文章编号:1671-4598(2017)10-0244-03 DOI:10.16526/j.cnki.11—4762/tp.2017.10.062 中图分类号:TP393 文献标识码:A

基于改进 Apriori 的网络安全感知方法

陆江东，郑 奋，戴卓臣
(第二军医大学 基础医学部，上海 200433)

摘要：针对网络安全态势评估过程中存在数据源单一、实时性不强、准确率不高的问题，提出一种基于改进关联规则算法（Apriori 算法）的网络安全态势感知方法；通过对数据的分析，发现在网络中存在关于安全态势的关联规则；通过网络攻击影响熵值序列的分析，对关联规则进行分类为空间正常和异常空间，进而对关联规则进行聚类分析；根据聚类后的规则划分网络安全态势等级；将改进后的算法应用到网络安全态势感知当中，实验结果表明，该方法满足了网络安全危险预警和实时监控的要求；改进的算法用于安全态势感知是可行的、有效的。

关键词：网络安全；关联规则；Apriori 算法；态势感知

Network Security Situation Awareness Method Based on Improved Apriori Algorithm

Lu Jiangdong, Zheng Fen, Dai Zhuochen
(College of Basic Medical Sciences, Second Military Medical University, Shanghai 200433, China)

Abstract: For the existing problems that data source is single, real—time is not strong, the accuracy rate is not high in the process of network security situation assessment, a network security situation awareness method based on algorithm of association rules is proposed. Through the analysis of the data, association rules about the security situation in the network can be found; based on network attack effect of entropy sequence analysis, association rules are classified for the space of normal and abnormal, and then the cluster analysis to association rules is carried on. Levels of network security situation are divided according to the clustered rules, the improved algorithm is applied to network security situational awareness, experimental results show that, the model can meet the requirements of the network security hazard warning and real—time monitoring. The improved algorithm used for security situational awareness is feasible and effective.

Keywords: network security; association rule; Apriori algorithm; situation awareness

0 引言

网络安全态势分析（NSSA）[1]是一门出现在网络当中的新兴技术，能够辅助安全专家对自己的网络进行实时、准确地安全评估。但是，由于网络安全的影响因素复杂多变，并且具有较大的不确定性，因此，构建一种网络安全态势评估系统对网络态势进行评估和预测成为一项重要的研究问题。

针对这一问题，研究者们提出了多种解决方法。Tim Bass[2]建立了融合多传感器数据的入侵检测框架，并且将其应用于新一代网络态势感知和入侵检测系统中。后来，Jason Shifflet 等[3]人也提出了类似的模型。2006 年，美国国防部预研发展署在本年的研究中，对网络态势感知系统所要研究的目的和一些相关技术做了详细的说明。陈秀真[4]等人通过研究提出了具有层次化的网络安全威胁衡量方法。赵国生等人[5]通过研究提出了基于灰色关联分析的网络可生存性衡量方法，将灰理论应用到网络态势感知中的态势预测模型中。

然而，现有方法仍然存在诸多问题。这类算法依赖于 Apriori 算法[7-8]，该算法有以下几个主要问题：1）多次扫描不是频繁项集的选项；2）进行候选间的自连接时，相同项目过多；3）未对数据库中的记录进行优化，重复扫描无用记录。因此，这类方法难以处理数据来源多样的情景，在实践中往往准确率较低、实时性较差。

针对这类问题，本文对经典的关联规则算法（Apriori 算法）进行改进，分别对候选项集的产生和判断过程进行优化，降低不必要的计算和存储过程。在此基础上提出了一种新网络安全态势的感知模型。该模型首先生成信息熵序列。接着对序列进行离散化，并运用本文改进的 Apriori 算法进行关联规则挖掘。然后对这些规则进行分类。最后根据分类后的关联规则进行态势风险指数生成。

