Journal of System Simulation

Volume 29 | Issue 4 Article 21

6-3-2020

Cohesion Based Algorithm to Manage IDS Alerts

Jinlei Huang
PLA Information Engineering University, Zhengzhou 450004, China;

Hengjun Wang
PLA Information Engineering University, Zhengzhou 450004, China;

Bin Yu
PLA Information Engineering University, Zhengzhou 450004, China;

Cohesion Based Algorithm to Manage IDS Alerts

Abstract

Abstract: On the basis of intrusion taxonomies and semantic similarity, *the concept of cluster cohesion as well as an algorithm was proposed to manage IDS alerts.* Based on cohesion, the proposed approach *used improved bisecting K-means to aggregate massive alerts, and extracted the abnormal alerts from clusters formed in aggregation*. The experimental results show that the approach *is effective in alerts aggregation and abnormal alerts detecting, and can generate understandable meta-alerts with higher accuracy*.

Keywords

alerts aggregation, anomaly extraction, semantic similarity, cohesion, improved bisecting k-means

第 29 卷第 4 期 系统仿真学报© Vol. 29 No. 4
2017 年 4 月 Journal of System Simulation Apr., 2017

一种基于凝聚度的报警处理算法

黄金垒，王衡军，郁滨
（解放军信息工程大学，河南 郑州 450004）

摘要：*在研究分类系统和语义相似度的基础上，给出了簇的凝聚度的概念，提出了一种基于凝聚度的报警处理算法。算法以凝聚度为基础，利用改进的二分 K 均值算法聚合报警，并对聚合结果进行异常提取。实验结果表明，提出的算法能有效聚合大量报警、发现异常报警，且聚合结果具有良好的语义和较高的准确性。*

关键词：报警聚合；异常提取；语义相似度；凝聚度；改进二分 K 均值
中图分类号：TP309.1 文献标识码：A 文章编号：1004-731X (2017) 04-0859-06
DOI: 10.16182/j.issn1004731x.joss.201704021

Cohesion Based Algorithm to Manage IDS Alerts

Huang Jinlei, Wang Hengjun, Yu Bin
(PLA Information Engineering University, Zhengzhou 450004, China)

Abstract: On the basis of intrusion taxonomies and semantic similarity, *the concept of cluster cohesion as well as an algorithm was proposed to manage IDS alerts.* Based on cohesion, the proposed approach *used improved bisecting K-means to aggregate massive alerts, and extracted the abnormal alerts from clusters formed in aggregation.* The experimental results show that the approach *is effective in alerts aggregation and abnormal alerts detecting, and can generate understandable meta-alerts with higher accuracy.*

Keywords: alerts aggregation; anomaly extraction; semantic similarity; cohesion; improved bisecting k-means

引言

报警聚合是解决入侵检测系统产生大量重复报警的重要手段，旨在将同一安全事件诱发的大量性质相同或相近的报警合并成一个报警(超报警)[1]，能有效减少报警冗余，便于网络安全人员分析攻击实质和排查误报原因。相似度是报警聚合的基础，一般的报警相似度主要通过各属性的简单数值运算进行定义[2-3]，是一种基于几何距离的相似度，其语义需人工进一步理解。Julisch[4]结合入侵检测系统所在网络的特征，通过构建分类系统确定了报警间的层次关系。Saad[5]等沿用类似的思想，提出了语义相似度的概念。区别于基于几何距离的相似度，语义相似度依据分类系统定义，具有语义清晰，利于生成超报警的优点。

在报警聚合实现方面，聚类技术能够根据相似度将数据集自动分簇，被广泛应用到报警聚合中来。文献[3, 5]为不同的攻击类型设置不同的阈值，采用顺序聚类的方法，将相似度高于阈值的报警进行合并，操作简单，适用范围广，但具有较强的次序依赖性。文献[6]通过 EM(Expectation Maximization)算法聚合报警信息，避免了聚类结果对报警处理次序的依赖，并通过引入 PCA(Principal Components

收稿日期：2016-05-22 修回日期：2016-08-04；
作者简介：黄金垒(1991-)，男，四川雅安，硕士生，研究方向为仿真、信息安全技术；王衡军(1973-)，男，湖南衡阳，博士，副教授，硕导，研究方向为人工智能、信息安全等。

