半监督学习在网络入侵分类中的应用研究*

赵建华[1,2]

(1. 西北工业大学 计算机学院, 西安 710072; 2. 商洛学院 计算机科学系, 陕西 商洛 726000)

摘 要: 为了解决网络环境中已标记入侵数据获取代价大的问题,将半监督学习引入网络入侵分类领域。根据网络攻击类型的不同,将少量的已标记入侵数据分为三部分,分别作为最初的训练集训练分类器,形成三个差异较大的初始化分类器。通过三个分类器协同学习,实现对未标记入侵数据进行标记。详细介绍了使用 KDD Cup 99 数据集构造半监督分类实验数据集的过程。实验结果表明,半监督学习能有效地挖掘未标记入侵数据信息,具有较高的入侵分类率。

关键词: 半监督学习; 协同训练; 入侵分类; 标记; KDD Cup 99 数据集

中图分类号: TP393.08 **文献标志码**: A **文章编号**: 1001-3695(2014)06-1874-03

doi:10.3969/j.issn.1001-3695.2014.06.064

Network intrusion classification based on semi-supervised learning

ZHAO Jian-hua[1,2]

(1. *College of Computer, Northwestern Polytechnical University, Xi'an* 710072, *China*; 2. *Dept. of Computer Science, Shangluo University, Shangluo Shaanxi* 726000, *China*)

Abstract: In order to solve the problem that it costs too much to obtain labeled intrusion data in the network environment, semi-supervised learning is applied into the field of network intrusion. According to the different types of network attack, this paper divided the limited labeled intrusion data into three equal training sets to form three different classifiers. Through training learning by three single classifiers, the unlabeled samples were labeled. It introduced the process of using KDD Cup 99 data sets to construct semi-supervised classification experiment data sets. The experimental results show that semi-supervised learning can effectively dig the unlabeled samples information of intrusion data and has a higher rate of intrusion classification.

Key words: semi-supervised learning; collaborative training; intrusion classification; labeling; KDD Cup 99 date set

入侵检测作为一种动态的网络安全技术,能够全面监控计算机网络或主机上的各种应用程序。依据智能的安全策略对大量数据进行审计分析,对系统中的大量入侵行为进行主动的识别和响应,有效地保障了系统的安全,已成为网络安全领域的一个研究热点[1]。早期的入侵检测算法是基于监督学习的,即要求所有的训练数据样本都是有类别标记的,这种监督学习方法检测率较高,但是无法有效地检测到未知攻击,且要求训练集中的数据被正确地标记为正常或者异常。然而在现实网络环境中,存在大量的未标记数据,获取标记数据难度较大、代价较高,要为所有数据作出标记几乎是不可能的[2]。

半监督分类[3,4]利用大量无标记数据扩大分类算法的训练集,主要研究从有监督学习的角度出发, 当已标记训练样本不足时, 如何利用大量未标记样本信息辅助分类器的训练。目前常见的半监督分类方法很多, 包括基于 EM 算法的生成式模型参数估计法[5]、协同训练方法[6~9]、基于流形的半监督分类方法[10]等。为了解决入侵检测领域中标记样本获取难度较大的问题,有不少研究者将半监督学习方法引入到入侵检测领域,并取得了较好的分类效果[11,12]。

本文将协同半监督分类算法 Co-S3OM[6] 应用到入侵检测领域中,结合实际应用,提出具体的半监督网络入侵数据分类方案。

1 相关知识

1.1 SSOM

自组织特征映射(SOM)[13]神经网络是由芬兰学者 Kohonen 于 1981 年提出的,是一种无监督聚类方法,它能将输入模式在输出层映射成一维或二维离散图形,识别环境特征并自动聚类,具有强大的模式识别能力。能够自适应环境变化,较好处理复杂的非线性问题,而且具有较好的稳健性和潜在的容错性,可获得很高的识别率。SOM 的结构包括输入层和竞争层两层结构,第一层为输入层,输入层的维数与输入样本向量维数一致,第二层为竞争层,竞争层节点一般呈二维阵列分布,一个竞争层节点代表一个神经元。

