一种无监督异常入侵检测的簇异常度量方法

陈 剑[1],蔡龙征[2]
(1. 广东科学技术职业学院 计算机工程技术学院,广东 珠海 519090;
2. 中南民族大学 工商学院,湖北 武汉 430065)

摘 要:文中主要研究用 Pearson 相关系数计算记录与簇、簇与簇间符号属性距离的方法;在这个方法中,提出了一种新的簇异常度量—近似平均距离 AAD, AAD 综合了一个簇的局部异常度,即簇的内部点密度,和该簇在整个簇结构中的全局异常度,即该簇与其它簇的距离;提出了依据 AAD 对聚类后的簇分类,并以已分类簇结构作为检测模型进行无监督异常检测的方法,通过异常检测能及时地对每个记录分类,从而能及时发现入侵行为,减小由入侵造成的损失;最后用 KDD 99 评估数据集所作的实验表明,用 AAD 作为簇的分类度量的方法比其它相关研究具有更高的检测率和更低的误警率。

关键词:无监督异常检测;入侵检测;网络安全;聚类
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2013)04-0131-04
doi:10.3969/j.issn.1673-629X.2013.04.032

A Cluster Anomaly Measure Approach for Unsupervised Anomaly Intrusion Detection

CHEN Jian[1], CAI Long-zheng[2]
(1. Computer Engineering Technical College, Guangdong Institute of Science and Technology, Zhuhai 519090, China;
2. Engineering and Commerce College, South-central University for Nationalities, Wuhan 430065, China)

Abstract: Mainly study the method of Pearson correlation coefficient to calculate the symbol attribute distance between record and cluster, cluster and cluster. A new metric, Approximate Average Distance (AAD), is proposed as cluster anomaly measure. AAD combines a cluster's local anomaly, the number of members, and its global anomaly, the distance with other clusters. An approach of unsupervised anomaly intrusion detection is also studied, in which records are checked with the classified clusters as detection models. To timely find intrusion behavior, reduce the loss caused by the invasion. Empirical experiments with the KDD 99 data set show that AAD can detect intrusions with relatively high detection rate and low false alarm rate compared with other researches.

Key words: unsupervised anomaly detection; intrusion detection; network security; clustering

0 引 言

无监督异常入侵检测所要解决的问题是:现有大量的过去记录,要区分这些记录哪些是正常记录,哪些是攻击。无监督异常检测的方法很多,其中广泛使用的一种是聚类分析法[1,2]。用聚类算法将数据集划分为簇后,需要对这些簇分类,确定哪些簇包含的是攻击记录,哪些簇包含的是正常记录。在半径相同的情况下,可以用簇的成员数量代表其点密度,因此,比较常用的簇异常度量是每个簇的成员数量。成员数量少的簇点比较稀疏,所包含的记录更有可能是攻击。相反,成员多的簇点比较密集,所包含的记录更有可能是正常网络连接。与簇成员数比较类似的簇异常度量是与簇的中心点距离不大于聚类阈值 w 的点数,表示为簇的记数 n。从概念上说, n 比簇的成员数能更准确地反映簇的点密度。簇的成员数是所有与该簇最近且与簇的中心点距离不大于聚类阈值 w 的点数。实际上,无论是簇的成员数还是簇的记数 n 都是一个簇内部的点密度,是一个簇异常性的局部性度量,而簇与其它簇的距离反映的是在整个簇结构中一个簇与其它簇的位置关系,是簇异常性的全局度量。簇的异常性应将两种度量结合起来。为此,文中提出新的簇异常度量:近似平均距离 AAD (Approximate Average Distance)。AAD 综合了簇的局部异常性和全局异常性。

1 簇异常度计算和簇分类

用基于距离的聚类算法将数据集划分为簇后,需

收稿日期:2012-08-08;**修回日期**:2012-11-11
基金项目:广东省科技计划项目(2010B060100056)
作者简介:陈 剑(1965-),男,江西吉安人,副教授,硕士,主要研究方向为数据通信、网络安全。

要对这些簇分类,确定哪些簇所包含的数据记录是攻击记录,哪些簇所包含的记录是正常记录。分类的依据是每个簇的异常度。异常度大的簇是由攻击记录组成的簇,异常度小的簇是由正常记录组成的簇。比较常用的簇异常度量是每个簇的成员数量,比如文献[3~6]就采用该度量。另一种常用的簇异常度量是与簇中心距离不大于聚类阈值 w 的点数 n [7~9]。

