

改进 k-means 算法的网络数据库入侵检测

韩占柱

(呼伦贝尔学院 传媒学院, 内蒙古 海拉尔区 021008)

摘 要: 提出改进的 k-means 算法, 加入过滤优化功能, 通过簇候选集中攻击簇的数目优化, 删除掉非最优聚类数据集中的攻击数据, 生产最优簇, 提高后期网络数据库入侵检测的时效性, 降低漏检率. 实验结果表明, 本文的方法能够优化聚类后生成攻击簇的数目的数目, 为网络数据量入侵检测提供便利, 提高了检测的准确性, 降低了漏检率.

关键词: 网络数据库; 入侵检测; 改进算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)03-0144-03

Improved K-means Algorithm Network Database Intrusion Detection

HAN Zhan-zhu

(Hulunbeir College Media College, Inner mongolia Hailar 021008, China)

Abstract: This paper proposed the improvement k-means algorithm, join filter optimization function through the cluster set the number of candidates clusters, deleted the optimal cluster the data in the data set, the optimal cluster, improve production later network database intrusion detection, reduce the efficiency of the miss rate. The experiment results show that the method can optimize the clustering to create the number of clusters, for the number of network data quantity intrusion detection provides convenience, improve the detection accuracy, and to reduce the miss rate.

Key words: network database; intrusion detection; improved algorithm

1 引言

当前, 网络数据库被入侵的情况时有发生, 各种入侵方法层出不穷, 手段日益增多. 传统的网络数据库检测方法是基于滥用检测^[1]. 这种算法是通过把攻击过数据库的方式方法作为数据, 存储到知识库中, 使得如果下次有相同攻击进行网络数据库入侵时, 系统会迅速做出反应, 对其进行相应的阻拦. 这种方法存在一个缺陷, 就是在进行入侵方法识别的时候, 先要对已存的入侵方法进行分类, 分类结果的好坏不但关系着入侵识别的速度, 也关系着识别的准确度. 传统的分类方法会分类出太多的无用项, 导致识别时间过长, 检测结果不准确.

基于以上因素, 为了提高网络数据库入侵检测的准确度, 本文提出改进的 K 均值聚类算法, 该算

法能够克服传统算法的缺陷, 对入侵检测方法进行很好的分类, 形成最优分类结果, 方便下一步的入侵检测, 并且实现简单高效, 具有很强的实用价值.

2 网络数据库入侵检测原理

网络数据库入侵检测, 就是通过一定的手段检测那些非正常使用计算机系统的个体和滥用其权限的正常用户. 主要是通过滥用检测方法完成的. 通过记录攻击类型知识库, 当新攻击来临时, 在知识库中进行查找, 完成攻击判断^[2].

这种数据库入侵检测方法的优点是能检测一些过去未出现过的的攻击类型, 但是缺点是需要对知识库进行分类, 分类结果影响着检测结果, 如果分类结果不好, 检测的误检率较高. 因为现有的网络数据库入侵检测大都基于统计的方法, 这需要较多的经

验,分类算法的优势是能够从大量数据中提取对系统有用处的、事先未知的知识和规律,而不依赖经验利用此技术.实现网络数据库安全在国内外都是一种新尝试,将聚类方法应用在基于数据库的入侵检测系统中,可以从大量的观察数据中发现有助于检测的知识和规则通过长期的实现^[3-6].

但是在实际运用中,发现该方法在存在一定的问題,主要反映在聚类过程中:

(1) 多次分析聚类知识数据库需要很大的负载.根据聚类算法可知,每产生一个新的攻击类型,都需要对其所在的知识数据库进行一次全面的扫描来判断它是否是新的攻击项目集.如果要生成 k -频繁攻击项目集,则需要扫描 k 次知识数据库,来确定它是否应加入到公积项目集中.当攻击知识数据库比较庞大时,生成的候选集攻击数量相应也就越多,扫描数据库的次数也就越多,数据库负载也就越大,因此,其效率就会很低.

(2) 产生大量的无用攻击候选集.聚类函数利用之前攻击数据和 L_{k-1} 连接生成攻击候选集 C_k ,所以 C_k 是由大量的 k -项攻击集组成.随着 k 值的增加,相应的 k -项候选集的数量成指数增长,有些攻击是无用的,重复的,造成攻击候选集的数量庞大.

(3) 当知识数据库攻击项比较少时,网络数据库攻击检测算法具有较好的性能.当知识数据库中数据量较大时,在连接过程中会产生大量的候选攻击集,而在判断攻击集是否是频繁攻击集的时候,需要反复扫描知识数据库,因此算法效率会大大降低.

