

文章编号:1001-9081(2008)07-1781-03

一种新的半监督入侵检测算法

宋凌, 李枚毅, 李孝源

(湘潭大学 信息工程学院, 湖南 湘潭 411105)

(7145366@qq.com)

摘要: 针对无监督学习的入侵检测算法准确度不高、监督学习的入侵检测算法训练样本难以获取的问题, 提出了一种粒子群改进的 K 均值半监督入侵检测算法, 利用少量的标记数据生成正确样本模型来指导大量的未标记数据聚类, 对聚类后仍未能标记的数据采用粒群优化的 K 均值聚类, 有效提高分类器的分类准确性, 并实现了对新类型攻击的检测。实验结果表明, 算法的整体检测效果明显优于基于无监督学习和监督学习的检测算法。

关键词: 半监督聚类; 入侵检测; 粒群优化; K 均值

中图分类号: TP393.08 **文献标志码:**A

Novel intrusion detection algorithm based on semi-supervised clustering

SONG Ling, LI Mei-yi, LI Xiao-yuan

(Institute of Information Engineering, Xiangtan University, Xiangtan Hunan 411105, China)

Abstract: An anomaly intrusion detection algorithm based on semi-supervised clustering along with PSO K-means was presented. It could solve the problems of the low detection rate of the intrusion detection algorithms based on unsupervised learning, and the insufficiency of training samples of the intrusion detection algorithms based on supervised learning. The algorithm utilized minimal labeled data and lots of unlabeled data to improve its learning capability, and novelty detection could also be carried out. The experimental results manifest that the detection results of the algorithm outperforms both the one based on unsupervised learning remarkably and the one based on supervised learning.

Key words: semi-supervised clustering; intrusion detection; particle swarm optimization; K-means

0 引言

基于机器学习的异常检测方法有: 监督学习^[1]、无监督学习^[2]和半监督学习^[3]。基于监督学习的异常入侵检测算法, 首先需要标注样本的类别用来构成训练样本集, 由于训练样本集的建立主要依赖于安全技术领域的专家, 因而训练样本集的构建的代价较大; 同时, 为了提高分类精度, 在学习过程中需要的训练样本必须足够多, 这又增加了构建训练样本集的成本; 而且对大量样本的学习也需要耗费大量的机器学习时间。基于非监督学习的异常入侵检测算法根据数据的相似性进行分组, 克服了监督学习方法中需要标记数据的不足。然而, 非监督的检测算法的检测精度明显低于有监督的检测算法。而在现实应用中, 获得少量的标记数据是可能的, 因此, 利用少量的标记数据进行指导的半监督学习技术正在获得广泛的关注。已有的半监督入侵检测算法检测精度仍然可以得到提高, 特别是对新攻击类型的检测。

针对以上基于机器学习的异常入侵检测算法所出现的问题, 本文提出了一种基于粒群优化^[4]的半监督异常检测算法 (Semi-Supervised Clustering with PSO K-means, PSSC), 首先对少量标记数据进行监督聚类得到正确模型, 然后利用这些模型指导大量未标记数据的聚类, 扩充标记数据集合, 最后对仍没有确定标记的数据利用粒群优化的 K 均值算法进行聚类, 确定其标记类型。经入侵检测实验表明此算法对已知攻击检测率明显高于无监督的入侵检测算法, 接近监督学习的入侵检测算法, 并成功实现了对未知攻击的检测。

1 粒群优化的半监督聚类 (PSSC)

1.1 基于粒群优化的 K 均值算法

在 K 均值^[5]聚类算法中, 若其初始化落在了一个局部极小点附近, 就会造成算法在局部极小处收敛。因此初始聚类中心的随机选取可能会陷入局部最优解, 而难以获得全局最优解, 直接影响聚类的效果。为解决以上问题, 我们采用粒群优化来指导初始聚类中心的选取。用粒子代替每个数据向量, 依据性能函数驱动粒子在群中移动, 找出最佳中心点。目标就是用粒群优化来初始化 K 个聚类的中心点, 使其取值不会落入局部极值点, 从而使聚类算法能收敛到全局最优。

