

基于集成 PU 学习数据流分类的入侵检测方法

宋 群, 张 骏, 智永锋

(西北工业大学 自动化学院, 陕西 西安 710072)

摘 要: 入侵检测问题可以模型化为数据流分类问题, 传统的数据流分类算法需要标注大量的训练样本, 代价昂贵, 降低了相关算法的实用性. 在 PU 学习算法中, 仅需标注部分正例样本就可以构造分类器. 对此本文提出一种动态的集成 PU 学习数据流分类的入侵检测方法, 只需要人工标注少量的正例样本, 就可以构造数据流分类器. 在人工数据集和真实数据集上的实验表明, 该方法具有较好的分类性能, 在处理偏斜数据流上优于三种 PU 学习分类方法, 并具有较高的入侵检测率.

关键词: 入侵检测; 集成分类; 数据流; PU 学习

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2013)07-0173-04

Dynamic Classifier Ensemble for Intrusion Detection Based on Positive Unlabeled Learning

SONG Qun, ZHANG Jun, ZHI Yong-feng

(Department of Automation College, NorthWestern Polytechnical University, Xi'an 710072, China)

Abstract: The intrusion detection problem could be modeled as data stream classification problem. For traditional data stream classification algorithms, a lot of labeled training samples should be provided to train the classifier, with a high expense, and decrease the utility of the algorithms. For positive unlabeled learning (PU Learning) based algorithms, only positive and unlabeled samples are required to train classifier. In this paper, we propose to use dynamic classifier ensemble of PU learning algorithm to cope with intrusion detection, which only requires label some positive samples for building data stream classifier. The experiment result on both synthetic and real-life dataset shows that the proposed algorithm has excellent classification performance, and it outperforms 3 PU learning algorithms on data streams with skewed distribution, with high intrusion detection ratio.

Key words: intrusion detection; classifier ensemble; data stream; Positive Unlabeled Learning

1 引言

在网络安全防护中, 入侵检测系统具有动态性、实时性和主动性防御等特点, 是其它静态信息安全防御工具的有效补充, 现已成为信息安全机制当中一个重要的组成部分. 随着数据挖掘技术研究的日益深入, 将其引入到入侵检测研究中, 以提高系统检测性能, 成为了网络信息入侵检测技术研究的重要方向.

PU 学习(Positive and Unlabeled Learning, PU

Learning)是数据挖掘一个重要研究方向, 能够从一部分正例样本(目标样本)和大量未标注样本中学习得分类分类器^[1]. 传统的入侵检测数据流分类算法中, 需要提供入侵、正常两个类别的训练数据, 在实际的网络入侵检测中通常入侵行为相对正常行为比例低, 人工收集、标注海量训练样本开销大、代价昂贵. 将入侵检测模型化为 PU 学习分类问题, 只需人工标注入侵样本作为正例数据, 即可基于正例和未标注样本构造分类器, 进行入侵检测, 可大大降低标注成本.

收稿日期: 2012-10-02; 修回日期: 2012-11-13

基金项目: 国家自然科学基金项目(61201321)

然而在真实的网络信息海量数据中,一般入侵信息数据相对正常数据比例偏低,导致网络入侵数据分布的偏斜,干扰了分类器的分类精度^[2]. 对此本文提出一种动态的集成 PU 学习数据流分类的入侵检测方法 PUDCE (Positive Unlabeled Learning Based Dynamic Classifier Ensemble), 在只标注入侵样本(正例)条件下,在正例和未标注样本中学习得到分类器,可用于网络信息入侵检测. 实验证明该动态集成算法在对入侵数据分布偏斜的情况下进行分类检测时,比 PU 学习的单分类器分类精度高,实用性强.

2 PU 学习数据流集成分类方法

本文选取三种不同的 PU 学习单分类器, Positive Naive Bayes (PNB)、Positive Tree Augmented Naive Bayes (PTAN)、Positive Hidden Naive Bayes (PHNB) 作为基础分类器,其 PU 学习的分类方法借鉴了文献^[2].