综合考虑网络数据库攻击检测算法各方面的缺点,以及在各种传统方法的研究基础上,本文从以上方面出发,提出一种改进 K 均值聚类算法的网络数据库攻击检测算法,通过对现存的知识库进行分类优化,大幅减少数据库中的冗余攻击数目,优化知识库中的攻击分类结果,以满足网络数据库后期检测的要求.

3 基于 K 均值聚类的网络数据优化分类

3.1 传统 K 均值算法原理

k-means 算法的优点是简单、快捷.在处理大型网络数据库知识聚类问题时比其他聚类算法更加快速.该算法是利用计算代价函数与引入惩罚机制的方法得到最合适的簇数目值,从而避免了上面陈述的缺陷.同时也避免了“噪声”和孤立点数据对结果的不利影响,并且获取了此算法的时间复杂度.但起也存在一定的缺点,其缺点如下:

(1) 根据经验确定的攻击聚类数据集合时需要生成簇的数目 k 不一定是合适的,而且初始值不同导致获取的聚类结果也不同,这影响了算法的性能.

(2) 因为需要首先确定攻击簇的平均值,所以不适用于聚类分类属性的网络攻击数据对象.

(3) 不适应于数据对象的数据集差异性较大的情况.

在网络数据库攻击数据集合中会存在一些与其他数据对象有较大差异的攻击数据对象,称为“噪声”和孤立点数据,这些数据对平均值有一定影响,最后导致聚类网络数据库攻击的性能受到影响.为此,本文提出改进 k-means 算法,首先提取 k 个密度最大的区域,然后从其中每个区域都选取一个点,称为簇中心,此点到其他簇中心代表点的距离是最远的,然后选取攻击数据对象的中心点,对孤立点的数据对象进行处理,从而提高聚类划分的准确率,避免了上述的缺陷,使算法的性能得到优化.

3.2 改进的 K 均值聚类算法检测攻击

改进后的算法需要确定一个攻击簇是否是动态簇或者静态簇,从而决定是否将其从候选的攻击集合中删除,最终减少了攻击簇候选集合中簇的数目.其具体流程如下:

(1) 对全部候选攻击数据集,构建一个适当的数据结构,并且对该数据结构中的所有结点进行计算.

(2) 假设 CC^0 是初始攻击簇中心集,设置 rt . $SC=CC^0$,根结点候选中心集用 rt . SC 表示.设置 C_i . $count=0$ 和 C_i . $total=0$.

(3) 对于除了根节点外的所有结点 n :设置 n . $SC=pn$. SC ,在表达式中结点 n 的父节点的候选中心集用 pn . SC 表示.通过 n . SC 确定 n . $center$ 的最邻近中心 C_i ,并对 r'_1 和 r'_2 进行计算,其中结点 n 单元集的中心用 n . $center$ 表示.假设 n 是叶子结点,设置 C_i . $count=C_i$. $count+n$. $count$, C_i . $total+=n$. $total$;假设 n 不是叶子结点,确定所有攻击簇中心 $C \in n$. $SC/\{C_i\}$ 的顶点 V ,假设 $d(V,C)>d(V,C_i)$,则将 C 从 n . SC 中删除.假设 $|n$. $SC|=1$,则设置 C_i . $count=C_i$. $count+n$. $count$ 和 C_i . $total=C_i$. $total+n$. $total$.

(4) 设置 $p=1$,可以获取新的攻击簇中心集 CC^p ,其中迭代的次数用 p 表示.

(5) 通过 CC^p 获取活动攻击簇的中心子集 CC_{active} ,设置 rt . $SC=CC^p$ 和 rt . $SC_{active}=CC_{active}$,其

中根结点活动中心的候选集是 rt, SC_{active} . 设置 $C_i, count=0$ 和 $C_i, total=0$.

(6) 对除了根节点外的所有结点 n : 设置 $n, SC = pn, SC$ 和 $n, SC_{active} = pn, SC_{active}$. 通过 n, SC 中的 $n, center$ 获取最邻近攻击簇中心, 对 r'_1 和 r'_2 进行计算. 假设 n 是叶子结点, 则设置 $C_i, count = C_i, count + n, count$ 和 $C_i, total = C_i, total + n, total$.

