

文章编号:1001-9081(2010)07-1916-03

高效的混合聚类算法及其在异常检测中的应用

李建国¹, 胡学钢²

(1. 淮北师范大学 计算机科学与技术学院, 安徽 淮北 235000; 2. 合肥工业大学 计算机与信息学院, 合肥 230009)

(ljg002002@126.com)

摘要:将聚类算法应用于异常检测,算法的有效性是关键。为了提高异常检测能力,提出了一种新的聚类算法,该算法运用窗口管理机制对网络数据采用分批实时处理,同时对算法中运用到的 DBSCAN 算法和 K-means 算法进行改进并组合。实验证明该算法可以提高异常检测的检测率,降低误报率,并提高系统的实时响应能力。

关键词:入侵检测;异常检测;聚类分析;K-means 算法;DBSCAN 算法

中图分类号: TP393.08 **文献标志码:** A

Efficient mixed clustering algorithm and its application in anomaly detection

LI Jian-guo¹, HU Xue-gang²

(1. School of Computer Science and Technology, Huaibei Normal University, Huaibei Anhui 235000, China;

2. School of Computer and Information, Hefei University of Technology, Hefei Anhui 230009, China)

Abstract: High efficiency of the clustering algorithm is the key when it is applied to anomaly detection. In order to improve the anomaly detection, this paper advanced a new clustering algorithm that deals with the network data partially by real-time processing, improving and integrating the DBSCAN algorithm and K-means algorithm. It is proved by experiments that the new algorithm can improve the detection rate, reduce the false positive rate, and enhance the real-time responding ability of the system.

Key words: intrusion detection; anomaly detection; cluster analysis; K-means algorithm; Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm

0 引言

随着当今计算机网络技术的快速发展,很多信息通过网络来进行传输与存储,网络安全成为倍受关注的研究领域。入侵检测(Intrusion Detection)技术^[1]是继“防火墙”、“数据加密”等传统安全保护措施之后,一种很重要的主动防御手段,异常检测^[2-3]是入侵检测的分类之一,通过分析异常行为和使用计算机资源的情况来检测出入侵行为,可以发现未知的攻击行为。聚类^[4]可以在没有任何先验知识的前提下检测出未知的攻击行为,同时根据聚类的特点分析,该技术比较适合高速网络中对大量网络数据的处理。

目前,通过对各种聚类算法的应用与改进,将聚类分析技术应用到异常检测系统,以提高对网络未知攻击的检测水平,已成为入侵检测领域中研究的一大热点。文献[5]提出了一种基于信息论的概率聚类算法(EMicro),给出了一种在多数数据流情况下的异常趋势检测方法;文献[6]提出了基于K-means的数据流聚类算法,利用分批处理的办法通过反复迭代的过程,使在有限的内存空间内对高速网络数据流的聚类成为可能;文献[7]提出了基于层次聚类的模糊聚类算法HFC,证明了该算法在入侵检测系统中的适用性;文献[8]提到将一种快速的基于高密度连接区域聚类(Density-Based Spatial Clustering of Application with Noise, DBSCAN)算法应用到异常检测中,该算法通过删除核心点,降低了算法的执行时间;文献[9]提出用聚类技术与关联规则技术进行联合挖掘的算法,较好地解决了数值属性的分类问题;文献[10]提出

一种二次聚类的异常检测方法,该方法可以存储大量的原始网络数据,并利用了最能反映当前网络行为的统计信息检测入侵行为;文献[11]提出基于概率密度的在线数据流聚类算法;文献[12]提出基于密度的进化数据流聚类算法,实现了对任意形状的数据流聚类,可以将潜在的离群点以及簇区分开;文献[13]提出了一种基于模糊连通度的聚类算法,用于检测已知攻击及其变种等;文献[14]提出了 IF-DBSCAN 算法,算法通过选取核心对象领域中的代表对象来扩展类,减少了查询次数,提高了算法的性能。

由于当前聚类算法大多只适用于具有特定分布的数据受噪声、孤立点的影响较大,并且通常对包含全部历史数据的整个数据集进行等同学习,因此难以真实描述网络数据随时间实时变化的特点,同时检测结果受历史数据的影响也较大。另外,网络中传输的数据往往具有属性较多、高速、连续到达等特点,所以有限的内存空间、时间复杂度和检测率等成为入侵检测算法面临的主要问题。本文结合当前网络异常检测的要求,及以往算法的不足,提出了一种高效的混合聚类算法(Efficient Mixed Clustering Algorithm, EMCLA),以提高异常检测系统的性能。

1 高效的混合聚类算法

EMCLA 依据滑动窗口管理机制^[15]将从网络中采集到并经过预处理的数据分批存储到内存中,采用 DBSCAN 算法对每个窗口的数据进行处理,根据基于密度聚类的基本思想,大部分聚类数据为正常数据,将聚集成一个高密度簇数据集,对