SSOM(supervised SOM)[14]在 SOM 两层结构的基础上,增加了第三层,即输出层,变成有监督的神经网络,其结构如图 1 所示。输出层的个数同数据分类类别一致,每个输出节点代表一种数据类别。在进行网络学习训练时,根据每个输入样本的预测类别和实际类别是否相等,选取不同的权值调整公式对权值进行调整。在 SSOM 中,不仅调整 SOM 中输入层节点和竞争层优胜节点领域内的权值,同时调整输出节点和竞争层优胜节点领域内的权值。通过两个权值的组合,便可以很容易实现

收稿日期: 2013-06-11; **修回日期**: 2013-07-25 **基金项目**: 陕西省教育厅科研计划资助项目(12JK0748); 商洛学院教育教学改革资助项目(13jyjx111); 商洛学院科研项目(10SKY1001)

作者简介: 赵建华(1982-),男,陕西商洛人,讲师,博士研究生,主要研究方向为机器学习、网络安全(zhaojh2009@aliyun.com).

对输入样本的类别进行分类和标记,具有较高的分类率。

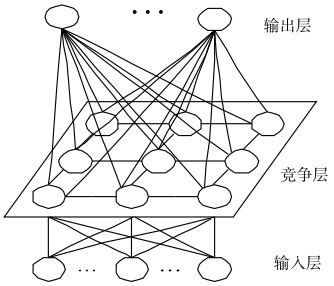

图1　SSOM结构

1.2　Co-S3OM

Co-S3OM (coordination semi-supervised SOM) 算法[6] 将 SOM 神经网络引入到半监督分类中,采用 SSOM 作为分类器实现对样本的训练和预测。Co-S3OM 既不要求充分冗余视图,也不要求使用不同类型分类器。Co-S3OM 将标记样本三等分为无重复的三个训练集,每个训练集使用有监督 SSOM 进行训练,生成三个分类器。各个分类器获得的新标记示例都由三个分类器协作投票提供,即如果三个分类器预测结果一致,才把数据及其类别标记加入训练集。在扩充新样本及其类别标记的过程中,每次将标记样本只添加到一个训练集,三个训练集轮流获取新增加的标记样本,能保证各训练集训练样本至始至终没有重复性,从一定程度保证了三个分类器的差异性。

使用 Co-S3OM 算法进行分类,最重要的是增强协同训练过程中三个分类器之间的差异性,以便提高分类性能。

2　基于 Co-S3OM 的网络入侵分类

网络入侵检测实际上是一个分类问题,把检测数据分为正常数据和异常数据。但是入侵检测中需要分类的数据比较复杂,往往呈现高维性和不可分性。将 SOM 应用到入侵检测领域,能较好地解决入侵分类问题[14]。

本文将协同半监督分类算法 Co-S3OM 应用到网络入侵检测领域,解决有标记样本较少的网络入侵检测问题。用 Co-S3OM 解决入侵检测问题,主要工作包括:a)选择有效的标记样本完成三个 SSOM 分类器训练集的初始化工作,保证三个分类器的差异性;b)通过三个 SSOM 分类器的协同工作,实现对未标记样本的标记;c)扩充三个分类器的训练集,将新标记的样本增加到三个分类器中,反复迭代,形成最终的分类器进行分类。其具体操作步骤如图 2 所示。

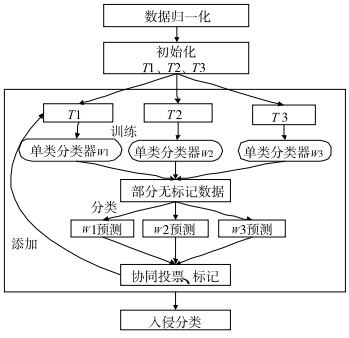

图2　基于半监督学习的网络入侵分类

1)数据归一化　实验选取 KDD Cup 99 数据集进行入侵检测分类实验,将实验数据分为训练集和测试集两个集合。训练集分为已标记集和未标记集,为了方便统计分类准确率,测试集全部使用已标记样本。未标记集样本中,每条数据有 41 个不同的属性(其中 32 个连续属性和 9 个离散属性);已标记样本中每条数据共有 41 个不同的属性和 1 个攻击类型标签。