基本粒子群算法的进化方程可表示如下:

$$V_i(t) = \omega V_i(t-1) + \rho_1(x_{pbest_i} - x_i(t)) + \rho_2(x_{gbest_i} - x_i(t)) \quad (1)$$

$$\begin{cases} X_i(t) = X_i(t-1) + V_i(t) \\ t = t + 1 \end{cases} \quad (2)$$

其中: $\rho_1 = r_1 c_1$; $\rho_2 = r_2 c_2$; r_1, r_2 为 $[0, 1]$ 中的一个随机数; c_1, c_2 为正常数; ω 为权重系数, 取值 $[0, 1]$ 。

粒群优化算法所用到的目标函数公式定义如下:

$$F(x_i) = 1/f(D_{\max}(x_i, x_j), B) \quad (3)$$

其中: $D_{\max}(x_i, x_j)$ 指的是点 x_i 与聚类中其他任意点 x_j 之间的欧几里德距离的最大值。B 是一个常系数, 通过测试选 $B = 1/8$ 时结果理想。在某个点分布的空间中, 点分布得越密集的地区, 越可以看成一个聚类, 因此可以用点的分布密度作为每个粒子性能指标。我们把点密度定义为: 与某个点的距离 (欧几里德距离) 小于指定的半径 (d) 的点的个数。半径取空间

收稿日期: 2008-01-21; 修回日期: 2008-04-01。

基金项目: 湖南省自然科学基金资助项目(06JJ5106); 湖南省教育厅科学研究项目(06C841); 湘潭大学研究生创新基金资助项目。

作者简介: 宋凌(1978-), 男, 湖南株洲人, 硕士研究生, 主要研究方向: 智能系统、模式识别; 李枚毅(1962-), 男, 湖南湘乡人, 教授, 博士, 主要研究方向: 人工智能、智能计算; 李孝源(1978-), 男, 湖南涟源人, 硕士研究生, 主要研究方向: 智能计算。

任意两点间距离最大值乘以系数 B , 即 $d = D_{\max}(x_i, x_j)/8$, 当目标函数值取得最小时, 所得点就是最优中心点。

1.2 PSSC 算法描述

输入: 已标记数据集 $S_l = \{(x_i, l_i) | i = 1, 2, \dots, n\}$, 未标记数据集 $S_u = \{x_i | i = 1, 2, \dots, m\}$, $n \ll m$, 数据集 $S = S_l \cup S_u$

输出: 数据 $x \in S_u$ 的数据类型(正常或异常)

第 1 步 对已标记数据集 S_l 进行监督聚类, 计算每个聚类的中心 O_λ , λ 为生成的簇的数目, 并计算各聚类的最大半径 R_λ 。

第 2 步 $\forall x \in S_u$ 计算 x 与各聚类中心 O_λ 的距离 $r_i, i = 1, 2, \dots, \lambda$, 令 $r' = \min(r_i | i = 1, 2, \dots, \lambda)$ 。若 $r' \leq R_\lambda$, 则将 x 分配给聚类 C_λ , 否则将 x 分配给聚类 C_r 。

第 3 步 对于聚类 C_r , 将其所有数据均分为 K 个子集, 针对所选用粒子数 A , 对每个粒子 $x_i, i = 1, 2, \dots, A$ 进行粒子初始化, 确定每个粒子 x_i 的初始位置与速度。

第 4 步 根据粒子当前位置 $x_i(t)$ 计算每个粒子的性能 $F(x_i(t))$, 若达到最大迭代次数 MAXITER, 转第 3 步学习下一样本, 当全部样本学完转第 5 步, 否则继续执行。

1) 比较粒子的当前性能 $F(x_i(t))$ 与其有过的最好性能 $pbest_i$, 如果 $F(x_i(t)) \leq pbest_i$, 则:

$$\begin{cases} pbest_i = F(x_i(t)) \\ x_{pbest_i} = x_i(t) \end{cases} \quad (4)$$