为验证本文算法的有效性，通过在校园局域网上设计实验，进行网络安全监测。实验结果表明，本文提出的算法可以有效的达到实时监测要求。

1 改进的关联规则算法

传统的 Apriori 算法[7-8]面临以下几个主要问题：1）多次扫描不是频繁项集的选项；2）进行候选间的自连接时，相同项目过多；3）未对数据库中的记录进行优化，重复扫描无用记录。针对这些问题，本文从候选集产生、频繁项集判断及数据库操作 3 个方面对该算法进行改进。

1.1 候选项集产生过程的改进

从 k 项集生成 $k+1$ 项时，由于 L_k 的自连接操作中会产生很多与 k 项集相同的候选项，只要在生成前进行判断，如果两个项集的前 $k-1$ 项不同，则直接放弃对该记录的计算，移置下一个记录进行计算。

1.2 频繁项集判断过程的改进

首先，计算出 $|L_{K-1}(i_j)|$，同时，计算出 L_{K-1} 中所

收稿日期:2017-04-26； **修回日期**:2017-05-11。

作者简介:陆江东(1983-),男,江苏盐城人,硕士,讲师,主要从事计算机教学,数据挖掘方向的研究。

有事务项集的频率；然后，计算频率小于 $k+1$ 的所有项集，即 $I' = \{i \mid < \mid L_{K-1}\ (i_j)\ \mid < k+1\}$；通过删除 L_{K-1} 中所有与 I' 集合中的相同的频繁项集，得到一个数量更小的项集集合 L'_{K-1}；最后，通过 L'_{K-1} 的自连接操作直接生成 k 项集的候选集合。

1.3　数据库的优化

由于每次进行项集判断时，会频繁的对数据库进行连接，这将消耗大量资源、浪费很多时间。这是本文通过对事务进行标记，从而达到减少连接数据库的目的。

1.4　算法步骤

第一步：设置最小支持度 min_support 和最小置信度 min_confidence，连接数据库，获取 1 项集合的各种支持度，生成 1 项集合 L_1。

第二步：候选项集的产生。依据 2.2 节当中所描述的改进方法，对第一步所产生的 L_1 进行项集频率的排序（升序或者降序均可），生成 $k-1$ 项集 L_{k-1}。

第三步：频繁项集判断。按照 2.3 节所描述的改进方法，获取 L_{k-1} 中所有项集的频率信息，记录所有频率小于 $k+1$ 的项目 $I' = \{i \mid < \mid L_{k-1}\ (i_j)\ \mid < k+1\}$。然后，判断 L_{k-1} 中的项集是否是 I' 当中的元素，从而生成 C_k。

第四步：添加标记信息。根据 2.4 节所描述的改进方法，为不在 C_k 中的项集的事务添加标记信息，比如用“1”表示，然后删除标记为“1”的所有项集，得到新的候选项集信息。

重复第二步到第五步的计算过程，当所有事务均不能产生新的频繁项集的时候，停止循环。得到的结果即挖掘出来的规则。通过与最小置信度 min_confidence 进行对比，得到有效的关联规则。

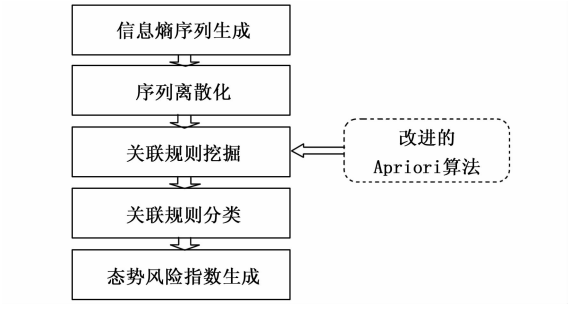

图 1　基于改进 Apriori 算法的网络安全态势感知方法

2　新的网络安全态势感知方法

网络当中受到的攻击进行检查和反馈的网络安全态势感知系统是加强网络安全的一种手段。针对网络安全态势感知的要求，本文在改进的关联规则的基础上提出了一种新的网络安全趋势的预知模型，如图 1 所示。该模型首先生成信息熵序列。接着对序列进行离散化，并运用本文改进的 Apriori 算法进行关联规则挖掘。然后对这些规则进行分类。最后根据分类后的关联规则进行态势风险指数生成。

