势;另外,为增强处理过程的准确性及合法性,该分离过程应尽量避免人工干预,实时地、自动化地对冗余数据进行删减。

针对这些需求,从目前的研究现状来看,利用当前各种方法能够在一定程度上实现自动删减冗余数据,减轻分析者负担。但是现有方法多处于实验阶段,实用性不高,尤其是人工智能技术等算法开销往往比较大,因此未来这一领域还有很多方面尚待深入研究和完善:(1)在算法的性能和复杂度方面,如何在快速准确地删减冗余数据的同时保证合理的算法开销是未来需要关注的问题。(2)在算法的自动化程度方面,现有的算法多少仍需要人工干预,无法完全做到自动化处理。(3)在应用场景方面,数据删减技术在入侵检测及取证领域可应用的地方很多,但目前仅见到对IDS误报删减的研究和网络、日志证据筛选的研究。实际上,如何利用数据删减技术在IDS检测之前筛选数据,从而提高检测机制的效率及准确率,以及如何利用数据删减技术识别入侵相关的文件、程序等各类证据等都是未来值得研究的课题。另外,在入侵证据删减方面,现有方法往往针对特定类型的候选数据,缺乏具有通用性的方法及统一的处理框架。(4)在删减的思想方面,除了现有基于分类甄别的直接删减方法,传统基于数据集抽象的思想是否也可以引入到入侵检测及取证数据删减过程之中,是否还有其他更好的删减思路,也是值得研究的问题。

参考文献

[1] Guest G, MacQueen K M. Data Reduction Techniques for Large Qualitative Data Sets [M]// Namey E, Guest G, Thairu L, et al. Handbook for Team-based Qualitative Research. Lanham: Altamira Press, 2008: 137-162

[2] MacQueen KM, McLellan E, Metzger D, et al. What Is Commu-

nity? An Evidence-based Definition for Participatory Public Health [J]. American Journal of Public Health, 2001, 91(12): 1929-1937

[3] Denzin N, Lincoln Y. Data Management and Analysis Methods [M]//Ryan G, Bernard R. Handbook of Qualitative Research. CA: Sage Press, 2000: 769-802

[4] Guest G, McLellan E. Distinguishing the Trees From the Forest: Applying Cluster Analysis to Thematic Qualitative Data [J]. Field Methods, 2003, 15(2): 186-201

[5] Barnett G, Danowski J. The Structure of Communication: A Network Analysis of the International Communication Association [J]. Human Communication Resources, 1992, 19(2): 264-285

[6] Richards D, Barnett G. Network Analysis of Message Content [M]// Danowski J. Progress in Communication Science. Norwood N J: Ablex Publishing Corporation, 1993: 198-221

[7] Pool IDS. The Representational Model and Relevant Research Methods [M]//Osgood C. Trends in Content Analysis. Urbana: University of Illinois Press, 1959: 33-88

[8] Aldenderfer M S, Blashfield R K. Cluster Analysis [M]. CA: Sage Press, 1984: 234-236

[9] Schiffman S, Reynolds SLM, Young FW. Introduction to Multidimensional Scaling: Theory, Methods, and Applications [M]. New York: Academic Press, 1981: 31-37

[10] Natick MA. ANTHROPAC 4.0 Methods Guide [M]//Borgatti SP. Analytic Technologies. USA: Columbia Press, 1996: 137-143

[11] 范明,孟小峰. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2007:98-102

[12] Pietraszek T. Using Adaptive Alert Classification to Reduce False Positives in Intrusion Detection [C]// French Riviera, France. Proceedings of RAID 2004. Heidelberg: Springer, 2004: 102-124

[13] Law KH, Kwok LF. IDS False Alarm Filtering Using KNN Classifier [C]//Proceedings of WISA 2004. Leju Island, Korea, Heidelberg: Springer, 2005: 114-121

[14] Alharby A, Imai H. IDS False Alarm Reduction Using Continuous and Discontinuous Patterns [C]// Proceedings of ACNS 2005. New York, USA, Heidelberg: Springer, 2005: 192-205

[15] Davenport M A, Baraniuk R G, Scott C D. Controlling False Alarms with Support Vector Machines [C]// Proceedings of ICASSP 2006. Toulouse, France, New York: IEEE Press, 2006: 589-592

[16] Julisch K, Dacier M. Mining Intrusion Detection Alarms for Actionable Knowledge [C]// Proceedings of KDD'02. Alberta, Canada, New York: ACM Press, 2002: 366-375