定量地分析簇的异常度,需要定义一个定量的计算方法,该方法要平衡地使用簇的成员数量和簇间的距离所包含的有关簇的异常度信息。

设网络连接记录集合 $E=\{E_1,E_2,\cdots,E_N\}$,共 N 个记录。聚类后被分成 $C_1,C_2,\cdots,C_k$ 共 k 个簇。可以近似地认为一个簇内的所有点的性质是相同的,因此,一个簇内的所有点的异常度也是相同的,可以用簇内的一个点的异常度来表示该簇的异常度。一个点的异常度可以用该点到所有其它点的平均距离表示。平均距离越大,说明该点与其它点的距离越远,因此该点的异常度越大。若用 $d(E_i,E_j)$ 表示属性空间中点 E_i、E_j 间的距离,则点 E_i 异常度量 AF_i(Anomaly Factor)用它与所有其它点的平均距离表示为:

$$\mathrm{AF}_i=\frac{1}{N-1}\sum_{\substack{j=1\sim N\\ j\neq i}} d(E_i,E_j) \tag{1}$$

由此,簇 C_i 的异常度用它的一个点与所有其它点的近似平均距离 AAD 表示为:

$$\mathrm{AAD}_i=\frac{1}{N-1}\times\sum_{\substack{j=1\sim k\\ j\neq i}} |C_j|\times d(C_i,C_j) \tag{2}$$

表达式(2)只考虑了簇间的距离因素,决定簇异常度的另一个重要因素是它的成员数量,簇的异常度与它所包含的成员数成反比,因此应除以簇的成员数。同时,簇分类的依据是各簇间的相对异常程度,与异常度的绝对大小没有关系,因此可以省略掉表达式(2)中对所有簇都相同的系数 $\frac{1}{N-1}$。最后得到的簇 C_i 的异常度(仍称作近似平均距离 AAD)为:

$$\mathrm{AAD}_i=\frac{1}{|C_i|}\sum_{\substack{j=1\sim k\\ j\neq i}} |C_j|\times d(C_i,C_j) \tag{3}$$

表达式(3)中,簇 C_i 的异常度与它和其它簇的距离 $d(C_i,C_j)$ 成正比。与它本身的成员数 $|C_i|$ 成反比。

簇分类方法是:

第一步:按表达式(3)计算所有簇的异常度;

第二步:按异常度由大到小排列簇 $\{C_1,C_2,\cdots,C_k\}$,得到簇序列 $\{C_{(1)},C_{(2)},\cdots,C_{(k)}\}$,满足 $\mathrm{AAD}_{(1)}\geqslant\mathrm{AAD}_{(2)}\geqslant\cdots\geqslant\mathrm{AAD}_{(K)}$;

第三步:找到满足 $\frac{\sum_{i=1}^{m}|C_{(m)}|}{N}\geqslant\tau$($\tau$ 是给定的阈值)的最小 m。$C_{(1)},C_{(2)},\cdots,C_{(m)}$ 是由攻击记录组成的簇,其它簇是由正常记录组成的簇。

2 簇表示法

在固定宽度聚类算法[7]中,簇是用中心点表示的,而中心点是该簇的某个记录所对应的点,也就是说簇的中心点是数据记录集合中真实记录在属性空间的对应点。计算点到簇的距离实际上是计算两点间的距离。实际上,一个真实记录所对应的点只是簇的近似中心点。在极端情况下,该点与簇的真正中心的距离可能达到聚类阈值的一半,即 $w/2$。

一个包含 p 个数字属性和 m 个符号属性的结构来描述聚类中的簇,每个属性都是簇中所有点该属性的统计值。称这种结构为 CSI(Cluster Summary Information,簇概要信息)。簇 C_j 的 CSI 表示为:

$$\mathrm{CSI}_j=(cx_1^j,cx_2^j,\cdots,cx_p^j,cy_1^j,cy_2^j,\cdots,cy_m^j) \tag{4}$$

表达式(4)中,$cx_k^j(k\in\{1,2,\cdots,p\})$ 是簇 C_j 中所有点的第 k 个数字属性值的算术平均值。$cy_1^j,cy_2^j,\cdots,cy_m^j$ 是簇的 m 个符号属性的值。每个符号属性的值用簇中所有的点该属性各类型值的频率(也就是各类型值的出现次数)组成的一个向量表示。