Analysis)对报警进行特征降维，使问题得到了一定的简化，但该方案收敛速度较慢，难以适用于分布式、高速网络产生的大量报警。文献[7]使用混沌粒子群与 K 均值算法相结合的方式对报警进行聚合，能较好的发现初始聚类中心，有利于寻找全局最优解；文献[8]采用自组织特征映射(Self-Organizing Map, SOM)聚合报警，SOM 具有加强邻近簇相关性的特点，为发现报警间的关系提供了新的思路。文献[7-8]都将重点放在报警减少上，取得了较高的报警精简率，但对聚合准确性关注较少。

目前大多数的研究还很少涉及异常报警的问题，异常报警也常因报警聚合而被掩盖。然而异常报警往往预示着某种新型的攻击或特殊的网络行为，因而对异常报警的处理有着更加重要的意义和价值。

综上，聚类技术应用于报警聚合中存在准确性有待提高、异常报警易被忽略以及语义欠佳等问题。本文在分类系统和语义相似度的基础上，提出了一种基于凝聚度的报警处理算法。算法结合二分 K 均值[9](Bisecting K-Means)快速有效的特点，使用改进的二分 K 均值算法聚合报警，并对聚合结果进行了异常提取。数据集测试验证了算法的有效性，结果表明，本文提出的算法能以较高的准确性聚合大量报警，有效发现异常报警，并生成语义良好的超报警。

1 相关概念及定义

本节首先定义了报警与簇的概念，并对报警进行了形式化表示，再基于分类系统和语义相似度，给出了超报警满足的条件及超报警生成算法，最后提出了簇的凝聚度的概念。为方便描述，所用符号及其含义如表 1 所示。

定义 1　一条报警 a 是由攻击类型、源 IP、源端口、目的 IP、目的端口、时间戳等 6 种属性构成的六元组，形式化表示为：
a=($Type$, $SrcIP$, $SrcPort$, $DstIP$, $DstPort$, $Time$)。

表 1　符号含义表
Tab. 1　List of symbols' meaning

符号	含义
$Alerts$	报警集
a,b	报警集中的任意两条报警
C	簇
Clusters	簇集
A	报警属性集
A_i	报警属性集中的第 i 个属性
$dom(A_i)$	属性 A_i 的值域
$a[A_i],b[A_i]$	报警 a,b 在属性 A_i 上的值
T_i	属性 A_i 对应的分类系统
$node_i^1 \lhd node_i^2$	T_i 中 $node_i^2$ 是 $node_i^1$ 的祖先

定义 2　报警集中相似的报警构成的集合称为簇。

报警间的相似程度用相似度衡量，本文采用语义相似度(见定义 4)。

定义 3[4]　属性 A_i 的分类系统是一棵由 $dom(A_i)$ 中元素构成的树，记为 $T_i=\{x \mid x \in dom(A_i)\}$，用于描述 $dom(A_i)$ 中元素间的关系。

定义 4[5]　报警 a,b 在语义上的相似程度称为报警 a,b 的语义相似度，记为 $Sim(a,b)$，计算公式如式(1)所示。

$$Sim(a,b)=\frac{\sum_{i=1}^{n} Sim(a[A_i],b[A_i])}{n} \tag{1}$$

其中 $Sim(a[A_i],b[A_i])$ 表示报警 a,b 关于属性 A_i 的相似程度，可依据相应的分类系统 T_i 确定。

定义 5　簇 C 的超报警 $MetaAlert(C)$ 是一条能概括 C 典型特征的报警，满足条件：

(1) 对 $\forall a \in C$，$\forall A_i \in A$，有 $a[A_i] \lhd MetaAlert(C)[A_i]$ 成立。

(2) $max(\sum_{a \in C} Sim(a, MetaAlert(C)))$。

超报警 $MetaAlert(C)$ 与普通报警具有相同的格式，但 $MetaAlert(C)$ 可能只是一条虚拟的报警，并不一定存在于 C 中。超报警可由如下算法生成。