收稿日期:2010-03-18;修回日期:2010-05-09。 基金项目:安徽省高校自然科学基金项目(KJ2008B125)。

作者简介:李建国(1975-),男,安徽砀山人,讲师,硕士,主要研究方向:网络安全、数据挖掘; 胡学钢(1961-),男,安徽当涂人,教授,主要研究方向:知识工程、数据挖掘。

于异常数据不仅数量少而且差异大,这部分数据将聚集成一个低密度簇数据集。因此,根据异常检测的特点,为了避免正常数据重新参与聚类而耗费系统时间,可以在首次聚类过程中将产生的高密度簇数据删除,并保存每个低密度簇的均值点,然后以所有的均值点为初始聚类中心,利用优化的K-means算法对剩余的低密度簇数据进一步聚类,最终确定异常聚类。

假定网络中传输的数据包以队列 $M_1, M_2, \dots, M_i, \dots, M_n$ ($i = 1, 2, \dots$) 的形式被存储到内存中,其中,每个队列中包含 X 条连接记录。那么对于在 t 时刻采集到的数据队列 M_i ,用DBSCAN算法进行处理,将得到 Y_i 个低密度簇聚类,其均值点用 O_1, O_2, \dots, O_i 表示,按照窗口管理机制保存每次获得簇均值点,每个窗口保存 $W = X$ 个簇均值点。

EMCLA 算法描述如下。

输入:队列中连接记录数 X ;邻域半径 e_{ps} ,核心点至少包含的点数 $s_{minipits}$; K-means 聚类参数 $Max-distance$ 。

输出: K 个异常聚类。

- 1) $i = 1$ i 代表窗口标号
- 2) do while $i \leq$ 窗口总数
- 3) 导入 1 个队列
- 4) 用基于密度的 DBSCAN 算法对该队列聚类
- 5) 删除高密度簇聚类中的连接记录,计算出低密度簇的均值点并保存在 i 窗口
- 6) $i = i + 1$
- 7) end do
- 8) 以所有的低密度簇均值点为初始聚类中心,对所有的低密度簇数据进行聚类;
- 9) 输出异常聚类

在 EMCLA 算法中,首先针对异常检测的特点和基于密度聚类的基本思想,在对窗口中的数据运用 DBSCAN 算法进行聚类时,考虑到大批正常数据对系统资源的占用,所以从中删除了一部分高密度簇数据,使得这部分数据不再参与下一轮的聚类操作,进而减少了系统对冗余数据的执行时间,提高了算法的效率。

另外,由于 K-means 算法本身存在着对初值比较敏感、需要预先确定聚类个数以及在聚类过程中容易产生空聚类的三方面的缺点,本文给出了优化措施:1) 设置一个聚类参数 $Max-distance$,然后分别求出当前记录到所有聚类中心的距离,当此值大于聚类参数时,就把当前记录作为一个新聚类的聚类中心,这样最终的聚类个数可以根据具体情况而自动调整,同时还可以将异常记录划分到新的聚类当中,从而得到较好的聚类效果;2) 对于在聚类结果中产生空聚类的情况,采取把离聚类中心距离最远的记录排除出该记录所属聚类的措施,以此记录作为新聚类的聚类中心,用新聚类取代空聚类;3) 对于初始聚类中心选取的问题,解决的方法是先对数据队列分批使用 DBSCAN 算法进行聚类,这样可以避免“孤立点”的影响,提高初始聚类中心的代表性。

优化后的 K-means 算法描述如下:

- 1) 在原始数据集中选出 m 个最优的初始聚类中心 $\{w_1, w_2, \dots, w_m\}$, 其中 $w_j = x_i, j \in \{1, 2, \dots, k\}, i \in \{1, 2, \dots, n\}$;
- 2) 让聚类 C_i 与 w_j 对应;
- 3) 求出另外的记录 $x_i (i \in \{1, 2, \dots, n\})$ 到聚类中心距离的最小值,记为 $Min-distance$;
- 4) 如果 $Min-distance < Max-distance$,就把 x_i 归到最近的 w_{j*} 所属的聚类 C_{j*} 。即 $|x_i - w_{j*}| \leq |x_i - w_j|, j \in \{1, 2, \dots,$

$n\}$;反之将产生一个用 x_i 作为聚类中心的新聚类;

5) 转到 2), 遍历所有记录;

6) 计算每个聚类中所有记录的平均值,并以此值来代替前一次的聚类中心,如式(1)所示:

$$w_j = \frac{\sum_{x_i \in C_j} x_i}{|C_j|} \quad (1)$$

7) 当有空聚类的情况发生时,就把离聚类中心距离最远的记录剔除出该记录所属的聚类,并以此记录作为新聚类的聚类中心,同时用新聚类取代空聚类;

