

主动学习半监督聚类入侵检测算法

李永忠, 李正洁, 荆春伟, 胡 翰

(江苏科技大学 计算机科学与工程学院, 江苏 镇江 212003)

摘 要: 针对聚类的入侵检测算法误报率高的问题, 提出一种主动学习半监督聚类入侵检测算法. 在半监督聚类过程中应用主动学习策略, 主动查询网络中未标记数据与标记数据的约束关系, 利用少量的标记数据生成正确的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改进的 K -近邻法进一步确定未标记数据的类型, 实现对新攻击类型的检测. 实验结果表明了算法的可行性及有效性.

关键词: 入侵检测; 半监督聚类; 主动学习

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2011)10-0028-04

Research of Intrusion Detection Algorithm Based on Semi-Supervised Clustering

LI Yong-zhong, LI Zheng-jie, JING Chun-wei, HU Han

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

Abstract: Aiming at the problem for labeled data that intrusion detection algorithms based on supervised learning, the ASCID algorithm for intrusion detection based on semi-supervised is proposed in this paper, by applying active learning strategy to semi-supervised clustering process. Active learning queries constrains on labeled data and unlabeled data, which uses minimal labeled data to generate the correct sample data model and guide lots of unlabelled data clustering, the algorithm performance is improved by use an improved K -nearest neighbor algorithm to further define the type of unlabeled data after clustering. Finally, the experiment results show the feasibility and validity of the algorithm.

Key words: intrusion detection; semi-supervised clustering; active learning

1 引言

利用少量标记数据的半监督学习技术^[1]引起了广泛关注. 由于标记数据所占比例小, 在利用半监督聚类算法^[2]对数据集进行聚集并标记后, 仍会有部分数据可能未被正确地标记类型. 主动学习策略对数据的标记情况和未标记数据的分布进行分析, 通过查询标记数据来引导采样过程, 利用较少的标签数据来提升分类器的性能. 针对监督学习的入侵检测算法在标记数据不足的情况下的无法检测未知攻击的问题和非监督学习的入侵检测算法无需标记样本可以检测出未知攻击但误报率高的问题, 本文提

出一种主动学习结合改进的 k -近邻法半监督聚类的入侵检测算法 ASCID (Active-learning Semi-supervised Clustering Intrusion Detection), 在检测过程中采用主动学习策略选择最有用的标记数据, 查询标记数据与未标记数据的约束关系, 利用前面聚类所得到的结果, 计算未知样本与各簇中心的距离, 提高了检测率, 降低了误报率.

2 主动学习与半监督聚类

2.1 主动学习

在主动学习中, 学习器自行挑选出一些未标记数据并通过这样的过程查询获得这些数据的标记,

收稿日期: 2011-04-20; 修回日期: 2011-06-25

基金项目: 江苏省高校自然科学基金资助项目(05KJD52006); 江苏科技大学科研资助项目(2005DX006J)

然后再将这些有标记数据作为训练例来进行常规的监督学习. 主动学习的基本原理是用少量的标记数据建立初始分类器, 每次学习过程中 分类器可以主动在未标记的数据集中选择最有利于分类器性能的数据, 并将这些数据以一定的方式加入到训练集中, 来进一步训练分类器. Freund 等^[3]从理论上证明, 在一定假设条件下, 采用合理的主动采样策略达到相同的学习效果(即相同泛化误差, generalization error), 所需训练样本可减少到任意采样情况下的对数倍, 而假设用户查询总是可以得到高的期望信息, 主动学习的泛化误差将以所查询样本数量指数倍的速度下降^[4].

利用未标记数据来改善学习性能, 假设所有数据服从于某个由 L 个高斯分布混合而成的分布^[5], 即

$$f(x | \theta) = \sum_{l=1}^L \alpha_l f(x | \theta_l) \quad (1)$$

式中, $\sum_{l=1}^L \alpha_l = 1$ 为混合系数, $\theta = \{\theta_l\}$ 为参数. 这样, 标记就可视为一个由选定的混合成分 m_i 和特征向量 x_i 以概率 $P(c_i | x_i, m_i)$ 决定的随机变量. 于是, 根据最大后验概率假设, 得到最优分类式:

$$h(x) = \arg \max_k \sum_j p(c_i = k | m_i = j, x_i) p(m_i = j | x_i) \quad (2)$$

其中

$$P(m_i = j | x_i) = \frac{\alpha_j f(x_i | \theta_j)}{\sum_{l=1}^L \alpha_l f(x_i | \theta_l)}$$