分别提取训练集和测试集中每条数据的输入特征值和对应的类型标签,实验中对离散属性进行数字编号,实现数值化。比如将攻击类型分为正常数据和攻击数据,分别用 1 和 2 进行数值化,对协议 TCP、UDP、ICMP 等属性分别用 1、2、3 等进行数值化,对服务类型 aol、auth、…、whois 等使用 1、2、…、68 等进行数值化,对 flag 中的类型名称 OTH、REJ、…、SH 等使用 1、2、…、11 等进行数值化。为了取消各维数据间数量级差别,避免因为输入输出差别较大而造成网络预测误差较大,采用式(1)所示数据归一化函数对输入特征值进行归一化,将其归一化到[0,1]之间。在这里,为了保证具有较高的测试率,本文把训练集和测试集放在一块作为一个整体数据集进行归一化:

$$x_k = (x_k - x_{\min})/(x_k - x_{\max}) \tag{1}$$

式中:$x_{\min}$表示数据中的最小值, $x_{\max}$表示数据中的最大值。

2)初始化　初始化时,最重要的工作是如何在已有的入侵数据集选取有效的已标记样本集,将选取的已标记样本分为三等分,分别作为三个 SSOM 分类器的初始化训练集。为了保证三个 SSOM 分类器对未标记样本进行投票且具有较高的正确率,应该使得这三个 SSOM 分类器训练集的样本具有很高的差异性、多样性和不同性。在网络入侵检测中,由于攻击类型较多,本文从多种攻击数据中选取一类攻击类型数据完成对一个 SSOM 分类器的初始化。比如让 DoS 攻击的数据实现对 SSOM 分类器 1 的初始化,R2L 攻击实现对 SSOM 分类器 2 的初始化,Probe 完成对 SSOM 分类器 3 的初始化。这样,三个 SSOM 分类器的训练集完全不一致且差异性非常大,训练 SSOM 神经网络生成的网络模型也不一致。经过三个分类器的投票,实现对未标记样本的标记,标记准确率较高。

3)数据标记　该过程主要使用半监督分类算法 Co-S3OM 实现对未标记样本进行标记,挖掘未标记样本的隐含信息,扩充有标记样本的数目。通过初始化,三个 SSOM 分类器具有各自的标记样本训练集 T_1、T_2 和 T_3。使用三个训练集分别训练分类器 W_1、W_2 和 W_3,三个 SSOM 分类器协同工作,三个 SSOM 分类器对某个未标记样本预测一致时,使用该预测标记对预测样本进行标记,将其添加到有标记样本集中,不断扩充有标记样本的数目形成新的训练集。

在 KDD Cup 99 数据集上进行实验时,由于每条样本有 41 个特征属性信息,第 42 个属性为样本的类别,如果三个 SSOM 预测某个未标记样本为正常样本,则直接给该样本的第 42 个属性标记为 1;如果预测为异常样本,则直接给第 42 个属性标记为 2;如果预测不一致,选择未标记集中的下一个样本进行预测标记。对于新增加的标记样本,每次只将其依次添加到 T_1、T_2、T_3 之一,形成新的标记样本训练集,保证三个分类器的训练样本不一致。在标记样本训练集更新后,重新使用 T_1、T_2、T_3 训练生成新的分类器 W_1、W_2 和 W_3,使用新的分类器反复迭代,直到未标记样本集为空为止。

4)入侵分类　将扩充后的三个 T_1、T_2 和 T_3 进行合并,作为最终的入侵检测训练集,训练 SSOM 神经网络,生成最终分类器,使用该分类器实现对测试数据的测试。

