DOI :10.11721/cqnuj20130118

一种基于 GFKM 的集群入侵检测模型*

徐艳群[1],张 斌[1],秦小铁[2]
(1. 南阳理工学院 计算机科学与技术系,河南 南阳 473004;2. 重庆科技学院 图书馆,重庆 401331)

摘要:为了提高入侵系统的检测率和检测速度,论文提出一种基于灰色 K 均值聚类算法的集群入侵检测模型。利用灰色关联分析理论对原始数据进行预处理,根据 $\eta_{ij} = \frac{1}{n-1}\sum_{2}^{n}\xi ij(k)$ 计算相关度,再对原始数据集合进行聚类;最后引入集群技术,将 GFKM 算法装入集群系统中的每个检测结点形成集群入侵检测模型。最后,通过仿真实验对该模型进行了验证,结果表明,GSFK 算法应用于入侵检测模型中出现的误报率为 0.31%,漏报率为 0.34%,而且该模型呈现出较好的泛化性,应用于网络入侵检测中具有较好的性能。

关键词:集群系统;入侵检测;K 均值聚类算法;聚类分析
中图分类号:TP309 文献标志码:A 文章编号:1672-6693(2013)01-0081-03

1 基本概念

1.1 灰色关联分析理论

灰色关联分析理论的原理:通过对统计序列几何关系的比较来计算系统中各个因素之间的关联程度,统计序列曲线的几何形状越相近,则他们之间的关联程度越大[1]。

自 1982 年邓聚龙教授提出邓氏关联度模型以来,灰色关联分析理论在各个领域得到了广泛应用,在应用过程中,学者们发现邓氏关联度模型存在一些缺陷,学术界为此展开了一场对灰关联度的深入研究。经过数十年的发展,关联度分析理论得到了很大的发展和改进,表 1 描述了各种关联度计算方法的比较。

表 1 各种灰色关联度计算方法的比较

关联度	规范性	无量纲化后的保序性	正负相关性
邓氏关联度	不满足	不具备	不体现负相关性
绝对关联度	不满足	不具备	不体现负相关性
斜率关联度	不满足	不具备	不体现负相关性
T 型关联度	不满足	满足	体现
改进 T 型关联度	满足	满足	体现

1995 年唐五湘通过增量来描述两序列的关联度,提出了 T 型关联度计算公式;此后,有很多学者对 T 型关联度公式进行研究和改进,其中,较为有影响力的是 2008 年孙玉刚、党耀国提出的改进 T 型关联度。从表 1 中可以看出原有的关联度模型普遍不满足规范性,在量纲化后不能保持原来顺序,同时,也不能体现正负相关性。改进 T 型关联度能够克服这些缺陷,具有规范性,无量纲化后保序性,正确体现正负相关性等优点,论文采用该灰关联度计算公式来计算数据流相关度。

1.2 数据流关联度

网络攻击主体上可以分为 4 类:分布式拒绝服务攻击(Distributed denial of service,DDOS)、扫描攻击、木马攻击和非法提升权限攻击[2]。扫描攻击和 DDOS 攻击的特点是向目标主机连续发送数据,可以通过检测数据流通信信息来检测是否存在这两种攻击类型。木马攻击的特点是受害主机向外发送违反安全规范的数据包,可以通过检测数据包的内容和非法开放端口来检测木马攻击。非法提升权限攻击包括远程到本地 R2L 和普通用户到管理员用户 U2R 共 2 种,其中 R2L 是通过获取远程用户的信息登录到远程计算机,而 U2R 是普通用户获取管理员用户的密码并用管理员账号登录到计算机系统,因此,权限提升攻击可以通过数据包的连接特征来检测[3]。

综合上述分析,本文提取数据流中数据通信、内容和网络连接 3 种特征,其中数据流通信信息提取连接统计信息,即在固定时间内到达同一主机的的网络连接数、内容特征提取操作系统报错信息和非法开放端口。网络连接特征可以使用 TCPDUMP 工具分析获取,主要提取源 IP 地址、目的 IP 地址、连接持续时间、服务类型、协议和连接标志位[4]。

对处理后的数据流格式进行数学建模:对于任意数据流 F,属性集合抽象成 10 维向量:$f = \{m(F),$

* 收稿日期:2012-05-02 修回日期:2012-10-15 网络出版时间:2013-01-18 15:05
资助项目:河南省科技厅 2011 年软科学项目(No. 112400450137)
作者简介:徐艳群,女,讲师,硕士,研究方向为数字图像处理及计算机应用,E-mail:85661336@qq.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20130118.1505.201301.81_017.html

$s(F)$,$d(F)$,$a(F)$,$b(F)$,$t(F)$,$e(F)$,$p(F)$,$f(F)$,$n(F)$}。其中每个属性变量所表示的意义如表2所示[5]。