算法 1　超报警生成

输入：簇

输出：超报警

Step 1：根据定义 1，对于报警的每一属性 A_i，

依据先验知识构建相应的分类系统 T_i。

Step 2：从簇中任取一条报警作为超报警。

Step 3：从簇中取下一条报警。

Step 4：更新超报警：对于每一属性 A_i，用两报警在相应的属性分类系统 T_i 中的最近共同祖先作为新的超报警的属性值。

Step 5：若簇中报警处理完毕，算法结束；否则，转到 Step 3。

定义 6　超报警 $MetaAlert(C)$ 与簇 C 中所有报警的平均语义相似度称为簇 C 的凝聚度，记为 $cohesion(C)$，计算公式如式(2)所示。

$$cohesion(C)=\frac{\sum_{a\in C} Sim(a,MetaAlert(C))}{size(C)} \tag{2}$$

其中：size(C) 表示簇 C 包含的报警数量。簇的凝聚度反映了簇中报警与其超报警的紧密程度，相比于一般的基于均方误差的度量方式，凝聚度更能反映实际的报警聚集情况，因而更具合理性。

从簇中删除一条报警后会引起簇的凝聚度的变化，对簇 C 中的一条报警 a，删除 a 后簇 C 的凝聚度增长率记为 $\Delta\overline{cohesion(C,a)}$，计算公式如式(3)所示。

$$\Delta\overline{cohesion(C,a)}=\frac{cohesion(C-\{a\})-cohesion(C)}{cohesion(C)} \tag{3}$$

2　算法描述

算法主要分为报警聚合与异常提取两个部分，其整体结构如图 1 所示。在报警处理过程中，簇的凝聚度用于从簇集中选择待分裂的簇，而簇的凝聚度的增长率用于对异常报警进行提取。

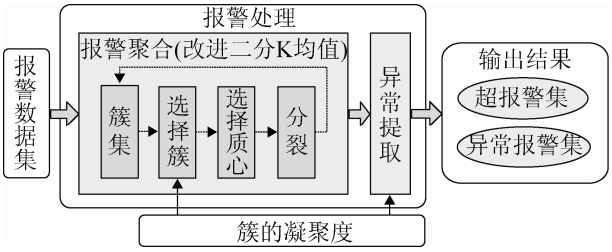

图 1　算法整体结构

Fig. 1　The algorithm's structure

2.1 报警聚合

基于凝聚度，使用改进的二分 K 均值算法对报警数据集进行聚合。二分 K 均值是对 K 均值的改进，它通过多次有选择的执行质心数为 2 的 K 均值算法，实现对数据集的分裂式聚类，能有效适用于多种数据类型，并具有与层次聚类算法相当的准确性[9]，比较符合报警聚合准确性要求高，报警数量大的特点。但二分 K 均值在每次分裂时会受到初始质心选取的干扰，且以均方误差确定待分裂簇的方式并不能反映实际的报警凝聚程度。

针对二分 K 均值算法在报警聚合中存在的不足，作以下两方面的改进：(1) 使用簇的凝聚度代替均方误差选择待分裂的簇，提高报警聚合的准确性；(2) 使用改进的质心生成方式，以减少随机选择初始质心对聚合结果的干扰，并提高算法的收敛速度。

算法 2　报警聚合

输入：报警集

输出：簇集

Step 1：设置凝聚度门限 *Th_cohesion*，并将簇集置为空集。

Step 2：将报警集加入簇集，作为待分裂簇；转到 Step 4。

Step 3：选择簇：根据公式(2)计算簇集中各簇的凝聚度，并从簇集中选择凝聚度最小的簇作为待分裂的簇。

Step 4：为待分裂簇生成质心：从待分裂簇中随机取一条报警 a，根据公式(1)计算 a 与簇中其余报警的语义相似度，并找出使得语义相似度 $Sim(a,b)$ 最小的报警 b，报警 a, b 即为进行分裂所需的质心。

Step 5：以 a, b 为初始质心执行质心数为 2 的 K 均值算法，二分选定的簇，得到两个新簇。

Step 6：将新簇加入簇集。

Step 7：若簇集中所有簇的凝聚度均大于或等于门限 *Th_cohesion*，算法结束；否则，转到 Step3。

在 Step 1 中，凝聚度门限 *Th_cohesion* 用于控制报警的聚合程度，本文采用如下方式确定：设 *Th_cohesion* 的初始值为 0.15，以 0.025 为步长，在不同的 *Th_cohesion* 下多次运行算法，得到不同的簇数量。以 *Th_cohesion* 为自变量，画出 *Th_cohesion* 与簇数量的关系图，簇的数量增长明显加快处所对应的 *Th_cohesion* 即为较好的凝聚度门限值。