设符号属性 Y_k 共有 n_k 种类型值,则簇 CSI_j 的符号属性 $cy_k^j(k\in\{1,2,\cdots,m\})$ 是一个 n_k 维的向量,该向量的每个分量是簇 C_j 中所有点符号属性 Y_k 各类型值的出现次数。cy_k^j 用向量的形式表示为:

$$cy_k^j=(cy_{k1}^j,cy_{k2}^j,\cdots,cy_{kn}^j) \tag{5}$$

其中 cy_{k1}^j 表示簇 C_j 中所有点符号属性 Y_k 取第 1 个类型值的次数。

3 簇与簇间及点与簇间的距离函数

距离函数的定义应保证数字属性和符号属性对总距离的贡献是均衡的。用类似于欧氏距离的表示法将簇与簇间的距离表示为所有数字属性和符号属性间距离平方和的开方。

设簇 C_i 的 CSI 为:

$\mathrm{CSI}_i=(cx_1^i,cx_2^i,\cdots,cx_p^i,cy_1^i,cy_2^i,\cdots,cy_m^i)$

簇 C_j 的 CSI 为:

$\mathrm{CSI}_j=(cx_1^j,cx_2^j,\cdots,cx_p^j,cy_1^j,cy_2^j,\cdots,cy_m^j)$

则簇 C_i 与 C_j 间的距离表示为:

$$d(C_i,C_j)=\sqrt{\sum_{k=1}^{p}\|cx_k^i-cx_k^j\|^2+\sum_{k=1}^{m}\|cy_k^i-cy_k^j\|^2} \tag{6}$$

设点 $E_i=(x_1^i,x_2^i,\cdots,x_p^i,y_1^i,y_2^i,\cdots,y_m^i)$,则点 E_i 与簇 C_j 间的距离表示为:

$d(E_i,C_j)=$

$$\sqrt{\sum_{k=1}^{p} \| x_k^i - cx_k^j \|^2 + \sum_{k=1}^{m} \| y_k^i - cy_k^j \|^2} \quad (7)$$

3.1　数字属性间的距离

数字属性对簇与簇间距离平方的贡献是所有数字属性间距离的平方和,表示为[7]:

$$\sum_{k=1}^{p} \| cx_k^i - cx_k^j \|^2 = \sum_{k=1}^{p} \left(\frac{cx_k^i - cx_k^j}{\text{shorth}_k} \right)^2 \quad (8)$$

3.2　符号属性间的距离

符号属性的距离实际上是描述的符号属性间的相异程度。符号属性的相异程度正好与其相似度相反。因此可以先计算两个符号属性的相似度,然后由相似度计算其相异度,即符号属性间的距离。对符号属性两个对象的距离度量的基本要求是:如果两个对象相同或很相似,则它们间的距离为 0 或很小;如果两个对象相反或有很大的差别,则它们间的距离应取很大的值。

聚类的结果依赖于相似性函数的定义,因此该函数的选择很重要。事实上,计算两个向量相似性的方法很多[10],每种定义都有它们各自的优缺点。以下采用 Pearson 相关系数作为两个符号属性的相似性度量。

Pearson 相关系数的计算公式有很多种,其中一种常用的公式如下[10]:

$$r_{ij} = \frac{\sum X_i X_j - \frac{\sum X_i \sum X_j}{m}}{\sqrt{\left(\sum X_i^2 - \frac{\left(\sum X_i \right)^2}{m} \right) \times \left(\sum X_j^2 - \frac{\left(\sum X_j \right)^2}{m} \right)}} \quad (9)$$

用公式(9)计算簇 C_i、C_j 的符号属性 cy_k^i 和 cy_k^j 相似性的公式如下:

$$\text{sim}(cy_k^i, cy_k^j) = \frac{\sum_{r=1}^{n_k} cy_{kr}^i cy_{kr}^j - \frac{\sum_{r=1}^{n_k} cy_{kr}^i \times \sum_{p=1}^{n_k} cy_{kr}^j}{n_k}}{\sqrt{\left(\sum_{r=1}^{n_k} (cy_{kr}^i)^2 - \frac{\left(\sum_{r=1}^{n_k} cy_{kr}^i \right)^2}{n_k} \right) \times \left(\sum_{r=1}^{n_k} (cy_{kr}^j)^2 - \frac{\left(\sum_{p=1}^{n_k} cy_{kr}^j \right)^2}{n_r} \right)}} \quad (10)$$