算法采用通过计算得到的 ROC 曲线下面积 AUC 的值,比较、分析单个分类器的性能,并作为更新分类器组的标准^[3]. 设训练数据流训练一次可在三种分类算法上得到的分类器分别为 PNB_i 、 $PTAN_i$ 、 $PHNB_i$, 训练数据每个批次为 T_i , 验证数据每个批次为 C_i 算法描述如下:

输入: 训练数据 $D_i = P_i \cup N_i = \{(X_1, y_1), (X_2, y_2), \dots, (X_k, y_k)\}$, 其中 P_i 表示正例, N_i 表示负例, y_k 为类标记; z 为分类器组中单个分类器个数; Q 为选取距离最近相邻的百分数.

输出: 集成的分类器组 E_i .

(1) 训练单个分类器, $T_i = \{PNB_i; PTAN_i; PHNB_i\} = \text{training}(D_i)$

(2) 当 $E_{i-1} = \emptyset$ 时, 设置 $T_i.weight$ 的权重等于 1, 输出 $E_i = T_i$;

(3) 统计集成分类器 E_{i-1} 的个数, 如果小于 Z , 那么 $T_i.weight$ 的权值为 1, $E_i = E_{i-1} \cup T_i$, 否则执行(4);

(4) 取 E_{i-1} 中分类器的 AUC 的最小值 $AUC(T_m) = AUC(PNB) + AUC(PTAN) + AUC(PHNB)$, 如果 $AUC(T_i) > AUC(T_m)$, 那么 $E_i = \{E_{i-1} - T_m\} \cup T_i$, 执行(6), 否则输出 E_i ;

(5) 取验证集 P 、 N 的质心 $Centroid_P$ 和 $Centroid_N$, 其中均方差为验证数据集中数据与两个质心产生, 而后选取距离数值较小的最近的部分数据, 邻域计算有, $Neighbor = \text{FindNeighbors}(P/N,$

$Q)$.

(6) 对于每个 $T_n \in E_i$, 设权值 $T_n.weight = 0$, 分类邻域的每一个数据, 如果分类正确, 那么 $w_n = T_n.weight + \frac{1}{\exp(\text{distnse}(X_n, f))}$, 最终得到分类器集合权重序列, 归一化后, 输出:

$$E_i = \{(T_1, w_1) \dots (T_i, w_i)\}.$$

(7) 输出集成分类器 E_i .

3 实验与分析

基于 Weka3. 5. 8 机器学习平台开发了本研究的集成算法, 在人工数据集和真实数据集上进行了实验. 人工数据集实验中, 采用数据挖掘研究界广泛使用的分类精度、AUC 和 F1 指标来评估分类器的分类性能. 每组实验重复 10 次, 并将 10 次实验数据的平均值作为最终实验结果. 真实数据集实验中, 采用 1999 KDD CUP 中的部分数据用作分类数据集, 并与 Wenke Lee 入侵检测方法相比较.

3.1 人工数据集分类试验

在数据流的模拟中, 本文通过移动一个 d 维的超平面的来实现^[3]. 构造方法如下:

$$\sum_{i=1}^d a_i x_i = a_0 \quad (1)$$

当一个样本满足 $\sum_{i=1}^d a_i x_i > a_0$ 时, 该样本被标注为正例, 否则为负例.

分类概率应用均方差法作为评价标准, 数据流中应用的方差公式为:

$$L = \frac{1}{n} \sum_{i=1}^n (f(x_i) - \rho(+|x_i))^2 \quad (2)$$

其中 $f(x_i)$ 为分类输出的估计概率, $\rho(+|x_i)$ 为真实的分类概率, 计算如下:

$$\rho(+|x) = \frac{1}{1 + \exp(g(x))} \quad (3)$$

由于上述方法, 因此本实验构造了三个离散数据集, 表 1 给出了人工数据集的详细信息. 包括每个数据集的大小, 数据中的包含的属性, 以及数据集中正例与负例的比值.

表 1 人工数据集

数据集	大小	#属性	Pos/Neg
A	3158	20	32%/68%
B	3092	28	10%/90%
C	7926	37	1%/99%

在下面实验中, 数据集的 70% 用做训练集, 30% 用作为测试集^[4], 实验结果如表 2 所示.