2.2 异常提取

异常报警与簇中大多数报警具有显著不同的特征，在空间上表现为远离簇中大部分的报警，当异常报警被移除后，会使得簇的凝聚度显著增加。基于上述思想，同时考虑不同簇的大小有所不同，使用凝聚度增长率与簇中报警数量的乘积作为判别报警是否为异常的依据。异常提取算法描述如下。

算法 3　异常提取

输入：簇集 $Clusters=\{C_0,C_1,\cdots,C_k\}$

输出：超报警集和异常报警集

Step 1：设置异常门限 *Th_abnormal*，并令 $i=0$。

Step 2：从 *Clusters* 中取簇 C_i，统计 C_i 中包含的报警数量 $size(C_i)$；对簇 C_i 中每条报警 a，根据公式(3)计算删除 a 后提供的凝聚度增长率 $\Delta\overline{cohesion(C_i,a)}$。

Step 3：将 C_i 中所有满足 $\Delta\overline{cohesion(C_i,a)}\cdot size(C_i)$ 大于 *Th_abnormal* 的报警 a 标记为异常报警，加入异常报警集，并从原有簇 C_i 中移除。

Step 4：根据算法 1，为簇 C_i 重新计算超报警 $MetaAlert(C_i)$，并将其加入超报警集。

Step 5：若簇集 *Clusters* 中所有簇均处理完毕，算法结束；否则，$i=i+1$，并转到 Step 2。

在 Step 1 中，异常门限 *Th_abnormal* 用于控制算法的异常提取程度，本文采用如下方式确定：将 *Th_abnormal* 的初始值设为 0.52，以 0.02 为步长，在不同的 *Th_abnormal* 下多次运行算法，得到不同的异常报警数量，画出 *Th_abnormal* 与异常报警数量的关系图，曲线拐点处或异常报警数量明显减少处所对应的 *Th_abnormal* 即为较好的异常门限值。

3　实验及结果分析

实验原始数据采用 MIT Linclon 实验室提供的 DARPA99 数据集。DARPA99 数据集是入侵检测领域广泛使用的基准评测数据集，共包含五周模拟数据，实验对 DARPA99 中的部分数据进行了重放，共收集到 31 278 条报警信息。

算法使用凝聚度门限 *Th_cohesion* 控制报警聚合情况，在不同的凝聚度门限下，算法的执行会得到不同的簇数量，其相互关系如图 2 所示。当凝聚度门限线性增加时，簇的数量近似指数增长，并且在凝聚度门限大于 0.475 后增长速度显著加快。簇的数量过多或过少都不利于进一步处理，由以上分析知，当凝聚度门限 *Th_cohesion*=0.475 时聚合效果较理想，此时算法得到 1 126 个簇，最终生成 1 126 个超报警，报警精简率为 96.4%。

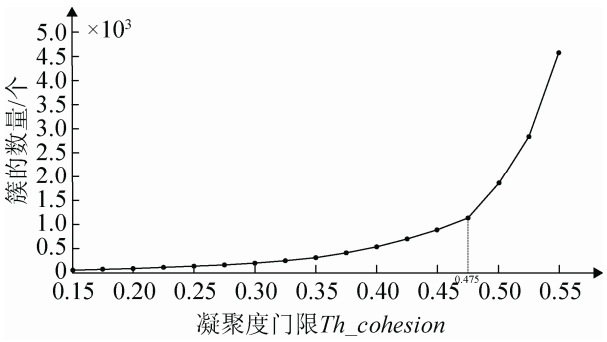

图 2　簇的数量随凝聚度门限变化情况

Fig. 2　The number of clusters changes with *Th_cohesion*

对报警聚合得到的 1 126 个簇进行异常提取，异常报警数量随异常门限 *Th_abnormal* 的变化情况如图 3 所示。随着 *Th_abnormal* 的增大，提取到的异常报警数逐渐减小，异常报警与簇的偏离程度也越大。在异常门限 *Th_abnormal*=0.64 处，曲线出现拐点，即此时的异常门限能更好地划分正常报警与异常报警。