假设 $C_i \neq C'_i$ 且 $d(n, center, C_i) \geq r'_2$, 确定所有中心 $C \in n, SC / \{C_i\}$ 的顶点 V . 如果 $d(V, C) > d(V, C_i)$, 则将 C 从 n, SC 中删除. 假设 $|n, SC| = 1$, 设置 $C_i, count = C_i, count + n, count$ 和 $C_i, total = C_i, total + n, total$.

在 $C_i \neq C'_i$ 且 $d(n, center, C_i) < r'_2$ 或 $C_i = C'_i$ 的情况下, 通过 $C \in n, SC_{active}$ 或者 $(C \in n, SC_{static}$ 且 $C \notin n, SC')$ 获取顶点 V . 假设 $d(V, C) > d(V, C_i)$, 则将 C 从 n, SC 中删除. 假设 $|n, SC| = 1$, 则设置 $C_i, count = C_i, count + n, count$ 和 $C_i, total = C_i, total + n, total$.

(7) 设置 $p = p + 1$ 可以获取新的攻击簇中心集 CC^p . 跳到步骤(5)继续执行, 当攻击簇中心不发生变化时结束.

过滤攻击数据算法需要在迭代的各个阶段确定 n, SC 中 $n, center$ 最邻近攻击簇中心 C_i , 并且确定 $n, SC / \{C_i\}$ 中各个成员是否需要删除. 可以利用 n, SC_{active} 确定 $n, center$ 的最邻近攻击簇中心 C_i , 并且确定 n, SC_{active} 中各个攻击成员是否需要删除. 通过上面的方法可以在迭代的各个阶段, 将算法的计算复杂度减小约 $\sum |n, SC| / \sum |n, SC_{active}|$ 倍. 其中 n, SC 是通过将攻击簇中心集合 CC 删除一些成员获取的, n, SC_{active} 是通过将动态攻击簇中心集 CC_{active} 删除一些成员获取的. $|n, SC|$ 和 $|CC|$ 是成比例的, 并且 $|n, SC_{active}|$ 和 $|CC_{active}|$ 成比例, 即 $(\sum |n, SC| / \sum |n, SC_{active}|) \approx (|CC| / |CC_{active}|)$.

通过以上的聚类算法, 能够得到优化之后的网络数据库攻击的先验分类, 利用优化后的分类结果, 可以更好的实施抗攻击检测, 提高检查的准确度.

4 实验结果及其分析

为了验证算法的优化效果, 本文选取网络数据库, 进行攻击实仿真实验, 仿真的攻击分为伪装攻击和合法攻击.

检测准确率 = 检测出的入侵事件个数 / 总的入

侵事件个数 $\times 100\%$

错误率 = 被误检为异常的事件个数 / 总的事件个数 $\times 100\%$

漏报率 = 漏报的入侵事件个数 / 总的事件个数 $\times 100\%$

统计结果如下表 1 所示.

表 1 实验数据统计

	检测准确率	错误率	漏报率
本文算法	89	7%	4%
传统算法	67%	17%	16%

首先, 分析两种算法在不同的攻击事务数下, 它们统计差异, 这里取定攻击事务是频繁项目集, 选取 400 个到 2000 个不同的攻击事务数据, 用传统算法和改进的算法分类其中的频繁项目集. 统计趋势如图 1 所示.

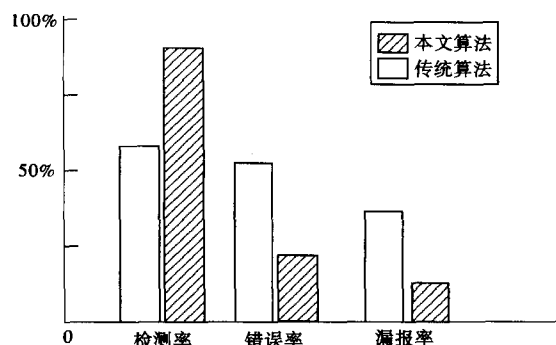


图 1 事务数不同时统计对比

通过以上统计可以看出, 经过优化分类后的网络数据库入侵检测取得的效果要明显好于优化之前. 因此本文算法有效.

5 结束语

本文的主要内容是对网络数据库的入侵检测方法进行研究, 先详细描述了入侵检测算法的原理, 再结合原理分析出其存在的缺点, 然后提出一种基于 K 均值聚类的改进算法, 改进的算法是通过减少网络知识库中攻击的种类, 以及在生成候选攻击集的数量, 然后连接产生新的候选集. 便于后期的数据检测. 最后实验也证实了改进的算法的优越性.

参考文献:

- [1] Krista Rizman, Zalik. An efficient K-means clustering algorithm [J]. Pattern Recognition Letters, 2008(29): 1385-1391.