这样, 学习目标就变成了利用训练例来估计 $p(c_i = k | m_i = j, x_i)$ 和 $p(m_i = j | x_i)$, 这两项中的第一项与类别标记有关, 而第二项并不依赖于数据的标记, 因此, 如果有大量的未标记数据可用, 则意味着能够用于估计第二项的数据数显著增多, 这会使得第二项的估计变得更加准确, 从而导致式(2)更加准确, 也就是说, 分类器的泛化能力得以提高^[5]. 本文提出的 ASCID(Active-learning Semi-supervised Clustering Intrusion Detection)算法通过分析标记数据和未标记数据, 用户通过对约束进行查询引导采样过程. 大部分标记数据的约束在初始阶段具有较高的有用性, 迭代过程的进行提高了分类器的预测能力, 此时仅需较少数据的约束就能提升分类器的性能.

2.2 半监督聚类过程

在 K-Means 算法^[6]中, 根据距离远近数据被分

配到最近的簇里. 假设数据服从 Gaussian 分布并具有单位协方差矩阵, 参数集 θ 包含均值 $\mu_1, \dots, \mu_k$, 则参数 θ 的对数似然估计量可以表示为

$$\begin{aligned} \text{Log}(L(\theta; x, t)) &= \log(p(x, t; \theta)) \\ &= \sum_{x_i \in D_l \cup D_u} \log \pi p(x_i | t_i; \theta) \\ &= \sum_{i=1}^{n+m} \log \left(\pi \cdot \frac{1}{(2\pi)^{d/2}} e^{-\|x_i - \mu_{t_i}\|^2} \right) - \\ &\quad \sum_{i=1}^{n+m} \|x_i - \mu_{t_i}\|^2 + r \end{aligned} \quad (3)$$

式中, r 为常数.

由式(3)可知, 模型是否符合数据的实际分布, 将显著影响算法的性能. 对于数据集 $X = \{x_i\}_{i=1}^n$, 可以得到 k 个不相交的 $\{x_h\}_{h=1}^k, \mu_h$ 为划分的中心点, K-means 算法的关键是选择好的初始中心点. 对于满足最小化目标函数 $\|x_i - \mu_{t_i}\|$ 的数据, 根据 K-Means 算法的相似性原则其应具有与簇中心相同的类型标记. 因此, 在同一簇中可扩展标记数据的类型信息来标记簇中的未标记数据的类型. 半监督聚类^[2]利用部分数据的类别标记或约束来指导非监督的聚类过程, 根据标记信息判定实例所属的类别, 而约束指明两个实例应属于相同的簇 must-link 或属于不同的簇 cannot-link^[7]. 假定约束集 $C = (c_{12}, c_{13}, \dots, c_{n-1, n}), c_{ij}$ 取值 $(-1, 0, 1), c_{ij} = 1$ 表示 $(x_i, x_j) \in C_{ML}, C_{ML}$ 是 must-link 约束集, 对于 $(x_i, x_j) \in C_{ML}$ 是指 x_i 和 x_j 必须属于同一个簇, 属于 must-link 约束的数据不能作为新的簇中心; $c_{ij} = -1$ 表示 $(x_i, x_j) \in C_{CL}, C_{CL}$ 是 cannot-link 约束集, $(x_i, x_j) \in C_{CL}$ 表示 x_i 和 x_j 不能共存于同一簇; $c_{ij} = 0$ 表示 x_i 和 x_j 不存在约束关系. C_{ML} 和 C_{CL} 中的数据是无序的, 即 $(x_i, x_j) \in C_{CL} \Rightarrow (x_j, x_i) \in C_{CL}$, C_{ML} 中的数据同样如此.

3 基于主动学习的半监督聚类入侵检测算法

改进的基于主动学习半监督聚类的入侵检测算法 ASCID 的数据集有: 输入数据集 $X = \{x_i\}_{i=1}^n, x_i \in R^d$, must-link 约束集 C_{ML} , cannot-link 约束集 C_{CL} , 簇数目 k_1, k -近邻参数 k_2, K 表示最终得到数据 x 的 k -近邻的集合. 输出为网络数据 $x \in D_u$ 的数据类型(攻击或正常). 算法如下:

(1) 初始化 λ 为近邻集的数目;

(2) 选择第一个标记数据 x , 将其加入到 $N_1, \lambda \leftarrow 1$;