2.1　网络安全态势感知的数据源分析

为了合理高效的使用收集到的网络数据，本文引入信息熵[6-8]这一概念及其技术。信息源被看作是一组随机事件的集合，该集合具有不确定性和随机性的特点。通信系统具有统计特征性，各个信息源信号出现的几率就形成了信息熵。

将流经某一个地址的不同类型事件进行分类，每一类视为一组随机事件，那么，这样运用信息理论的相关技术对这一系列随机事件进行分析。假设 $p\ (xi)$ 为某一随机事件的概率分布，$i=1, 2, \cdots n$，那么信息熵表示为：

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

而 $p(x_i)$ 由某一种测量数据中某个地址发生的次数 n_i 进行计算：

$$p(x_i) = \frac{n_i}{\sum_{i=1}^{n} n_i} \tag{2}$$

将 NetFlow 数据源与信息理论中的信息熵进行映射，形成新的数据源信息，下文的分析主要基于转换后的数据源信息。

2.2　信息熵数据源的生成

要通过信息论算法[9]进行处理，首先要将 NetFlow 数据转化为信息熵序列。网络信息被视为与时间相关的函数，将信息熵 H 假定为与时间 t 的函数：$H = f(t)$。通过计算时间 t_0 前所发生的信息熵时间序列规律，对时间 t_0 以后的信息熵进行预测，从而达到对网络的安全态势感知。信息熵时间序列表示为：

$$x = \{x_i \mid x_i \in \boldsymbol{R}, i = 1, 2, \cdots, L\}$$

2.3　信息熵的离散化

信息熵序列[10]是一组连续的信息，即一组连续函数，而在计算机当中，连续值的形式不易于处理，故要进行离散化处理。离散化是一个将连续模型或者函数转化为相同效果的离散值的过程。

2.4　关联规则挖掘与分类

该步骤使用 2.4 节所描述的改进后的 Apriori 算法对 3.2 节所获取的离散化后的数据源进行关联规则的挖掘与分类。

首先，设置合理的最小支持度和最小置信度。设置信息要根据不同的网络和不同安全级别进行设置。其次，在信息熵数据中应用 2.4 节改进后的 Apriori 算法，对网络安全态势关联规则进行挖掘，得到满足关联规则的最小可信度和最小支持度。然后，将关联规则进行相关分类。划分为正常空间和异常空间两大类，主要的依据：通过已知网络攻击的信息熵变化值进行类别判断。

2.5　态势的可视化

根据 3.3 节挖掘出来的关联规则信息，虽然能够很好的反应网络当中的一些安全态势，但是对于管理人员来说，这种数值化的信息不能很直观的进行阅读。故本文将已经挖掘出来的关联规则进行可视化表示。主要是通过对关联规则进行等级划分来实现。

要对已获取的关联规则进行网络态势进行划分，首先定义出安全级别的划分标准。3.3 节描述了如何将关联规则划分为两大类，即正常空间和异常空间，但是，两个级别显然不能描述网络当前的状态，故还要进行更进一步的划分。规定：若规则属于正常空间，认为网络目前处于一个安全级别；若规则处于正常空间，但对攻击行为不能确定，则判定网络风险位于一个中等级别；否则，判定此时网络风险位于一个危险级别。根据以上描述，依据网络的具体情况，定义不同级别的风险指数，即：

$$Risk\_Index = \begin{cases} 0 & t \text{ 时刻网络处于安全级别} \\ T_t & T_t \text{ 为 } t \text{ 时刻网络的危险度} \end{cases}$$

