

半监督聚类算法及其在入侵检测中的应用

张念贵 李永忠 王汝山

(江苏科技大学计算机科学与工程学院, 镇江 212003)

摘要 针对现有入侵检测技术的不足, 对基于机器学习的异常入侵检测系统进行了研究, 提出了一种基于半监督聚类的异常入侵检测算法。此算法通过利用少量的标记样本, 生成用于初始化算法的种子聚类, 然后辅助聚类过程, 对数据进行检测。实验表明, 与以往入侵检测算法相比, 此算法可以明显地改善入侵检测系统的性能。

关键词 入侵检测 半监督学习 聚类

中图法分类号 TP393.08; **文献标志码** A

随着人工智能和数据挖掘技术的快速发展, 聚类分析在各个方面得到了广泛的重视^[1]。聚类分析试图将未知数据按照一定的准则分为几个不同的数据集, 同一组中的数据相似度较大, 不同组之间数据相似度较小。通常情况下, 聚类分析被作为一种无监督的机器学习方法。在处理大量数据问题时, 我们经常会得到部分数据的先验信息。充分利用这些信息, 有利于对数据进行更好的分类。在处理数据过程中用聚类方法分析数据, 这就是半监督聚类学习算法。现有的入侵检测系统大都是基于监督学习算法, 需要获得大量的标记数据。虽然利用大量标注过的样本信息可以有效地提高系统的检测率, 降低误报率, 但却忽略了未标记数据的作用, 而有效地利用未标记数据的信息无疑将在一定程度上提高入侵检测系统的性能。半监督聚类入侵检测系统能充分利用部分已知数据信息和未知数据, 形成聚类并对数据进行检测。本文对此进行研究, 发现半监督聚类算法可以明显地改善入侵检测系统的性能。

1 机器学习

机器学习属于人工智能的范畴, 是指机器获得新知识和新技能, 并识别现有知识的能力^[2]。这里将对机器学习进行简单介绍。

在传统的异常检测中, 训练数据集上的所有数据必须确保是“正常”的数据单元, 并利用机器学习算法在这些训练数据集上对检测模型进行训练, 系统通过对正常数据的抽象, 对未知数据进行检测, 偏离检测模型的数据便被认为是异常数据。其中抽象是指剖析对象具有的特征, 抛弃与目前工作不相关的细节, 保留具有实际处理意义的特征。异常检测的优势在于它可以检测出新的入侵, 因为这些新的入侵往往偏离建立在正常数据之上的检测模型。这些算法被称为监督学习算法^[3], 它需要一个由“完全干净”的正常数据组成的数据集来对模型进行训练。但在实际中, 我们很难获得这种数据集, 或者获得的代价很大。这就使得传统的基于监督学习的检测算法实现起来很复杂。

无监督学习^[4]是一种自学习方式, 不需要对学习样本做类别标记, 能处理不带标记且含异常数据的训练数据, 分类过程无需人的干预, 可以在较低的误报率下检测出新的攻击类型。但在完全不提供监

2009年9月4日收到

江苏省教育厅、江苏科技大学

课题(2005DX006J)资助

第一作者简介: 张念贵(1983—), 男, 江苏徐州人, 硕士研究生, 研究方向: 网络安全。

督信息的情况下,该算法学习得到的模型不够精确,不能获得令人满意的学习结果。

半监督学习^[5]是机器学习领域中一个新的研究热点,它是介于监督学习和无监督学习之间的一种学习方式,学习样本既包括已标记类别样本也包括未标记类别样本。在已标记类别样本提供的监督信息的“引导”下,处理未标记样本。半监督学习方式是以假设同类别的未标记数据与已标记数据在特征空间上的某种距离最近为基础。它只需要提供少量的标记样本,而通过全部样本的学习又可以获得相对于无监督学习更好的学习效果。

2 半监督聚类入侵检测算法

2.1 半监督聚类

聚类是在预先不知道目标数据库到底有多少类的情况下,希望将所有的记录组成不同的“聚类”或者说簇,并且使得在这种分类情况下,以某种度量为标准的相似性,在同一聚类之间最大化,不同聚类之间最小化。在解决实际问题时我们可以轻易获得少量数据的部分信息。研究表明,在聚类搜索过程中充分利用已知数据的信息会显著提高聚类算法的性能。通过利用少量数据的先验信息来改善无监督聚类算法的性能,所提出的算法被统称为半监督聚类算法^[6]。

k-近邻法^[7]是一种非参数估计方法,存在着大量计算问题,时间复杂度为 $O(n^2)$,难以满足异常入侵检测的速度要求。本文采用了一种改进的半监督 *k*-近邻算法,充分利用聚类所得到的结果,提高了对新攻击类型的检测率。