假设Fi是任意待聚类的数据流,Fj是已聚类的数据流,则参考序列为Fi的属性集合fi,比较序列为Fj的属性集合fj,Fi对Fj中元素的关联系数记为ξij,关联度记为ηij[6]。

$$fi = \{m(Fi), s(Fi), d(Fi), a(Fi), b(Fi), t(Fi), e(Fi), p(Fi), f(Fi), n(Fi)\}$$

$$fj = \{m(Fj), s(Fj), d(Fj), a(Fj), b(Fj), t(Fj), e(Fj), p(Fj), f(Fj), n(Fj)\}$$

$$\xi ij(k) = \begin{cases} \operatorname{sgn}(Zi(k)Zj(k)) \dfrac{1}{1+\frac{1}{2}\| Zi(k)| - |Zj(k)\| + \frac{1}{2}\left(1 - \frac{\min(Zi(k), Zj(k))}{\max(Zi(k), Zj(k))}\right)}, & \text{当 } Zi(k), Zj(k) \text{ 不同时为 } 0 \\ 1, & \text{当 } Zi(k), Zj(k) \text{ 同时为 } 0 \end{cases} \quad (1)$$

其中,式中涉及变量计算如下[7]

$$Zi(k) = \frac{yi(k)}{Di}; Di = \frac{1}{n-1}\sum_{n=2}^{n} | yi(k) |$$

$$yi(k) = Xi(k) - Xi(k-1)$$

关联系数反映数据流中2个对应元素之间的关联程度,所有元素关联系数的均值(即关联度)反映2个数据流的关联程度,关联度的计算公式为[8]

$$\eta ij = \frac{1}{n-1}\sum_{2}^{n} \xi ij(k) \quad (2)$$

当关联度$-1 \leqslant \eta ij < 0$时,fi,fj负相关;当关联度$0 < \eta ij \leqslant 1$时,fi,fj正相关;当关联度$\eta ij = 0$,则fi,fj不相关。

表2 数据流各个属性特征描述

属性	意义	属性	意义
$m_{(F)}$	系统提示信息	$t_{(F)}$	连接持续时间
$s_{(F)}$	源IP	$e_{(F)}$	服务类型
$d_{(F)}$	目的IP	$p_{(F)}$	协议类型
$a_{(F)}$	源端口	$f_{(F)}$	连接标志位
$b_{(F)}$	目的端口	$n_{(F)}$	固定时间内连接主机数

2 基于GFKM的集群入侵检测模型

2.1 GFKM算法

在K均值算法基础上改进的GFKM算法仍然是一种迭代的启发式算法,在迭代过程中更新聚类成员和聚类中心,直到得到理想的簇集。GFKM算法基本过程如下。

1)初始化n个聚类,没有聚类A,聚类中心为Fa,聚类权重为m。

2)对于任意Fi,按照(1)式和(2)式计算2个数据流的关联度,若关联度小于预定阀值r,跳转到步骤5。

3)若关联度大于r,将Fi并入聚类A,同时调整聚类中心(该聚类中所有对象的均值),将聚类的权重加1,即$m = m + 1$。

4)依据(3)式计算聚类准则函数G是否收敛,若收敛,算法结束,否则跳转至步骤1。

$$G = \sum_{i=1}^{N}\sum_{F \in Ai} | f - fa | \quad (3)$$

5)建立新的聚类B,聚类数目$n = n + 1$。

6)判断聚类数目n是否大于设定的聚类数目上限$n_{\max}$,若是,则删除聚类权重最小的$n - n_{\max}$个聚类,然后跳转至步骤1。

2.2 基于GFKM的集群入侵检测

本文提出的基于GFKM的集群入侵检测模型体系结构如图1所示,其中GFKM算法置于负载均衡装置中。

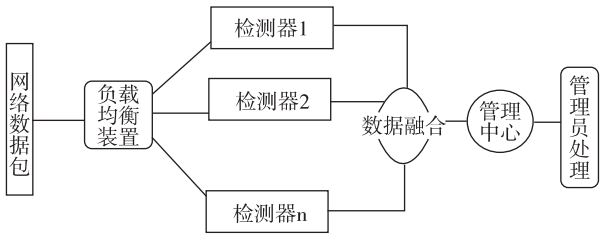

图1 GFKM集群入侵检测体系结构

3 仿真试验

本文采用KDD CUP1999数据集来验证GFKM集群入侵检测模型的有效性。KDD CUP1999数据集包括全部数据集、10%数据集和corrected.gz数据集,其中全部数据集和10%数据集都包括训练集和测试集,corrected.gz为含有攻击标记的测试数据集。数据以网络连接的形式保存,每条记录包括42个属性,包括41个固定属性和1个类标识。

根据实验需要选取10%数据集作为实验数据。实验前要对数据进行预处理,根据参考文献[5-6]中描述的方法对数据进行量化和标准化处理,然后,利用测试数据集验证论文GFKM集群入侵检测模型的有效性。试验结果如表3所示。

表3 试验检测结果 %

类型	正确率	误报率	漏报率
DOS	98.31	1.12	0.57
Probe	97.65	2.35.	0
U2R	99.76	0.1	0.14
R2L	98.80	0.9	0.3