如果目前网络处于安全级别，此时网络的风险指数值为0。如果某一时刻同时存在一种或一种以上的网络攻击，则风险指数等于各种威胁度之和。网络安全专家通过对不同的网络定义来设定危险指数。根据风险指数，通过一些可视化软件，生成描述风险级别的可视化信息。根据可视化信息，网络的安全管理人员将能够更直观、便利的发现网络当中存在的威胁，能够依据这些可视化信息进行危险预测。

3 实验与分析

为了验证本文提出的网络安全态势感知模型的实践性和合理性，本文在实际环境中设计网络安全监测实验。

3.1 实验环境

实验通过校园局域网进行安全性检测。硬件设备包括Firewall，高性能千兆交换机，IDS系统，另外有模块化的多业务交换机和路由器[11]。所使用的设备均支持设定的SNMP、日志文件、NetFlow等多源数据源[12-14]，满足所提出模型数据源的要求。

3.2 实验过程及结果分析

本文将所提出的在关联规则基础上的网络安全态势预知模型运用到网络安全监测中。首先进行NetFlow数据流量信息的收集，以便将其转换为信息熵序列。在学校局域网中，取其中骨干网上面的一个节点Netflow数据进行分析，选择7天中的2 000个数据作为我们的测试样本点。通过3.1节描述的方法，生成信息熵序列，部分信息如图2和图3所示。

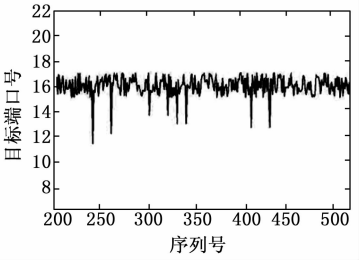

图2 源IP地址的部分信息熵序列

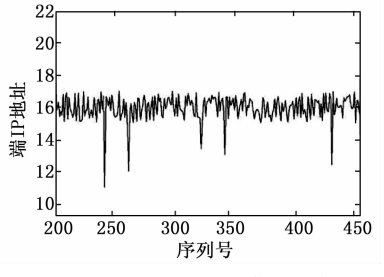

图3 目标端口号的部分信息熵序列

根据经验，定义已知网络攻击对熵值序列的影响，如表1所示。

表1 已知网络攻击对熵值序列的影响

攻击类型	影响熵值序列
SQL注入攻击	源IP地址、源端口号
UDP FLOOD攻击	源IP地址、目的IP地址
MSBLLAST蠕动攻击	目的IP地址、目的端口号
……	……

为了提高计算的效率，采用非均匀区间值离散化的方法，将每个信息熵序列分成5个区间A－E，每个区间的大小是不确定的。同时，对不同的序列，使用不同的序列号进行标注以示区别，分别为数字1－5。

然后，采用3.3节提出的方法对该离散化后的信息熵数据进行基于Apriori算法的关联规则挖掘。根据以往经验，设置最小支持度为0.01，最小置信度为0.02。通过计算，得到不同的项集信息，即不同项数的关联规则信息。表2显示了频5－项集和相关的支持度计数结果。

表2 频繁5－项集和支持度计数

频繁项集	支持度
{1A，2B，3C，4B，5E}	24
{1A，2B，3C，4E，5D}	35
{1B，2A，3E，4D，5C}	165
…	…

通过选择满足最小置信度的项集，得到网络安全态势的感知信息，即将该实验结果结合表1当中的信息，对当前网络的安全状态进行一个整体的评估。例如：{1A，2B，3C，4B，5E}的支持度为24，这个节点位置的目的端口地址熵值较大，但是目的地址的熵值相反，熵值非常小，由此可见，该节点的目的地址是比较聚集的，但是其目的端口却是分散的，将其称之为端口扫描攻击模式。

为了管理员更直观地了解当前网络的安全状况，对网络安全态势进行可视化描述。根据3.4节所描述的方法，根据这些关联规则对网络态势进行分级。网络安全态势感知的实时监控信息如图4所示。可见，本文提出的方法可以在较短时间内监控网络安全风险。最多可以在10秒内监控到11级风险。有效满足网络安全实时监控的需求。