8) 当聚类中心的值不再发生改变时结束,并输出新的 k 个聚类。

K-means 算法中, n 代表数据库中数据的个数, M 代表初始聚类个数, $Max-distance$ 代表聚类参数。

2 算法分析

设队列中连接记录的个数为 x , 低密度聚类簇数为 k , p 为循环次数, z 为得到 w 个簇均值点需处理的队列数。由于 EMCLA 分两步进行,所以时间开销由两部分组成。在第一阶段(对队列分批聚类阶段),对每个队列的 x 个连接记录进行 DBSCAN 聚类生成密度簇均值点,时间复杂度为 $O(x^2)$,对 w 个均值点用优化的 K-means 算法聚类,时间复杂度为 $O(xkp)$,因此总的时间复杂度为 $O(x^2z + xkp)$ 。在实际应用中队列中的连接记录个数 x 较小,所以总的时间花费不多。在对 w 个簇均值点进行 K-means 优化聚类时,由于该算法具备高效、简单的优点,且通常 $k \ll x$,故该阶段时间花费较小。对于算法的空间复杂度,由于在算法设计中采用了窗口管理机制,使得 EMCLA 可以在有限的内存空间中实现对队列的聚类。

另外,该算法还有 3 个方面的优点:1) 针对目前网络攻击变化多端、流量较大、网速较快的情况,该算法使得系统不再依赖庞大的历史数据库就可以实时地发现异常攻击,通过对 DBSCAN 算法的改进,不仅减少了入侵检测的时间,还较容易发现新的攻击;2) 通过对 K-means 算法的优化,克服了原有 K-means 需要预先确定最终聚类个数、初始聚类中心选择难以及容易产生空聚类等缺点,进一步保证了聚类的质量;3) 通过分批实现对网络数据的聚类,以及双重聚类的有效结合,使得异常数据很难逃过检测,消除了噪声和孤立点的影响,在提高系统的入侵检测效率的同时也降低了误报率。

3 实验分析

实验硬件环境基本配置:CPU 为 Intel Pentium Dual-Core E5200,内存为 2 GB,软件环境基本配置:操作系统采用 Windows 2000,编程语言采用 VC++ 6.0。

实验数据采用 KDD Cup 99 数据集^[16],该数据集来自 MIT 的 Lincoln 实验室,其中共收集了局域网中 494 020 条连接记录,实验共进行三次。每次从数据集中随机抽取包含一定数目攻击记录的连接记录组成数据集,分别包含 5 000 条、7 000 条、10 000 条。从每条记录的 41 个属性中选取其中具有入侵检测价值的 31 个属性作为测试样本。

实验假设在记录总数为 N 的数据集中,异常记录的比例为 Q ,当聚类中对象数目小于 QN 时,将其视为异常聚类,同时该聚类中的所有对象都被认为是异常的。在本实验当中采用的 Q 值为 1.5%,队列中连接记录个数为 200;邻域半径 e_{ps} 为 0.5,核心点至少包含的点数 $s_{minipits}$ 为 7;K-means 聚类参数

Max-distance 为 15。分别用 DBSCAN 算法、K-means 算法和 EMCLA 算法对以上三个数据集进行聚类,分别从入侵检测率、误报率以及算法执行时间三个方面进行比较,如表 1~3 所示。

表 1 三种算法的异常检测检测率对比 %

算法	数据集		
	Data1	Data2	Data3
DBSCAN	90.80	91.70	90.20
K-means	92.30	91.50	90.60
EMCLA	97.80	98.10	98.70

表 2 三种算法的异常检测误报率对比 %

算法	数据集		
	Data1	Data2	Data3
DBSCAN	11.02	10.70	10.20
K-means	11.30	9.50	10.60
EMCLA	1.80	1.10	1.06

表 3 三种算法的执行时间对比 s

算法	数据集		
	Data1	Data2	Data3
DBSCAN	0.8	1.3	1.6
K-means	0.9	1.7	1.8
EMCLA	1.6	2.3	2.7

通过以上实验结果可以看出,EMCLA 由于采用了分批处理、双重聚类的思想在入侵检测率方面取得了非常好的效果,误报率也保持在较低的水平。另外由于算法首先采用了改进的 DBSCAN 算法分批处理,然后又采用了改进的 K-means 算法对低密度数据进行聚类,在算法执行时间上虽然有所增加,但是经过前期的分批处理后,有很多安全数据已被过滤,这部分时间增加不多,相比在系统安全性更重,不影响整体性能的情况下,这部分时间的增加是可以接受的。