2) 比较每个粒子的性能 $F(x_i(t))$ 与全局最优粒子的性能 $gbest_i$, 如果 $F(x_i(t)) \leq gbest_i$, 则:

$$\begin{cases} gbest_i = F(x_i(t)) \\ x_{gbest_i} = x_i(t) \end{cases} \quad (5)$$

3) 按式(1)和(2)分别改变粒子速度和位置。

第 5 步 重复执行第 6 步、第 7 步, 直到各个聚类中心不再发生变化。

第 6 步 从 $i = 1$ 到 n 分别取 $x_i \in C_r$, 然后从已有的 K 个中心找到一个距离 x_i 最近的中心点 O_j , 即求出使 $D(o_j, x_i)$ 最小的 O_j , 并将 x_i 加入聚类 C_j 。

第 7 步 重新计算各个新的聚类的中心点 $O_j (j = 1, 2, \dots, k)$, 即计算新的聚类成员的平均值, 用这个平均值点作为新的聚类的中心点。

第 8 步 $\forall x \in C_r$, 若 $C_r . sunnum \leq M$, 则 $x_{.l} =$ 异常, x 分配给 C_a , 否则 $x_{.l} =$ 正常, x 分配给 C_n 。

第 9 步 重复第 8 步, 直到 C_r 为空。

2 实验及分析

2.1 实验数据选取及预处理

基于聚类的入侵检测算法建立在两个假设条件上:

条件 1 正常数据的数目远远大于异常数据的数目。

条件 2 异常数据从本质上与正常数据不同。

该方法的基本思想就是由于异常数据是和正常数据不同的并且数目相对很少, 因此它们在能够检测到的数据中呈现出比较特殊的特性。

为了测试本文提出的基于粒群优化的半监督入侵检测算法的效果, 我们采用了 KDD CUP 1999 数据集^[6]作为测试集。该数据集是在入侵检测领域广泛使用的测试集, 采用 1998DARPA 入侵检测数据集来构造连接记录及提取特征。每个连接共有 41 种定性和定量的特征, 其中有 8 个属性是离散型的变量, 其余是连续型的数字变量。入侵数据有 4 大类, 24 小类。4 大类分别是: 1) DoS(拒绝服务攻击); 2) R2L(对远程主机的未授权的访问); 3) U2R(对本地超级用户权限的未授权的访问); 4) Probe(扫描与探测行为)。

KDD 原始数据集包含 4 900 000 多条连接记录, 过于庞大, 为了对算法进行验证, 本文从整个数据集中选择了 62 000 条数据作为测试集。其中正常数据占 61 448 条, 异常数据 552 条。正常数据所占比例大于 99%, 基本满足检测算法的第一个假设要求。

样本中的入侵数据类型及数目见表 1。

我们将选取的样本分成 6 组, 每一组包含 10 000 多条数据, 前 4 组作为训练样本, 训练结果取各组的平均值, 第 5、6 组作为测试样本。测试时我们对于每一组数据标记其中 1% 的数据作为已标记样本, 标记后测试集中包含了 20 种已知攻击类型和 4 种未知攻击类型。

表 1 样本中的入侵数据种类及数目

类型	小类型及数目	合计
DoS	back(48), land(6), neptune(54), smurf(54), teardrop(12), pod(30)	204
U2R	Buffer_overflow(36), perl(6), rootkit(30)	72
R2L	ftp_write(6), guess_passwd(30), multihop(12), phf(6), warezmaster(12), snmpguess(12), snmpgetattack(18), xsnoop(6), named(12)	114
Probe	Ipsweep(60), nmap(30), satan(12), portsweep(60)	162
总计		552

由于 KDD Cup 1999 Data 是由连续型数据和离散型数据组成的混合数据, 为了便于以后各分类器的构造, 我们在进行特征选择前对数据库中的数据进行了预处理, 将符号型字段进行离散化处理, 转换成数值型记录, 另由于数据间数值差异较大应该对原始数据进行规范化处理, 以免某个属性对结果的权值过大, 出现大数属性淹没小数属性的情况。设输入数据为 $S = \{x_1, x_2, \dots, x_n\}$, 则规范化处理步骤如下:

1) 计算平均值

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (6)$$

2) 计算标准差

$$\sigma(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

3) 计算归一化后的值

$$x'_i = \frac{x_i - \bar{x}}{\sigma(x)} \quad (8)$$

2.2 实验结果

在实验中所用到有关参数如表 2。

表 2 实验参数的选取

参数名	值	取值原因
MAXITER	300	经测试迭代次数大于 300 次后结果基本无变化
A	30	经测试 30 个粒子已能够满足要求
ω	0.95	根据经验
r_1, r_2	[0, 1]	使用 rand() 函数取[0, 1] 中的一个随机数
c_1, c_2	2	根据经验
M	/	在实验中取得
K	/	在实验中取得

为了评价入侵检测算法的性能, 我们主要考虑与入侵检测性能相关的统计量: 检测率和误检率。分别定义如下:

$$\text{检测率} = \frac{\text{检测到的异常数据个数}}{\text{样本中异常数据总数}}$$

$$\text{误检率} = \frac{\text{误报为异常的正常数据个数}}{\text{样本中正常数据总数}}$$

聚类数 K 和阈值 M 的大小直接影响聚类的效果, 但又没

有很好的方法进行确定。因此只能试探性地选取不同的 K 和 M 值来进行实验。我们将 K 值从一个较小值开始不断增大。而对于每个固定的 K 值,把 M 值从小到大依次进行聚类。

表 3 列出了选取不同的 K 和 M 时算法的入侵检测结果。从表 3 中可以看出,PSSC 算法对已知攻击检测率达到 88.11%,对未知攻击的检测率平均达到了 62.04%,误检率平均只有 0.72%。当 $K = 60, M = 15$ 时 PSSC 算法的整体效果较好,因此我们选取这两个值进行下一步实验。

表 3 选取不同的 K 和 M 时算法的入侵检测结果

聚类数 K	阈值 M	PSSC 算法			
		已知攻击 检测率/%	未知攻击 检测率/%	检测率/%	误检率/%
20	5	84.15	22.22	78.02	0.65
20	10	84.15	22.22	78.02	0.65
20	15	91.46	55.56	87.91	0.65
20	30	91.46	55.56	87.91	0.71
30	5	84.15	33.33	79.12	0.65
30	10	84.15	55.56	81.32	0.68
30	15	91.46	55.56	87.91	0.71
30	30	91.46	55.56	87.91	0.78
40	5	84.15	33.33	79.12	0.65
40	10	84.15	33.33	79.12	0.70
40	15	91.46	66.67	89.01	0.74
40	30	91.46	66.67	89.01	0.74
50	5	84.15	44.44	80.22	0.65
50	10	84.15	88.89	84.67	0.73
50	15	91.46	88.89	91.21	0.73
50	30	91.46	88.89	91.21	0.86
60	5	84.15	44.44	80.22	0.65
60	10	85.37	88.89	85.71	0.71
60	15	92.68	88.89	92.31	0.75
60	30	92.68	88.89	92.31	0.85
70	5	84.15	44.44	80.23	0.65
70	10	85.37	88.89	85.71	0.71
70	15	92.68	88.89	92.31	0.79
70	30	92.68	88.89	92.31	0.93
平均		88.11	62.04	85.53	0.72

表 4 PSSC 算法与其他入侵检测算法比较

攻击类型	PSSC		SAID		Wenke Lee	
	检测率	误检率	检测率	误检率	检测率	误检率
已知攻击	92.7	0.75	81.9	3.52	80.2	-
未知攻击	88.9	0.83	74.9	1.07	37.7	-

从表 4 可以看出,PSSC 算法在已知攻击的检测效果上略优于 SAID 算法^[7] 和 WENKELee 算法^[8],在未知攻击的检测

(上接第 1780 页)

4 结语

本文提出随机混沌动力系统组的概念,并将序列密码体系与之相结合,设计出一种基于随机混沌动力系统组的序列加密算法。由于混沌动力系统组比单一混沌动力系统复杂得多,加之混沌系统本身所具有的特性,该算法使明文、密钥及密文之间形成了复杂而敏感的非线性关系,其密钥空间很大且密文与明文的相关度很小,可有效抵抗各类攻击。