DOI: 10.16182/j.issn1004731x.joss.201704021

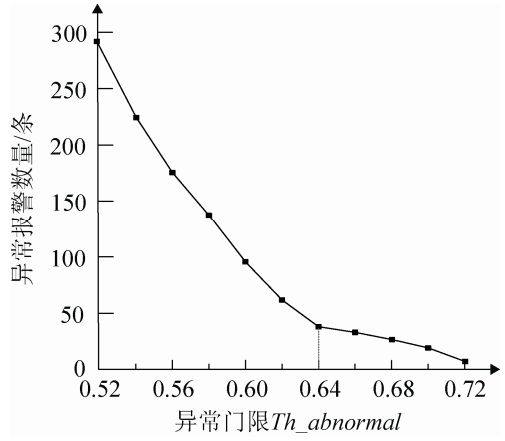

图 3 异常报警数量随异常门限变化情况
Fig. 3 The number of abnormal alerts changes with *Th_abnormal*

表 2 不同方案聚合准确性与时间开销对比
Tab. 2 Comparison of different approaches in accuracy and time cost

对比指标	二分 K 均值	文献[8]	凝聚度方案
Purity	0.726	0.830	0.925
Time/s	344.5	681.2	423.8

考虑算法的聚合准确性与时间开销，以相同的报警集对普通二分 K 均值和文献[8]的 SOM 算法进行测试，并与本文的凝聚度方案进行对比(*Th_cohesion*=0.475)，结果如表 2 所示，其中纯度 *Purity* 表示簇由单个类的报警组成的程度，用于反映报警聚合的准确性。设报警集实际划分为 N 个类，聚合后得到 K 个簇，m 表示报警总数，m_i 表示簇 i 中报警的数量，m_{ij} 表示簇 i 中属于类 j 的报警数量，则簇 i 的纯度 $p_i = \max_j(\frac{m_{ij}}{m_i})$，报警聚合形成的簇集的总体纯度 $Purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i$。由表 2 可知，本文的凝聚度方案准确性最好，而时间开销略大于普通二分 K 均值算法，因此在报警聚合中具有优势。

部分异常提取结果显示在表 3 中，随异常门限 *Th_abnormal* 的增大，异常总数与异常类型种类均逐渐减少。当 *Th_abnormal*≥0.60 时，所有属于 DOS 类的 mail_bomb 和 teardrop 两种具体攻击类型从异常报警集中消失；当 *Th_abnormal*=0.72 时，所有属于 R2L 类的 guess_password、xlock、worm 3 种具体攻击类型以及属于 U2R 类的 perl 攻击类型从异常报警集中消失。此时报警集中仅余 U2R 类的 1 条 xterm 攻击和 Probe 类的 5 条 portsweep 攻击。

当 *Th_abnormal*=0.64 时(对应于图 3 中曲线的拐点)，算法提取到 3 条攻击类型为 xterm 的报警，2 条攻击类型为 worm 的报警，5 条攻击类型为 xlock 的报警，以及 27 条攻击类型为 portsweep 的报警，共计 37 条。依据 DARPA99 数据集的攻击类型分布情况，xterm、worm 和 xlock 属于新的攻击类型，因此算法具有较好的异常提取能力并能从异常报警中发现新型攻击。

表 4 显示了不同方案中 IP 扫描攻击生成的超报警的比较结果。文献[2-3]的超报警生成方法需进一步统计和分析源端口、目的 IP、目的端口以及时间跨度等信息，而本文生成的超报警清楚地表明外网 IP 地址为 10.104.56.32 的主机对内网中 192.168.X.X 的网段实施了 IP 扫描攻击，发起攻击和受攻击的主机端口均为私有端口，相应簇中共包含 136 条报警信息，时间跨度从 8:25:22 到 8:28:09，共计 167 s。相比于文献[2-3]，本文聚合得到的超报警清晰、直观，为分析网络攻击提供了更多便利。