(3) While 查询 and $\lambda < k_1$; 将与已有的近邻集 N 中标记数据最远的数据赋给 x ;

if while 查询 x 与近邻集 N 中的数据的关系, 如果 x 与近邻集 N 中所有标记数据属于 cannot-link 约束关系, 则 $\lambda \leftarrow \lambda + 1$, 生成包含 x 的新近邻集 N_λ ;

else

将 x 赋给与其属于 must-link 约束关系的近邻集. 同时得到 λ 个不相交的近邻集 $\{N_p\}_{p=1}^\lambda$

$\frac{n!}{r!(n-r)!}$, 其中 $\lambda \leq k_1$;

(4) 计算各簇中心为 $\mu_h \leftarrow \frac{1}{|C_h|} \sum_{x \in C_h} x, h = 1, \dots, \lambda$, 其中 λ 为标记数据生成的簇的数目, 并计算各簇的半径, 确定最大半径 $R_{\max}$.

(5) 对于包含标记数据的簇, 根据标记数据的类型确定该簇的类型 $\xi_i = \text{label}_i, 1 \leq i \leq s$, 其中 s 为数据集中已知的数据类型数目, 并将簇的类型 ξ_i 赋予该簇中所有的数据.

(6) 得到标记的簇集 CL 与未标记的簇集 CU , 且 $C = CL \cup CU$.

(7) 对于无法根据标记数据确定类型的数据 x , 其中 $x \in C_r, C_r \subseteq CU$, 根据基于 k -近邻的新攻击检测方法, 确定其类型:

While $|C| > 0$

(7a) 将簇 C_r 中的数据加入集合 P 中, 令 $C \leftarrow C - \{C_r\}$, 针对 C 计算 $d_{\min}$;

(7b) 对于 $\forall x_i \in P$, 如果 $d(x, x_i) < d_{\min}$, 当 $|K| < k_2$ 时, 将 x_1 加入集合 K 中; 当 $|K| = k_2$ 时, 若 $d(x, x_i) < \max_{x_j \in K} d(x, x_j)$, 则从集合 K 中去掉数据 x_j , 将 x_i 加入集合 K 中, 否则, 执行步骤(8);

(7c) 如果 $d(x, x_i) > d_{\min}$, 将集合 C 中距离数据 x 最近的簇 $C_{\min}$ 中的数据加入集合 P 中; $C \leftarrow C - \{C_{\min}\}$, 更新 $d_{\min}$; 执行步骤(7b).

(8) 由集合 K 中的数据进行多数投票, 确定数据 x 的类型 $\text{label}(x)$; $\text{label}(x) =$

$\begin{cases} \text{label}_i, & \text{if } |\text{label}_i| < |\text{label}_j|, 1 \leq i, j \leq s, i \neq j \\ \text{new-attack}, & \text{label}(x_i) \text{ 未知}, x_i \in K \end{cases}$

(9) 重复步骤(7), (8), 直至簇集 CU 为空, 所有未标记数据的数据类型标记完毕.

ASCID 算法从标记数据中初始化 k 个簇中心作为聚类的种子, 充分利用了标记数据中的监督信息, 对数据之间的约束关系进行主动查询, 即使少量的约束也对算法的性能有了明显改善. 通过有限的查询获得数据 (x_i, x_j) 约束的类型 (must-link 或 cannot-link), 目标是通过较少的查询次数而获得比随机选择约束算法更好的聚类结构.

4 实验与分析

为了检验算法的有效性, 本文采用入侵检测领域比较权威的 KDD CUP 1999 数据集^[8]作为测试集. 该数据集包含 4 种攻击类型 DoS 拒绝服务攻击、R2L 对远程主机的未授权访问、U2R 对本地超级用户权限的未授权访问、Probe 扫描与探测行为. 实验时随机抽取 kddcup. data 10 percent. gz 数据集中 10% 的数据, 约 4 940 条数据记录, 同时给出 2 048 条记录作为监督信息. 检测到入侵样本数与入侵样本总数的比值称为检测率 (Detection Rate, DR), 被误报为入侵的正常样本数与正常样本总数的比值称为误报率 (False Positive Rate, FPR). 通过测试算法对攻击的检测率与误报率, 能够很好地度量出算法的检测效果.

比较 ASCID 算法与 K-means、SVM 算法检测攻击的性能. 参数设置: ASCID 算法和 K-means 算法的 $k=100$, SVM 算法参数按缺省设置^[9], 实验结果如表 1 所示.