数据集 $X = \{x_1, x_2, \dots, x_n\}$, $c(x)$ 表示数据 x 所在的簇的中心, $r(x)$ 表示数据 x 所在簇的半径, $d(x, y)$ 表示点 x 和 y 之间的距离。设各簇的形状为不相交的圆,根据三角不等式,对于 $\forall x_i, x_j \in X$,有:

$$|d(x_i, c(x_j)) - d(x_i, x_j)| \leq d(x_j, c(x_j)) \quad (1)$$

根据 $r(x)$ 定义得:

$$d(x_j, c(x_j)) \leq r(x_j) \quad (2)$$

由式(1),式(2)得:

$$d(x_i, x_j) \geq d(x_i, c(x_j)) - r(x_j) \quad (3)$$

由式(3)可得,当 $d(x_i, c(x_j)) > r(x_j)$ 时,数据 x 至其它簇中数据距离下限为

$$d_{\min} = \min_{x_i, x_j \in X} \{d(x_i, c(x_j)) - r(x_j) \mid c(x_i) \neq c(x_j)\} \quad (4)$$

根据式(4),将未标记数据加入某簇,同时更新各参数,然后处理下一个未标记数据。通过反复迭代,确定所有未标记数据的类别。试验结果证明,此过程降低了算法的复杂性,提高了算法的性能。

2.2 基于半监督聚类的入侵检测算法

本文采用的是改进的基于 k -近邻的半监督聚类算法,全称为:Improved Semi-supervised Clustering Algorithm based on K-means (ISCA)^[8]。算法假设各簇大小基本相同,理想聚类形状为正六边形,实际操作时以圆形近似。算法具体过程如下:

ISCA 算法描述:

输入:已标记数据集合:

$D_{\text{label}} = \{(x_i, l_i) \mid i = 1, 2, \dots, n\}$, l_i 为数据 x_i 所在的簇;未标记数据集合:

$$D_{\text{unlabel}} = \{x_i \mid i = n + 1, n + 2, \dots, n + p\},$$

其中 $n \ll p$;数据集 $D = D_{\text{label}} \cup D_{\text{unlabel}}$,聚类 C , k -近邻参数 k ;

输出:未标记数据 $x \in D_{\text{unlabel}}$ 的类型(正常、已知攻击或新攻击);

算法:

(1) 对已标记数据集 D_{label} 进行聚类,使得每个簇中仅包含相同内容的数据,即 $\forall x_i, x_j \in C_h$, $C_h \subseteq D_{\text{label}}$,有 $l_i = l_j = h$,且各簇中心为 $\mu_m = \frac{1}{|C_m|} \sum_{x \in C_m} x$, $m = 1, 2, \dots, \lambda$,其中 λ 为标记数据生成的簇的数目,求 $r(x)$,确定最大半径 R_{\max} 。

(2) 对未标记数据 $\forall x_i \in D_{\text{unlabel}}$,计算 $l_i = \min \{||x - \mu_m|| \mid m = 1, 2, \dots, \lambda\}$;比较 l_i 和 R_{\max} ,若 $l_i < R_{\max}$,则将 x 加入 C_m ,更新 μ_m 与 $r(x_m)$;否则,令 $\lambda = \lambda + 1$,重新计算簇中心 $\mu_\lambda = \frac{1}{|C_\lambda|} \sum_{x \in C_\lambda} x$ 。

(3) 对包含未标记数据的簇,根据标记数据类

型确定该簇类型，并赋值给簇中未标记数据。

(4) 得到未标记簇集 C_u 和标记簇集 C_l , $C = C_l \cup C_u$ 。

(5) 对于未标记数据 x , $x \in C_r$, $C_r \subseteq C_u$, 确定 x 的类型:

a) 求 μ_r 至 μ_u 的距离共 $|C_u|$ 条, 确定最短距离 $\mu_r \mu_u$ ($\mu_r \in C_r$), 将 C_r 中数据加入 P , 令 $C \leftarrow C - \{C_r\}$, 求出 d_{\min} ;

b) 对于 $x_i \in P$, 如果 $d(x, x_i) < d_{\min}$, 当 $|K| < k$ 时, 将 x_i 加入 K 中; 当 $|K| = k$ 时, 若 $d(x, x_i) < \max d(x, x_j)$, 用 x_i 替换 x_j , 否则执行步骤(6);

c) 如果 $d(x, x_i) < \max d(x, x_j)$, 则将 C 中距离 x 最近的簇 C_n 中数据加入 P , 同时令 $C \leftarrow C - \{C_n\}$, 更新 d_{\min} 。

(6) 根据多数原则利用集合 K 中的数据确定 x 的数据类型:

$$label(x) = \begin{cases} label_j, & \text{当 } |label_i| < |label_j| \text{ 时,} \\ & \text{其中 } 1 \leq i, j \leq s, i \neq j; \\ new - attack, & \text{当 } label(x_i) \text{ 未知, 其中} \\ & x_i \in K. \end{cases}$$

(7) 重复步骤(5),(6), 直至 C_u 为空。

3 实验及结果分析

实验采用 KDD CUP 99 数据集^[9]进行实验。从数据集中选取 20,000 条数据组成实验的数据集, 其中有 200 条攻击数据, 这满足聚类假设: 正常数据远远多于入侵数据。然后对其中的 400 条(2%)数据进行标记, 用这些数据作为系统的种子, 生成初始聚类并对数据进行检测。