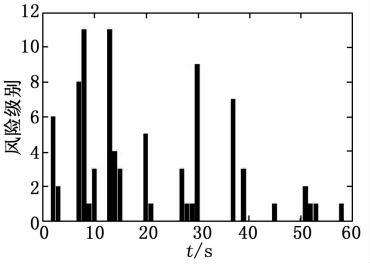

图4 网络安全态势感知的实时监控

4 结论

本文提出一种基于改进关联规则算法（Apriori算法）的网络安全态势感知方法。首先，该方法对所使用的数据源进行详细的描述；然后，以Apriori算法为理论基础，对相应的理论进行改善，得到改善后的Apriori算法。该算法通过对复杂项目集计数过程的中间结果进行过滤，把无效的数据进行删除，以此减小数据库规模。再者，基于改进的Apriori算法，给出了网络安全态势感知模型的具体实施步骤。最后，用实验验证了该框架的可行性并进行了结果分析。实验结果表明，本文提出的基于关联规则的网络安全态势感知模型能够实现了网络安全实时监控和危险预警的需求。

（下转第254页）

优选测试点。其中 T_1 的测试比较方便，选择 T_1 作为第 4 个优选测试点。

最终剩下 F_{10} 和 F_{12}，选择 T_{10} 作为第五个优选测试点。

根据上述分析，选择 T_{12}、T_9、T_{13}、T_1、T_{10} 作为故障隔离优选测试点。

表 4 简化的相关性数学模型(简化 $\boldsymbol{D}$ 矩阵)

		$T_1\sim T_2$	$T_3\sim T_8$	T_9	T_{10}	T_{11}	T_{12}	T_{13}
$\boldsymbol{D}$	$F_1\sim F_2$	1	1	1	1	1	1	1
	$F_3\sim F_8$	0	1	1	1	1	1	1
	F_{10}	0	0	0	1	0	1	0
$\boldsymbol{D}^{(1)}$	F_9	0	0	1	0	1	0	1
	F_{11}	0	0	0	0	1	0	1
	F_{12}	0	0	0	0	0	1	0
	F_{13}	0	0	0	0	0	0	1
$\boldsymbol{D}^{(1)}$	W_{FI}	4	6	6	6	4	6	4
$\boldsymbol{D}$	W'_{FI}	2	2	2	0	2	0	2

2.2.3.3 建立基层级测试诊断树

根据故障隔离测试点的选取结果和相关性矩阵建立测试诊断树如图 4 所示。

3 总结

从图 4 的测试诊断树可以看出，所有故障都可以被检测到；有 5 个组成单元可以被隔离到单个组成单元，即 F_9、F_{10}、F_{11}、F_{12}、F_{13}；有 7 个组成单元可以被隔离到两个组成单元，即 F_1、F_2、F_9、F_{10}、F_{11}、F_{12}、F_{13}。故障诊断能力计算也可以由相关性矩阵中只包含优选测试点的子矩阵计算得到。

故障检测率 $\gamma_{FD}=100\%$ 可以检测到所有组成单元的功能故障；隔离到单个组成的单元故障隔离率 $\gamma_{FI(1)}=5/13=38.5\%$