参考文献:

- [1] 冯登国.密码分析学[M].北京:清华大学出版社,2000:55~92.
- [2] SHANNON C E. Communication theory of secrecy systems [J]. Bell

效果上明显优于 WENKE Lee 算法,相比 SAID 算法也提高不少,同时误检率明显降低,因而更具实际利用价值。

3 结语

实验结果表明 PSSC 算法不仅对已知攻击的检测率上接近基于监督的入侵检测算法,明显高于无监督的入侵检测算法,而且对未知攻击的检测率也明显高于监督的入侵检测算法。因为 PSSC 算法利用少量的标记样本产生正确的样本模型来指导大量未标记样本进行监督聚类,对于聚类后仍没有标记的样本采用粒群优化的 K 均值算法进行无监督聚类,实现了对未知攻击的检测,从而有效弥补了单纯基于监督学习或无监督学习的入侵检测算法的不足。实验还将 PSSC 算法与其他相关的基于半监督学习的入侵检测算法进行了比较,其检测也效果优于其他的算法。但由于对实验中在入侵数据选取上的稍许不同,各算法的检测结果有可能有稍许偏差。

本算法是基于特征值的差别来进行聚类的。实验结果表明,要进一步提高算法的分类精度,另一个有效途径就是对特征值进行分类聚类。即:将 41 个特征值按其相关属性分成若干个小类,然后分别对每一类进行聚类,最后再按特征值的某种权值计算方法来确定被重复分类的数据的类别标识。

参考文献:

- [1] WESTON J, WATKINS C. Multi-class support vector machines [R]. Royal Holloway, Department of Computer Science: University of London, 1998.
- [2] FLANAGAN J A. Unsupervised clustering of symbol strings[C]// International Joint Conference on Neural Networks, IJCNN'03. Portland Oregon, USA: [s. n.], 2003: 3250~3255.
- [3] BASU S, BANERJEE A, MOONEY R. Semi-supervised clustering by seeding[C]// Proceedings of the 19th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers, 2002: 19~26.
- [4] KENNEDY J, EBERHART R C. Particle swarm optimization [C]// Proceedings of IEEE International Conference on Neural Networks. Perth Australia: [S. l.]: IEEE Press, 1995: 1942~1948.
- [5] SCLIM S Z, LEMAIL A. K-means-type algorithm: A generalized convergence theorem and characterization of local optima reality [J]. IEEE Transaction Pattern Analysis and Machine Intelligence, 1984, PAMI-6(1): 81~87.
- [6] The UCI KDD Archive. KDD99 cup dataset [EB/OL]. [2007-10-10]. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.
- [7] 俞研,黄皓.一种半聚类的异常入侵检测算法[J].计算机应用,2006,26(7):1640~1643.
- [8] LEE W, STOLFO S, MOK K. A data mining framework for building intrusion detection models [C]// Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland, USA: IEEE Press, 1999: 120~132.
- [9] Systems Technical Journal, 1949, 28(4): 656~715.
- [10] 陈鲁生,沈世镒.现代密码学[M].北京:科学出版社,2002: 100~101.
- [11] 郝柏林.从抛物线谈起——混沌动力学引论[M].上海:上海科技教育出版社,1993: 122~125.
- [12] 李红达,冯登国.复合离散混沌动力系统与序列密码体系[J].电子学报,2003,31(8): 1209~1212.
- [13] BAPTISTA M S. Cryptography with chaos[J]. Physics Letters A, 1998, 240(12): 50~54.
- [14] KOTULSKI Z, SZCZEPANSKI J. Application of discrete chaotic dynamical systems in cryptography —DCC method [J]. International Journal of Bifurcation and Chaos, 1999, 9(6): 1121~1135.
- [15] 黎全,赵凯,邓正才,等.对一种混沌加密图像方法的破译研究.国防科技大学学报,2007,3(29): 46~49.