表 1 入侵检测结果

	算法	检测率/%				
		DOS	U2R	R2L	PROBE	总体
误报率=4%	ASCID	76.17	55.23	25.72	98.61	84.16
	K-means	68.91	50.66	10.16	97.52	71.33
	SVM	93.25	0	35.64	72.35	68.63
误报率=8%	ASCID	76.36	96.72	92.15	99.07	90.57
	K-means	69.56	96.48	91.57	98.66	77.54
	SVM	94.01	0	30.58	73.47	92.45

可以看出, K -means 算法的检测率明显低于 ASCID 算法, 因为 ASCID 算法采用少量的标记数据对聚类过程进行指导, 提高了检测未知攻击的精确度. 当误报率小于 7% 时, ASCID 算法拥有比 SVM 算法更高的总体检测率. ASCID 算法在改进的 k -近邻法半监督聚类的基础上, 采用主动学习策

略能查询未知攻击的类别标记, 从而解决了 SVM 算法对未知攻击类别标记不足的问题, 降低了误报率.

表 2 选取了不同的 K 值, 对 ASCID 算法和 Combined 算法^[10]的已知和未知攻击的入侵检测结果进行比较.

表 2 选取不同 K 值时 2 种算法的入侵检测结果比较

K 值	ASCID 算法				Combined 算法			
	已知攻击/%		未知攻击/%		已知攻击/%		未知攻击/%	
	DR	FPR	DR	FPR	DR	FPR	DR	FPR
20	73.92	4.95	87.72	8.21	67.15	4.91	88.81	9.51
30	86.35	5.25	93.32	9.15	73.63	5.35	92.76	9.86
40	84.82	5.17	95.49	9.49	82.17	5.68	89.03	10.13
50	72.86	4.67	92.83	8.91	79.23	5.18	87.86	9.65
60	68.14	4.19	95.73	10.01	74.13	4.70	93.87	10.07
70	67.51	4.62	90.02	9.69	67.32	4.83	89.93	9.64
平均	75.77	4.81	92.51	9.24	73.94	5.12	90.37	9.81

由表 2 可知, 与 Combined 算法相比, ASCID 算法的平均检测率较高, 而平均误报率较低. 使用了标记数据指导聚类过程, 提高了检测未知攻击的能力, 同时降低了误报率.

5 结束语

ASCID 算法采用合理的主动采样策略, 通过标记数据和未标记数据的联合约束改善分类器的性能, 解决了实际网络环境中标记数据不足的问题, 提高了检测未知攻击类型的精确性. 实验结果表明 ASCID 算法提高了检测率, 降低误报率, 证明算法的可行性和有效性.

参考文献:

[1] Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering[C]// Proceedings of ACM CSS Workshop on Data Mining Applied to Security. USA: Philadelphia, 2001.

[2] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding[C]// Proceedings of 19th International Conference on Machine Learning (ICML-2002). Sydney, Australia, 2002:19-26.

[3] 王汝山, 李永忠, 张念贵, 等. 半监督学习在入侵检测系统中的应用[J]. 广西师范大学学报: 自然科学版, 2009, 27(3):179-182.

[4] 李永忠, 胡翰. 基于主动学习的半监督聚类入侵检测算法[J]. 江苏科技大学学报: 自然科学版, 2010, 24(2):

160-163.

[5] 李永忠, 孙彦. 基于 HMM 和 STIDE 的异常入侵检测方法[J]. 计算机工程, 2008, 34(3):181-182.

[6] Bin Zhang. Generalized K-harmonic means-boosting in unsupervised learning[R]. Hewlett-PASCARD Laboratories, 2000.

[7] Klein D, Kamvar S D, Manning C. From instance-level constraints to space level constraints: Making the most of prior knowledge in data clustering[C]// Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002). Sydney, Australia, 2002:307-314.

[8] The UCI KDD Archive. KDD99 cup dataset [EB/OL]. [2007-10-10]. <http://kdd.ics.uc.edu/databases/kddcup99/kddcup99.html>.

[9] 李雯睿. 基于半监督聚类的入侵检测算法的研究[D]. 开封: 河南大学, 2007.

[10] Shi Zhong, Taghi M Khoshgoftaar, Naeem Seliya. Clustering-based network intrusion detection[J]. International Journal of Reliability, Quality, and Safety Engineering (IJRQSE), 2005, 14(2):169-187.

作者简介:

李永忠 男, (1961-), 教授, 硕士生导师. 研究方向为网络安全、计算机应用、藏文信息处理.

李正洁 女, 硕士研究生. 研究方向为计算机网络安全.

荆春伟 男, (1984-), 硕士研究生. 研究方向为计算机网络安全.