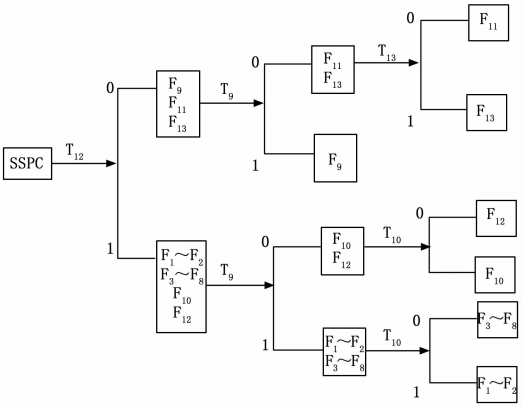

图 4 测试诊断树

隔离到单个组成单元（模糊度＝1）；隔离到 2 个组成单元的故障隔离率 $\gamma_{FI(2)}=(5+2)/13=53.8\%$（模糊度≤2）。

上述故障诊断能力计算结果表明 BIT 设计技术在配电系统中的应用可提高其系统的可靠性和智能化程度。

参考文献：

[1] 郑先成，等. 航天器新型固态配电技术研究 [J]. 宇航学报，2008，29 (4)：1430－1434.

[2] 姜东升，陈 琦，张 沛，等. 航天器供配电智能管理技术研究 [J]. 航天器工程，2012，21 (4)：100－105.

[3] 刘红奎. 基于 FPGA 的固态功率控制器的设计与实现 [D]. 西安：西安电子科技大学，2013：1－3.

[4] 赵 雷，王 磊，董仲博，等. 星载电子设备浪涌电流抑制以及浪涌电流的测试方法 [J]. 计算机测量与控制，2014，22 (9)：2730－2732.

[5] 杨文涛，张小林，吴建军. 无人机电源机内测试系统的设计与实现 [J]. 计算机测量与控制，2010，18 (7)：1509－1511.

(上接第 246 页)

参考文献：

[1] 郭方方，唐匀龙，修龙亭，等. 基于云计算的网络安全态势感知模型研究 [J]. 计算机学报，2014，47 (11)：149－151.

[2] Bass T. Intrusion detection system and multisensor data fusion：creating cyberspace situational awareness [J]. Communications of the ACM，2012，43 (4)：99－105.

[3] Shifflet J. A technique independent fusion model for network intrusion detection [A]. Proceedings of the Midstates Conference on Undergraduate Research in Computer Science and Mathematics [C]. 2012，3 (1)：13－19.

[4] 陈秀真，郑庆华，管晓宏，等. 层次化网络安全威胁态势量化评估方法 [J]. 软件学报，2013，17 (4)：885－897.

[5] 赵国生，王慧强，王 健. 基于灰色关联分析的网络可生存性态势评估研究 [J]. 小型微型计算机系统，2012，27 (10)：156－864.

[6] 杨启昉，马广平. 关联规则挖掘 Apriori 算法的改进 [J]. 计算机应用，2012，28 (12)：217－218.

[7] Shifflet J. A technique independent fusion model for network intrusion detection [A]. Proceedings of the Midstates Conference on Undergraduate Research in Computer Science and Mathematics [C]. 2013，3 (1)：13－19.

[8] 张春生. 改进的数据库一次扫描快速 Apriori 算法 [J]. 计算机工程与设计，2013，30 (16)：3811－3813.

[9] Ling X，Ren A S. Analysis on factors affecting quantity safety of agricultural products based on DEMATEL method [J]. Science & Technology and Economy，2012. 22 (1)：65－68.

[10] Sun B，Liu W，Tian D，et al. Application of time series date mining on security analysis [J]. Journal of Jilin University：Information Science Edition，2013，28 (3)：270－274.

[11] 刘华婷，郭仁祥，姜 浩. 关联规则挖掘 Apriori 算法的研究与改进 [J]. 计算机应用与软件，2012，26 (l)：146－148.

[12] Yu M，Hu M，Jin G，et al. Association rules algorithm applied to telecommunication network alarms [J]. Journal of Jilin University：Information Science Edition，2014，281 (3)：264－269.

[13] 钱光超，贾瑞玉，张 然，等. Apriori 算法的一种优化方法 [J]. 计算机工程，2012，34 (23)：196－198.

[14] Tang H Y，Zhan X Y. The research of the accessibility website design based log mining [J]. Computer Knowledge and Technology，2013，7 (4)：3261－3262.