比较 ISCA 算法与 Combined 算法^[10]、ACKID 算法^[11]检测攻击的性能。将 ISCA 算法和 ACKID 算法中 k 都设置为 100。

表 1 显示了三种不同算法应用于入侵检测的结果。从实验结果可以看出 ISCA 算法在同等情况下较另外两种算法有较低的误报率。无论面对已知攻击还是未知攻击, ISCA 的误报率都较低。同时从表

中也可以看出, ACKID 算法和 Combined 算法对未知数据攻击有较高的检测率, 但这种高检测率是以牺牲它们的误报率为代价: 系统确实能将大部分攻击检测出来, 但同时也给用户带来了很大的麻烦: 用户不得不面对系统提供的大量的虚假报警, 此时用户将疲于处理这些虚假报警, 从而降低了工作效率。ISCA 算法在面对已知攻击时, 与其它两种算法相比不仅有较低的误报率, 而且其检测率也较高; 在面对未知攻击时, 系统在一定范围内以降低较小的检测率为代价, 远远降低了系统的误报率, 而且此时的检测率是用户可以接受的。这样用户就可以不用面对大量虚假报警, 而将精力放在不断变化的新的攻击上, 在控制较低的误报率的情形下, 提高入侵检测系统的性能。

表 1 入侵检测算法比较

数据 类型	ACKID 算法 (%)		Combined 算法 (%)		ISCA 算法 (%)	
	检 测 率	误 报 率	检 测 率	误 报 率	检 测 率	误 报 率
已知 攻击	76.3	5	75.7	5	81.9	3.56
未知 攻击	93.1	10	92.5	10	74.9	1.07

4 结束语

本文对目前现存的入侵检测系统进行了粗略的分析比较, 通过对基于机器学习的入侵检测进行研究, 提出了一种基于半监督聚类的异常入侵检测算法。通过实验发现这种算法在面对已知攻击数据时不仅有较低的误报率, 而且有较高的检测率。在面对未知攻击数据时, 此算法与基于监督学习算法的入侵检测系统相比, 检测率低于其它入侵检测系统。但该算法是在较小的范围内放弃对部分攻击数据的检测, 并以此为代价远远降低了系统的误报率。虽然此时系统的检测率可以为普通用户所接受, 但相对于基于监督学习的入侵检测系统相比, 此算法的

检测率还有待提高。如何改进半监督聚类算法,将其更好的应用于入侵检测系统之中,并在控制较低的误报率的前提下,最大限度地提高系统的检测率,完善系统对各种攻击数据的检测,仍然是今后长期内需要研究和改进的问题。

参 考 文 献

- 1 李永忠,孙彦,罗军生. WINEPI 挖掘算法在入侵检测中的应用. *计算机工程*,2006;32(23):159—161
- 2 郑毅. 基于机器学习的 IDS 研究. *现代电子技术*,2006;21:98—99,102
- 3 李颖,彭广川. 非监督学习算法应用于异常检测的效果评估. *电脑知识与技术*. 2005;3:32—36
- 4 Weber M, Welling M, Perona P. Towards automatic discovery of object categories., *IEEE Conf on Computer Vision and Pattern Recognition*, Hilton Head Island,2000
- 5 孙广玲,唐降龙. 基于分层高斯混合模型的半监督学习算法. *计算机研究与发展*,2004;41(1):157—161
- 6 邵峰晶,于忠清. *数据挖掘原理与算法*. 北京:中国水利水电出版社,2003
- 7 边肇祺,张学工. *模式识别*. 第 2 版. 北京:清华大学出版社,2000
- 8 俞研,黄皓. 一种半聚类的异常入侵检测算法. *计算机应用*,2006;26(7):1640—1642
- 9 KDD99. KDD99 CUP dataset. <http://kdd.ics.uci.edu/databases/kddcup99.htm1>
- 10 Shi Zhong, Khoshgoftaar T M, Seliya N. Clustering-based network intrusion detection. In: *International Journal of Reliability, Quality, and Safety Engineering (IJRQSE)*,2005;14(2):169—187
- 11 李雯睿. 基于半监督聚类的入侵检测算法研究. 开封:河南大学硕士学位论文. 2007

Semi-supervised Clustering Algorithm and Its Application to Intrusion Detection

ZHANG Nian-gui, LI Yong-zhong, WANG Ru-shan

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, P. R. China)

[Abstract] Aiming at some problems in current technique of intrusion detection and studying the anomaly intrusion detection system based on machine learning, an anomaly intrusion detection system based on semi-supervised clustering is proposed. It uses a little quantity of labeled data to generate seeds which applied in initiating the algorithm, and then assists the clustering process to detect data. The experiment results manifest that comparing with the past algorithm, this algorithm can significantly improve the performance of intrusion detection system.

[Key words] intrusion detection semi-supervised learning cluster