表 2 算法分类结果

指标	分类器	A	B	C
F1	PNB	0.8435	0.8172	0.7672
	PTAN	0.8569	0.8287	0.7901
	PHNB	0.8612	0.8299	0.7939
	PUDCE	0.8911	0.8671	0.8361
Accuracy	PNB	83.98	82.33	—
	PTAN	85.62	82.68	—
	PHNB	86.57	83.02	—
	PUDCE	88.73	87.16	—
AUC	PNB	0.8897	0.8132	0.7791
	PTAN	0.8915	0.8227	0.7872
	PHNB	0.9150	0.8291	0.7899
	PUDCE	0.9192	0.8626	0.8315

表 2 中,由于数据集 C 正例比例过少,Accuracy 值不做参考^[2]. 可以看出,随着正例样本比例的减少,算法的分类精度逐渐下降,但 PUDCE 的分类精度变化较舒缓,尤其在正例比例极少的数据集 C 指标数据明显优于其它三种 PU 算法.

对于每个人工数据集 $D=PU N$,其中,P 表示正例样本集,N 表示负例样本集;在 PU 学习场景下训练阶段,正例样本集 P 中随机选取 $1-a\%$ 的样本作为正例样本,构成正例样本集 P' ;其余的 $a\%$ 样本与负例样本集 N 一起构成未标注样本集 U' ;从而构成 PU 学习场景下的训练数据集 $D'=P' \cup U'$ ^[5]. 进而,从未标注样本集 U' 中随机选取样本构成未标注样本集 U'' ,构造 PU 学习场景下的训练数据集 $D''=P' \cup U''$,用 UnLevel 表示 D'' 中未标注样本占的比例 $UnLevel=|U''|/|D''|$.

PUDCE 算法在不同 UnLevel 取值下的分类性能,如表 3 所示.

表 3 UnLevel 值变换实验结果

UnLevel	指标	A	B	C
40%	F1	0.9062	0.8697	0.8368
	Accuracy	88.73	85.58	—
	AUC	0.9125	0.8692	0.8457
50%	F1	0.9011	0.8531	0.8312
	Accuracy	87.63	84.59	—
	AUC	0.9017	0.8576	0.8363
60%	F1	0.8912	0.8391	0.8227
	Accuracy	86.83	83.17	—
	AUC	0.8915	0.8378	0.8252
70%	F1	0.8852	0.8321	0.8190
	Accuracy	84.926	82.65	—
	AUC	0.8893	0.8305	0.8213
80%	F1	0.8643	0.8137	0.8019
	Accuracy	84.18	81.23	—
	AUC	0.8687	0.8172	0.8162

可以看到,随着参数 UnLevel 变大,PUDCE 算法的 F1、Acc 和 AUC 指标总体在降低. 随着 UnLevel 变大,训练数据集中正例样本的量在降低,导致 PUDCE 算法对目标类别进行学习的分类性能逐步降低.

在下面这组实验中,取 $a=0.2,0.4,0.6$ 时,PUDCE 分类指标变化如表 4 所示. 可以看到,随着参数 a 变大,PUDCE 算法的分类性能在逐渐降低,说明训练中正例样本所占的比例和算法的分类精度成正比.

表 4 a 值变换实验结果

指标	a	A	B	C
F1	0.2	0.8862	0.8561	0.8153
	0.4	0.8657	0.8352	0.7982
	0.6	0.8381	0.8101	0.7867
Accuracy	0.2	87.62	85.34	—
	0.4	85.26	82.50	—
	0.6	82.85	80.71	—
AUC	0.2	0.9053	0.8643	0.8275
	0.4	0.8879	0.8431	0.8092
	0.6	0.8695	0.8207	0.8031

3.2 入侵数据集的分类实验

本文的入侵数据来自于 1999 KDD CUP 入侵检测评估数据集,是目前入侵检测研究中使用较多的数据源之一^[6],包含有四类入侵数据,分别是: R2L(未授权远程主机访问)、DoS(服务拒绝攻击)、Probing(扫描和探测行为)和 U2R(未授权超户权限访问)^[7].

实验构造了 A、B 两组数据集,A 组从 1999 KDD CUP 数据集中抽出 7625 条数据,4125 条为数据为训练数据,2300 条为正常访问数据,包含 R2L 攻击数据 125 条,DoS 攻击数据 860 条,Probing 攻击数据 735 条,U2R 攻击数据 105 条,与正常数据混合;测试数据中正常数据为 3100,四类入侵数据各 100 条. B 组数据集在 A 组数据基础上把入侵数据所占比例进一步降低,训练数据中正常数据保持不变,去掉 DoS 和 Probing 两类攻击数据,测试数据集中正常数据不变,入侵数据只包含 R2L 和 U2R 数据各 50 条. 采用 DR(检测率)和 FPR(误检率)两个指标来评价算法的性能^[8],取 $UnLevel=30\%$, $a=0.3$ 时,与 Wenke Lee 的入侵检测指标比较如表 5 所示.