Normal	100	0	0

同时,为了检测该入侵检测模型的性能,本文用基于传统 K 均值聚类算法的入侵检测模型进行对比实验,实验结果如图 2 所示。

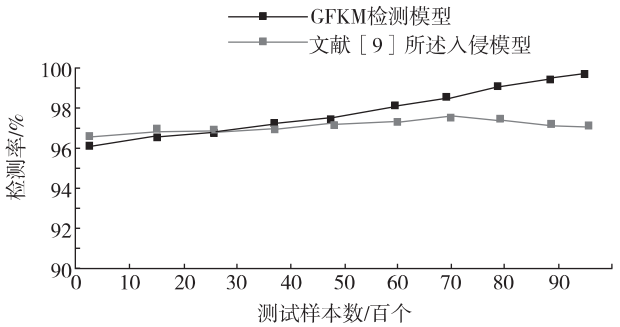

图 2 对比实验结果

根据表 3 的各项性能指标可以看出,GFKM 算法在进行入侵检测试验时,误报率为 0.31%,漏报率为 0.34%,可见该算法应用于入侵检测是可行且有效的。由图 2 对比试验结果可以看出,传统的基于聚类的入侵检测模型随着测试赝本数目的增加,模型的检测率逐渐降低,说明模型的泛化能力比较弱,相比较而言,基于 GFKM 的检测模型整体性能比较优越。

4 结束语

本文针对将聚类分析应用于入侵检测时忽略数据流关联度而导致的聚类结果不准确这一问题,提出了一种灰色 K 均值聚类算法,同时为了避免传统网络入侵检测系统由于无法匹配高速网络流量而导致的丢包漏报率高的缺陷,将 GFKM 应用于集群入侵检测系统中,从而提出了一种新的基于 GFKM 的集群入侵检测系统。最后,利用 KDD CUP1999 数据集进行仿真试验,试验结果验证了此入侵检测模型的正确性和有效性。

参考文献:

[1] 张宝军. 网络入侵检测若干技术研究[D]. 杭州:浙江大学,2010.
Zhang B J. Research on some techniques of network[D]. Hangzhou:Zhejiang University,2010.

[2] Han J W,Kamber M. 数据挖掘概念与技术[M]. 北京:机械工业出版社,2007:184-244.
Han J W,Kamber M. Data mining:concepts and techniques[M]. Beijing:China Machine Press,2007:184-244.

[3] Giacinto G,Roli F,Didaci L. Fusion of multiple classifiers for intrusion detection in computer networks[J]. Pattern Recognition,2003,24(12):1795-1803.

[4] 赵越,张为群. 一种基于 CFCM 的集群入侵检测方法研究[J]. 计算机科学,2010,36(6):176-178.
Zhao Y,Zhang W Q. Research based on a method of CFCM clustering intrusion detection[J]. Computer Science,2010,36(6):176-178.

[5] 肖轩. 灰色神经网络与支持向量机预测模型研究[D]. 武汉:武汉理工大学,2009.
Xiao X. A study on the predicting model of gray neutral network and support vector machine[D]. Wuhan:University of Technology,2009.

[6] Todd H L,Gihan D V,KarlLevitt N. A network security monitor[M]. Oakland:1990 IEEE Symposium on Security and Privacy,1991:296-304.

[7] 唐五湘. T 型关联度及其计算方法[J]. 数理统计与管理,1995,14(1):34-37.
Tang W X. The concept and the computation method of T's correlation degree[J]. Journal of Applied Statistics and Management,1995,14(1):34-37.

[8] 孙玉刚,党耀国. 灰色 T 型关联度的改进[J]. 系统工程理论与实践,2008,4(4):0135-0140.
Sun Y G,Tang Y G. Improvement on grey T's correlation degree[J]. Systems Engineering-Theory & Practice,2008,4(4):0135-0140.

[9] 谷保平,许孝元,郭红艳. 基于粒子群优化的 K 均值算法在网络入侵检测中的应用[J]. 计算机应用,2007,27(6):1368-1370.
Gu B P,Xu X Y,Guo H Y. Research of K-means algorithm based on particle swarm optimization in hetwerk intrusion detection[J]. Journal of Computer Application,2007,27(6):1368-1370.

Clustering Intrusion Detection Model Based on Grey Fuzzy *K*-mean Clustering

XU Yan-qun[1], ZHANG Bin[1], QIN Xiao-tie[2]

(1. Dept. of Computer Science and Technology, Nanyang Institute of Technology, Nanyang Henan 473004;
2. Library, Chongqing University of Science and Technology, Chongqing 401331, China)

Abstract: *K*-means clustering algorithm is applied to the field of intrusion detection, it has the following problems. First, cluster analysis does not consider the correlation degree of the data flow, so clustering accuracy is not high; the second is prone to packet loss omissions be- cause of data ove rload. In order to improve the clustering accuracy, this paper introduces grey analysis algorithm to improve the clustering accuracy. Me anwhile, in order to avoid the packet loss and underreporting phenomenon, we introduce clustering technology into the intru- sion dete ction system to process the load balancing of the data stream, thereby overcome the contradiction between the high-speed network data flow and low-speed intrusion detection system processing capabilities. After a comprehensive analysis, this paper proposed an intrusion detection model based on the gray *K*-means clustering algorithm.

Key words: clustering system; intrusion detection; *K*-mean clustering algorithm; cluster analysis

(责任编辑 欧红叶)