4 结语

本文所提出的 EMCLA,采用了滑动窗口管理机制分批处理的思想,将网络中的传输数据分批处理,通过对两种聚类算法的改进及有效结合,提高了系统的入侵检测率,降低了误报率;同时解决了原有聚类算法中聚类结果与初始聚类中心的个数联系紧密的问题,消除了空聚类与孤立点对聚类结果的影响,使得该算法具有高效处理大批实时数据的能力和适应网络正常行为变化的能力,可以执行异常检测,并通过实验证明了该算法的优越性和有效性。进一步优化该算法以及与其他入侵检测技术结合起来提高系统的检测性能,是下一步研究的目标和方向。

参考文献:

- [1] PATCHA A, PARKA J M. An overview of anomaly detection techniques existing solutions and latest technological trends[J]. *Computer Networks*, 2007, 51(12): 3448-3470.
- [2] SHARMA A, PUJARI A K, PALIWAL K K, *et al.* Intrusion detection using techniques with a kernel based similarity measure [J]. *Computer and Security*, 2007, 26(7): 488-495.
- [3] 赵阔,胡亮,李博,等. 基于 CASL 的入侵检测系统测试[J]. *吉林大学学报: 信息科学版*, 2005, 23(1): 50-58.
- [4] LUHR S, LAZARESCU M. Incremental clustering of dynamic data streams using connectivity based representative points [J]. *Data & Knowledge Engineering*, 2009, 36(1): 43-48.
- [5] 张晨. 数据流聚类分析与异常检测算法[D]. 上海: 复旦大学, 2009.
- [6] GUHA S, MISHRA N, MOTWANI R, *et al.* Clustering data streams[EB/OL]. [2007-03-20]. <http://citeseer.ist.psu.edu/guha00clustering.html>.
- [7] 李玲娟. 数据挖掘技术在入侵检测系统中的应用研究[D]. 苏州: 苏州大学, 2008.
- [8] 吴贞珍,黄建华. DBSCAN 聚类算法在异常检测中的应用[J]. *计算机安全*, 2007(8): 43-46.
- [9] ZANERO S, SACRARESI S M. Unsupervised learning techniques for an intrusion detection system [C]// *Proceedings of the 2004 ACM Symposium on Applied Computing*. New York: ACM, 2004: 203-209.
- [10] 俞研,郭山清,黄皓. 基于数据流的异常入侵检测[J]. *计算机科学*, 2007, 34(5): 66-71.
- [11] SONG MINGZHOU, WANG HONGBIN. Highly efficient incremental estimation of Gaussian mixture models for online data stream clustering [C]// *Proceedings of Intelligent Computing Theory and Applications III*. Bellingham: SPIE, 2005: 174-183.
- [12] CAO F, ESTER M, QIAN W, *et al.* Density-based clustering over an evolving data stream with noise [EB/OL]. [2008-01-10]. <http://www.siam.org/meetings/sdm06/proceedings/030caof.pdf>.
- [13] WANG Q, MEGALOOKONOMO V. A clustering algorithm for intrusion detection [C]// *Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2005*. Orlando, Florida, USA: SPIE, 2005: 283-289.
- [14] 王桂芝,王广亮. 改进的快速 DBSCAN 算法[J]. *计算机应用*, 2009, 29(9): 2505-2508.
- [15] BABCOCK B, DATAR M, MOTWANI R, *et al.* Maintaining variance and k-medians over data stream windows [C]// *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York: ACM, 2003: 234-243.
- [16] STOLFO S J, FAN WEI, LEE W K. KDD-CUP-99 TASK DESCRIPTION [EB/OL]. [1999-06-19]. <http://kdd.ics.uci.edu/databases/kddcup99/task.html>.

(上接第 1915 页)

参考文献:

- [1] 李静宜. 协同采购[J]. *企业管理*, 2009(4): 92-93.
- [2] 鲁奎,杨昌辉,戴道明. 运输能力受限与费用时变批量问题的拉格朗日松弛启发式算法[J]. *系统工程理论与实践*, 2008, 10(10): 47-52.
- [3] 赵耀华. 供应链上一种供方——需方综合采购模型[J]. *南京理工大学学报*, 2002, 26(1): 105-108.
- [4] 向晋乾,黄培清,李杰. 定期订货策略下企业集团集中采购的最优订货模型[J]. *上海交通大学学报*, 2005, 39(3): 474-478.
- [5] 孙晓林,仲德强,满大庆,等. Purchasing policy model based on components/ parts unification[J]. *Journal of Southeast University: English Edition*, 2003, 19(2): 168-173.
- [6] 胡耀光,范玉顺,王田苗. 基于二层规划模型的统一采购方法研究[J]. *中国机械工程*, 2007, 18(1): 52-55.
- [7] ROY R N, GUIN K K. A proposed model of JIT purchasing in an integrated steel plant[J]. *Production Economics*, 1999, 59(1/3): 179-187.
- [8] GAO ZHEN, TANG LIXIN. A multi-objective model for purchasing of bulk raw materials of a large-scale integrated steel plant[J]. *Production Economics*, 2003, 83(3): 325-334.