从表 5 可以看出,PUDCE 在入侵数据比例较低的情况下,检测率优于 Wenke Lee 方法,并且误

检率较低,证明了算法的有效性.

表 5 入侵检测比较表

数据集	PUDCE		Wenke Lee	
	DR	FPR	DR	FPR
A	91.72%	1.51%	82.37%	——
B	88.96%	1.73%	81.35%	——

4 结束语

本文针对 PU 学习分类算法在正例样本比例低时分类精度不高的问题,提出一种动态的集成 PU 学习数据流分类的入侵检测方法 PUDCE. 在只标注入侵样本(正例)条件下,在正例和未标注样本中学习得到分类器,可进行入侵行为识别. 在人工数据集和真实数据集上的实验表明,PUDCE 算法效果优于单个 PU 分类器,并在正例样本极低的条件下具有良好的分类效果,更加适合网络入侵检测中实际应用.

参考文献:

- [1] 康世瑜. 基于数据挖掘和特征选择的入侵检测模型[J]. 微电子学与计算机,2011(8):52-56.
- [2] Jiazhen He, Yang Zhang, Xue Li, et al. Bayesian classifiers for positive unlabeled learning[C]//Proceedings of the 12th international conference on Web-age information management, Berlin, Heidelberg: Springer-

Verlay,2011:81-93.

- [3] Yi Wang, Yang Zhang. Mining data streams with skewed distribution by static classifier ensemble[J]. Studies in Computational Intelligence, 2009(214):65-71.
- [4] Calvo B, Larranaga P, Lozano J A. Learning bayesian classifiers from positive and unlabeled examples[J]. Pattern Recognition Letters,2007, 28(16):2375-2384.
- [5] Zhengzheng Xing, Jian Pei, Eamonn Keogh. A brief survey on sequence classification[J]. SIGKDD Explorations Newsletter, 2010,12(1):40-48.
- [6] 田俊峰,刘仙跃. 入侵检测数据分类模型—PCANN[J]. 微电子学与计算机,2007(9):35-39.
- [7] 尹向东. 基于正态云模型的网络入侵检测[J]. 微电子学与计算机,2010(10):21-25.
- [8] Wenke Lee,Stolfo J. A data mining framework for constructing features and models for intrusion detection systems[D]. York: Columbia University New ,1999, 135-160.

作者简介:

宋 群 男,(1984—),博士研究生. 研究方向为数据挖掘、系统工程.

张 骏 男,(1964—),博士,教授,博士生导师. 研究方向为控制理论、系统工程等.

智永锋 男,(1978—),博士,讲师. 研究方向为信号与系统.

(上接第 172 页)

- [3] 李宏,阿玛尼,李平,等. 基于 EM 和贝叶斯网络的丢失数据填充算法[J]. 计算机工程与应用,2010,46(5):123-125.
- [4] JinlongSHI,ZhigangLUO. Missing value estimation for DNA microarray gene expression data with principal curves;2010[C]//International Conference on Bioinformatics and Biomedical Technology. Haesun Park: IEEE,2010:262-265.
- [5] XiaofengZhu, ShichaoZhang. Missing value estimation for mixed-attribute data sets[J]. IEEE transactions on knowledge and data engineering,2011,23(1):110-121.
- [6] 刘星毅,农国才. 几种不同缺失值填充方法的比较[J]. 南宁师范高等专科学校学报,2007,24(3):148-150.
- [7] 吕萍,钱进,王波,等. 一种快速差别矩阵属性约简算法

[J]. 计算机工程与应用,2010,46(20):164-167.

- [8] 刘星毅,檀大耀,曾春华,等. 基于马氏距离的缺失数据填充算法[J]. 微计算机信息,2010(9):225-226.

- [9] 杨涛,罗嘉伟,王艳. 基于马氏距离的缺失值填充算法[J]. 计算机应用,2005,25(12):2838-2871.

作者简介:

陈志奎 男,(1968—),博士,教授. 研究方向为无线认识网络、路由算法、数据挖掘.

吕爱玲 女,(1989—),硕士研究生. 研究方向为大规模异构网络不完全信息填充.

张清辰 男,(1988—),博士研究生. 研究方向为大规模数据处理、数据挖掘.