相似度 $\text{sim}(cy_k^i, cy_k^j)$ 的取值在区间[-1, 1]内,值越大表示两个向量越相似。两个向量的相似度正好与它们间的相异度(距离)相反,因此可以定义 cy_k^i、cy_k^j 间的距离为:

$$d(cy_k^i, cy_k^j) = 1 - \text{sim}(cy_k^i, cy_k^j) \quad (11)$$

$d(cy_k^i, cy_k^j)$ 的取值范围是[0, 2]。

由此可以得到所有符号属性对簇与簇间距离平方的贡献表示为:

$$\sum_{k=1}^{m} \| cy_k^i - cy_k^j \|^2 = \sum_{k=1}^{m} (1 - \text{sim}(cy_k^i, cy_k^j))^2 \quad (12)$$

将计算簇 C_i、C_j 间数字属性距离的公式(8)和计算符号属性间距离的公式(12)代入公式(6),就可以计算出簇与簇间的距离。

点与簇间的距离计算方法与簇与簇间的距离计算方法类似。设点 E_i 为:

$$E_i = (x_1^i, x_2^i, \cdots, x_p^i, y_1^i, y_2^i, \cdots, y_m^i)$$

则所有数字属性对点 E_i 与簇 C_j 间距离平方的贡献:

$$\sum_{k=1}^{p} \| x_k^i - cx_k^j \|^2 = \sum_{k=1}^{p} \left(\frac{x_k^i - cx_k^j}{\text{shorth}_k} \right)^2 \quad (13)$$

点 E_i 的第 k 个符号属性用向量 y_k^i 表示为:

$$y_k^i = (y_{k1}^i, y_{k2}^i, \cdots, y_{kn}^i)$$

向量 y_k^i 的各个坐标中,只有点 E_i 的第 k 个符号属性的取值所对应的那一维为 1(表示该类型值出现一次),其它维均为 0(出现 0 次)。

点 E_i 的第 k 个符号属性 y_k^i 和簇 C_j 的第 k 个符号属性 cy_k^j 的相似度计算方法为:

$$\text{sim}(y_k^i, cy_k^j) = \frac{\sum_{r=1}^{n_k} y_{kr}^i cy_{kr}^j - \frac{\sum_{r=1}^{n_k} y_{kr}^i \times \sum_{p=1}^{n_k} cy_{kr}^j}{n_k}}{\sqrt{\left(\sum_{r=1}^{n_k} (y_{kr}^i)^2 - \frac{\left(\sum_{r=1}^{n_k} y_{kr}^i \right)^2}{n_k} \right) \times \left(\sum_{r=1}^{n_k} (cy_{kr}^j)^2 - \frac{\left(\sum_{p=1}^{n_k} cy_{kr}^j \right)^2}{n_r} \right)}} \quad (14)$$

定义 y_k^i、cy_k^j 间的距离为:

$$d(y_k^i, cy_k^j) = 1 - \text{sim}(y_k^i, cy_k^j) \quad (15)$$

$d(y_k^i, cy_k^j)$ 的取值范围是[0, 2]。

所有符号属性对点与簇间距离平方的贡献表示为:

$$\sum_{k=1}^{m} \| y_k^i - cy_k^j \|^2 = \sum_{k=1}^{m} (1 - \text{sim}(y_k^i, cy_k^j))^2 \quad (16)$$

将计算点与簇间数字属性距离的公式(13)和计算符号属性间距离的公式(16)代入公式(7),就可以计算出点与簇间的距离。

4　实验结果及分析

该实验比较在相同数据集,相同聚类算法下,用三种不同的簇异常度量(簇成员数量、与簇中心距离不大于聚类阈值 w 的点数 n 及近似平均距离 AAD)时,检测攻击的能力。使用的数据集是 KDD 99 数据[11]中所有 logged in 属性为 1 (已成功登录的连接) 的数据子集,共 703066 条记录,其中 3377 条是入侵,入侵记录大约占总记录数的 0.48%。聚类算法是固定宽度聚类法。检测结果如表 1 所示。

表 1 表明,采用簇成员数和 AAD 作为异常量时,

得到的检测结果基本相同。并且采用这两种度量比用计数 n 作为簇异常度量的检测结果更好。

表 1 采用不同簇异常度量时检测结果比较

簇异常度量	检测率	虚警率
簇成员数	97.8%	11.3%
	93.8%	6.9%
	80.9%	3.3%
	68.7%	1.1%
n	97.8%	16%
	93.9%	11.7%
	93.8%	6.9%
	80.9%	3.3%
	68.7%	0.9%
	62.3%	0.8%
	59.8%	0.8%
AAD	97.8%	11.3%
	93.8%	6.9%
	80.9%	3.3%
	68.7%	1.2%
	21.7%	1.2%