(下转第 150 页)

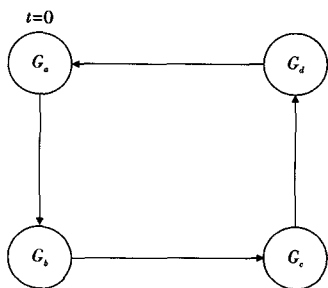


图4 拓扑结构切换规则

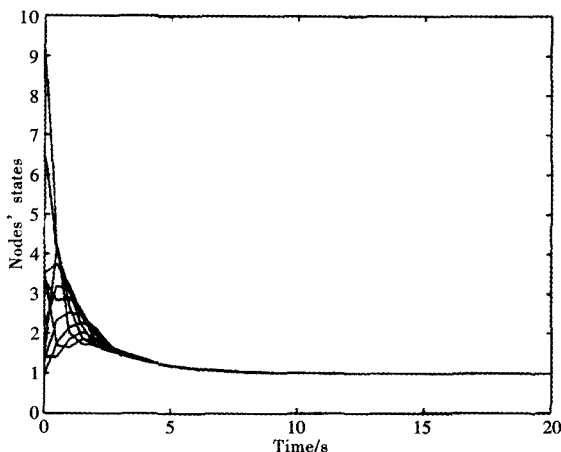


图5 可变拓扑结构下系统的曲线图

5 结束语

目前,关于多系统的分布式控制系统一致性问题的研究,大都基于单个信号,线性耦合情况为基础的.但是实际的系统,根据需求,并不能保证都是线性的.尽管这些系统可以用一个线性模型来近似的表示,但在很多种情况下,这种近似计算还存在着较

大的误差.

本文研究了非线性个体之间存在的一致性问题的,以及这种困难带来的问题;运用带非线性耦合的一致性,考虑了时不变拓扑结构以及时变拓扑结构的不同情况下的对应关系.对于前者,用紧缩分析来处理系统的非线性;对于后者,用多函数的思想来处理时变的拓扑结构.结果表明:在一些假设前提下,可以通过控制一个节点,来满足系统的一致性需求.

参考文献:

- [1] Watts DJ. Small worlds; the dynamics of networks between order and randomness [M]. Princeton: Princeton University Press, 1999;
- [2] Hernandez-Valencia EJ. Architectures for broadband residential IP services over CATV networks [J]. IEEE Network, 1997(11): 36-43.
- [3] Jasný B R, Ray L B. Life and the art of networks [J]. Science, 2003(301): 1863-1863.
- [4] Milgram S. The small world problem [J]. Psychology Today, 1967(2): 60-67.
- [5] Guare J. Six degrees of separation; a play [M]. New York: Vintage, 1990.
- [6] 徐淑萍,苏小会. 基于 Internet 的远程控制系统网络时延分析研究[J]. 微电子学与计算机, 2011(07): 49-50.
- [7] 胡晓娅,朱德森,汪秉文. 网络控制系统仿真模型的网络数据传输[J], 计算机仿真, 2005(11): 93-94.

作者简介:

田 屏 女, (1974—), 硕士, 讲师. 研究方向为计算机软件与理论.

刘彦宾 男, (1966—), 研究生. 教授研究方向为信息安全.

(上接第 146 页)

- [2] He D, Goker A. Detecting session boundaries from Web user logs [C]// Proceedings of the 22nd Annual Colloquium of IR Research. Cambridge: [s. n.], 2005.
- [3] 陶剑文. 一种分布式 Web 日志挖掘系统的设计与实现 [J]. 计算机仿真, 2006(10): 55-56.
- [4] Cooley R, Mobasher B, Srivastava J. Grouping web page references into transactions for mining world Wide web browsing patterns [C]// proceedings of the 1997 IEEE knowledge and Data Engineering Exchange Workshop. Washinytion, DC, USA: IEEE Computer Society, 1997.

- [5] Zhang Bofeng, Su Jinshu, Xu Xin. A class-incremental learning method for multiclass support vector machines in text classification [C]// Proceeding of the Fifth International conference on Machine Learning and Cybernetics. Dalian: IEEE, 2006.
- [6] He D, Goker A. Detecting session boundaries from web user logs [C]// Proceedings of the 22nd Annual Colloquium of IR Research. Cambridge: [s. n.], 2005.

作者简介:

韩占柱 男, (1961—), 副教授. 研究方向为计算机科学与技术、计算机网络与技术、数字媒体技术、流媒体技术.