表 3 不同异常门限下的异常提取结果
Tab. 3 Extraction results in different *Th_abnormal*

Th_abnormal	异常总数	描述(攻击类型：数量)
0.52	291	xterm:3\perl:2\worm:2\guess_password:88\teardrop:12\mail_bomb:117\xlock:5\portsweep:62;
0.56	174	xterm:3\perl:2\worm:2\guess_password:88\teardrop:12\xlock:5\portsweep:62
0.60	95	xterm:3\perl:2\worm:2\guess_password:34\xlock:5\portsweep:49
0.64	37	xterm:3\worm:2\xlock:5\portsweep:27
0.68	26	xterm:3\worm:2\portsweep:21
0.72	6	xterm:1\portsweep:5

表 4 超报警的生成结果比较
Tab. 4 Meta-Alert results comparison

	Type	SrcIP	SrcPort	DstIP	DstPort	Time	(Size)
本文	ip sweep	10.104.56.32	PRIV	192.168.X.X	PRIV	8:25:22-8:28:09	136
文献[2-3]	ip sweep	10.104.56.32	无	无	无	无	136

5 结论

给出了簇的凝聚度的概念并设计了一种基于凝聚度的报警处理算法。算法依据凝聚度选取待分裂的簇，改进了原有二分 K 均值的质心生成方式，并使用簇的凝聚度增长率对聚合结果进行了异常提取。实验结果表明，算法能准确聚合大量报警，有效提取异常报警，并生成语义良好的超报警。未来的工作包括误报警处理、算法的并行化研究以及基于语义的入侵响应自动化等。

参考文献:

[1] 穆成坡, 黄厚宽, 田盛丰. 入侵检测系统报警信息聚合与关联技术研究综述 [J]. 计算机研究与发展, 2006, 43(1): 1-8. (Mu Chengpo, Huang Houkuan, Tian Shengfeng. Survey of Intrusion-Detection Alert Aggregation and Correlation Techniques [J]. Computer Research and Development, 2006, 43(1): 1-8.)

[2] 王琢, 范九伦, 刘建华. 入侵检测系统报警信息聚合方法的改进 [J]. 计算机工程与应用, 2010, 46(7): 107-109. (Wang Zhuo, Fan Jiulun, Liu Jianhua. Improved Aggregation Algorithm for Intrusion-Detection Alerts [J]. Computer Engineering and Applications, 2010, 46(7): 107-109.)

[3] 郭帆, 余敏, 叶继华. 一种基于分类和相似度的报警聚合方法 [J]. 计算机应用, 2007, 27(10): 2446-2449. (Guo Fan, Yu Min, Ye Jihua. Alert Aggregation Algorithm Based on Category and Similarity [J]. Computer Applications, 2007, 27(10): 2446-2449.)

[4] Klaus Julisch. Using Root Cause Analysis to Handle Intrusion Detection Alarms [D]. Germany: University Dortmund, 2003.

[5] Saad S, Traore I. A Semantic Analysis Approach to Manage IDS Alerts Flooding [C]// Information Assurance and Security (IAS), 2011 7th International Conference on. USA: IEEE, 2011: 156-161.

[6] Siraj M M, Maarof M A, Hashim S Z M. Intelligent Clustering with PCA and Unsupervised Learning Algorithm in Intrusion Alert Correlation [C]// Information Assurance and Security 2009. Fifth International Conference on. USA: IEEE, 2009, 1: 679-682.

[7] 胥小波, 蒋琴琴, 郑康锋, 等. 基于混沌粒子群的 IDS 告警聚类算法 [J]. 通信学报, 2013, 34(3): 105-110. (Xu Xiaobo, Jiang Qinqin, Zheng Kangfeng, et al. IDS Alert Clustering Algorithm Based on Chaotic Particle Swarm Optimization [J]. Journal on Communications, 2013, 34(3): 105-110.)

[8] Ahrabi A A A, Navin A H, Bahrbegi H, et al. A New System for Clustering and Classification of Intrusion Detection System Alerts Using Self-Organizing Maps [J]. International Journal of Computer Science and Security (IJCSS) (S1985-1553), 2010, 4(6): 589-597.

[9] Steinbach M, Karypis G, Kumar V. A Comparison of Document Clustering Techniques [J]. KDD Workshop on Text Mining(S2095-2236), 2000, 400(1): 525-526.

DOI: 10.16182/j.issn1004731x.joss.201704021