[17] Julisch K. Mining Alarm Clusters to Improve Alarm Handling Efficiency [C]//Proceedings of ACSAC'01. Los Alamitos, CA, New York: IEEE Press, 2001: 12-21

[18] Julisch K. Clustering Intrusion Detection Alarms to Support Root Cause Analysis [J]. ACM Transactions on Information and System Security, 2003, 6(4): 443-471

[19] Manganaris S, Christensen M, Zerkle D, et al. A Data Mining Analysis of RTID Alarms [J]. Computer Networks, 2000, 33(4): 571-577

[20] Nehinbe JO. Optimized Clustering Method for Reducing Challenges of Network Forensics [C]// Proceedings of CEEC 2010 2nd. Colchester, UK, New York: IEEE Press, 2010: 1-6

[21] 刘 静,赵宇驰. 数据挖掘领域中的聚类分析[J]. 东北林业大学学报,2012,40(8):13-19

[22] Ertoz L, Eilertson E, Lazarevic A, et al. Detection of Novel Network Attacks Using Data Mining [C]//Proceedings of DMSEC 2003. Melbourne, FL, USA, New York: IEEE Press, 2003: 1-10

[23] Dokas P, Ertoz L, Kumar V, et al. Data Mining for Network Intrusion Detection [C]//Proceedings of NSF Workshop on Next Generation Data Mining. Baltimore, USA, Cambridge: AAAI/MIT Press, 2002: 21-30

[24] Fu Xiao, Xie Li. Using Outlier Detection to Reduce False Positives in Intrusion Detection [C]// Proceedings of NPC 2008. Shanghai, China, New York: IEEE Press, 2008: 26-33

[25] Iwan S. False Alert Reduction System using Outlier Detection Methods [R]. UK: University of Southampton, 2010

[26] Axelsson S. The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection [C]//Proceedings of 6th ACM Conference on Computer and Communications Security. Singapore, New York: ACM Press, 1999: 1-7

[27] Goldman R. A Stochastic Model for Intrusions [C]// Proceedings of RAID 2002. Zurich, Switzerland, Heidelberg: Springer, 2002: 199-218

[28] Puttini R, Marrakchi Z, Me L. Bayesian Classification Model for Real-Time Intrusion Detection[C]// Proceedings of 22th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Idaho, USA, 2002: 150-162

[29] Mukkamala S, Sung A H. Identifying Significant Features for Network Forensic Analysis Using Artificial Intelligence Techniques [J]. International Journal on Digital Evidence, 2003, 1(4): 1-17

[30] Liu Z Q, Lin D D, Feng D G. Fuzzy Decision Tree based Inference Techniques for Network Forensic Analysis [J]. Journal of Software. 2007, 18(10): 2635-2644

[31] Kim J S, Kim M, Noh B N. A Fuzzy Expert System for Network Forensics [C]// Proceedings of the ICCSA 2004. Assisi, Italy, Heidelberg: Springer, 2004: 175-182

[32] Quinlan J R. Induction on decision trees [J]. Machine Learning, 1986, 1(1): 81-106

[33] Zadeh L A. Fuzzy Sets [J]. Information and Control, 1965(8): 338-353

[34] Setnes M, Babuska R, Verbruggen HB. Rule-based Modeling: Precision and Transparency [J]. IEEE Transaction on Systems, Man, and Cybernetics, 1998, 28(1): 165-169

[35] Takagi T, Sugeno M. Fuzzy Identification of Systems and Its Applications to Modeling and Control [J]. IEEE Transaction on Systems, Man, and Cybernetics, 1985, 15(1): 116-132

[36] Hertz J, Krogh A, Palmer R G. Introduction to the Theory of Neural Computation [M]. Boston, USA: Addison -Wesley, 1991: 18-30

[37] Demuth H, Beale M. Neural Network Toolbox User's Guide [R]. Math Works, Inc. Natick, MA, 2000

[38] Sung AH. Ranking Importance of Input Parameters of Neural Networks [J]. Expert Systems with Applications, 1998, 15(3/4): 405-41

[39] Vladimir V N. The Nature of Statistical Learning Theory[M]. Heidelberg: Springer, 1995: 98-99

[40] Joachims T. Estimating the Generalization Performance of a SVM Efficiently [C] // Proceedings of the 17th International Conference on Machine Learning. CA, USA, San Francisco: Morgan Kaufman, 2000: 431-438