5 结束语

文章研究了用 Pearson 相关系数计算记录与簇、簇与簇间符号属性距离的方法,该方法根据向量的相似性计算符号属性量化的距离值,克服了许多研究中符号属性间距离只有两种值(0,或者是某个常数)的缺陷,提高了符号属性的分辨力。提出了一种新的簇分类度量:近似平均距离 AAD,AAD 综合了一个簇的局部异常度,即簇的内部点密度,和该簇在整个簇结构中的全局异常度,即该簇与其它簇的距离。用 KDD 99 评估数据集所作的实验表明用 AAD 作为簇的分类度量可以很好地区分由攻击记录组成的簇和由正常记录组成的簇。

参考文献:

[1] Singh S, Kaur G. Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters[C]//Proceedings of the National Conference on Challenges & Opportunities in Information Technology. Mandi Gobindgarh, India: [s. n.], 2007: 107-110.

[2] Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering[C]//Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001). Philadelphia, PA: [s. n.], 2001: 432-438.

[3] 关 健,刘大昕. 基于主成分分析的无监督异常检测[J]. 计算机研究与发展,2004,41(9):1474-1480.

[4] Nguyen B V. An Application of Support Vector Machines to Anomaly Detection[R]. Athens, Ohio, United States: Ohio University, 2002.

[5] 高朝勤,陈元琰,李 梅. 一种面向入侵检测的快速多模式匹配算法[J]. 计算机应用,2008,28(1):82-84.

[6] 陶善旗,李 俊. 入侵检测系统中模式匹配算法的研究与改进[J]. 计算机技术与发展,2010,20(2):167-170.

[7] Cai Longzheng, Chen Jian, Ke Yun, et al. A new data normalization method for unsupervised anomaly intrusion detection [J]. Journal of Zhejiang University - Science C, 2010, 11(10): 778-784.

[8] Eskine E, Arnold A. A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data [M]. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[9] 肖海军,王小非. 基于特征选择和支持向量机的异常检测[J]. 华中科技大学学报,2008,36(3):99-102.

[10] 任若恩,王惠文. 多元统计数据分析-理论、方法、实例[M]. 北京:国防工业出版社,1997:208-209.

[11] KDD[EB/OL]. [1999]. http://kdd.ics.uci.edu/databases/kddcup99/task.html.

(上接第 110 页)

便可以得到非常正确的检索结果。

参考文献:

[1] 任平红,陈 矗. 基于聚类主颜色和边缘直方图的图像检索方法[J]. 计算机技术与发展,2011,21(3):142-145.

[2] 叶宇光. 基于多特征信息融合的图像检索技术研究[D]. 福建:华侨大学,2006.

[3] Tumara H, Mori S, Yamawaki T. Texture features corresponding to visual perception[J]. IEEE Transactions on Systems, Man and Cybernetics, 1987, 8(6): 460-473.

[4] Young D C, Sang Y S, Namc K. Image retrieval using BDIP and BVLC moments[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(9): 951-957.

[5] Khotanzad A, Hernandez O J. Color image retrieval using multispectral random field texture model and color content features[J]. Pattern Recognition, 2003, 36(8): 1679-1694.

[6] Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(8): 837-842.

[7] 叶永伟,杨庆华,应富强. 基于小波和区域统计的纹理图像检索系统[J]. 浙江工业大学学报,2003,6(3):306-309.

[8] 朱明忠. 多尺度 Gabor 小波变换在图像检索中的应用[J]. 电子科技,2011,24(8):61-65.

[9] 汪祖媛,庄镇泉,何劲松,等. 基于形状的小波变换系数广义高斯分布图像检索算法[J]. 电子学报,2003(5):765-768.

[10] Gunn S R, Nixon M S. A Robust Snake Implementaion: A Dual Active Contour[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997, 17(5): 817-821.

[11] 张新民,沈兰荪. 基于小波和统计特性的自适应图像增强[J]. 信号处理,2001(3):227-231.

[12] 安 斌,陈书海,陈 华,等. 纹理特征在多光谱图像分类中的应用[J]. 激光与红外,2002(3):188-190.

一种无监督异常入侵检测的簇异常度量方法

作者：陈剑， 蔡龙征
作者单位：陈剑(广东科学技术职业学院 计算机工程技术学院,广东 珠海519090)， 蔡龙征(中南民族大学 工商学院,湖北 武汉430065)

刊名：计算机技术与发展
英文刊名：Computer Technology and Development
年，卷(期)：2013(4)