3　实验

实验平台选用 Intel Core2 Duo CPU 2.0 GHz、内存 2.0 GB 的

PC、安装 Windows XP 操作系统和 MATLAB 7.8.0（R2009.0a）编程环境。实验数据采用 KDD Cup 99 数据集，该数据集包括大量的正常网络流和各种攻击。攻击可以归类为以下四类：

a）DoS。拒绝服务攻击，如 SYN Flood、land 攻击等。

b）R2L。非授权的远程访问，如口令猜测等。

c）U2R。各种权限提升，如各种本地和远程 Buffer Overflow 攻击等。

d）PROBE。各种端口扫描和漏洞扫描。

实验从 KDD Cup 99 的“10% KDD”数据集中随机抽取 500 条正常数据和 500 条攻击类型数据作为未标记训练数据集，从 KDD Cup 99 的“10% KDD”数据集中随机抽取 36 条数据作为已标记训练数据集，从 KDD CUP 99 的“Corrected KDD”数据集中随机抽取 500 条正常数据和 500 条攻击类型数据作为测试数据集。对于已标记样本训练集，将其分为三份，各自作为三个 SSOM 分类器的初始化训练集，按照选取数目从少到多的顺序分别进行三次实验。每次实验中，各个分类器进行初始化时标记样本训练集的组成如表 1 所示。比如，第 1 次实验中，三个分类器都选取 5 个已标记样本作为训练集进行实验。其中分类器 1 的训练集由 2 个正常样本和 3 个 DoS 类攻击样本组成；分类器 2 的训练集由 2 个正常样本、2 个 R2L 类攻击样本和 1 个 U2R 类攻击样本组成；分类器 3 由 2 个正常样本和 3 个 Probe 类攻击样本组成。由于每个分类器的训练集都来自不同的攻击类别，三个分类具有明显的差异性。

表 1 实验数据集

实验	分类器	normal	DoS	R2L	U2R	Probe
实验 1	分类器 1	2	3			
	分类器 2	2		2	1	
	分类器 3	2				3
实验 2	分类器 1	3	6			
	分类器 2	3		3	3	
	分类器 3	3				6
实验 3	分类器 1	6	6			
	分类器 2	6		3	3	
	分类器 3	6				6

$$\text{rate1} = \frac{\text{标记正确的入侵数据样本的个数}}{\text{新标记的入侵数据样本的个数}} \times 100\% \quad (2)$$

$$\text{rate2} = \frac{\text{对测试集样本分类正确的入侵数据数目}}{\text{测试集的入侵数据样本总数}} \times 100\% \quad (3)$$

每次实验按照式(2)和(3)分别计算正确标记率和正确分类率，其中 rate1 表示未标记样本的正确标记率，rate2 表示对测试集样本的正确分类率。每种实验各进行 3 次，然后求平均值作为最终结果，实验结果如表 2 所示。图 3 为随着训练集数据的不断增长时，rate1 和 rate2 的变化曲线。

表 2 实验结果

实验	SSOM	Co-SSOM	
	rate2/%	rate1/%	rate2/%
实验 1	54.50	67.83	72.63
实验 2	57.03	77.45	79.03
实验 3	66.90	82.50	89.00

通过表 2 和图 3 可以看出：a）由于 Co-SSOM 实现了对标记样本的扩充，其分类率较仅仅使用初始化标记样本训练集的 SSOM 大大提高；b）随着初始化训练集中，已标记样本的数目不断增加，正确标记率和入侵检测率都不断提高。可见，基于 Co-SSOM 的半监督入侵检测方法能有效地提高网络的入侵检测分类率。

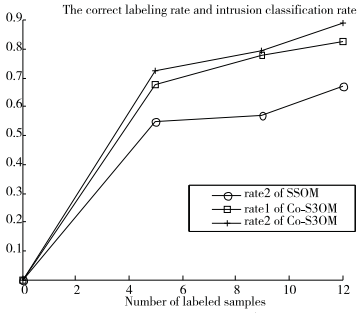

图3 实验结果对比

4 结束语

本文将半监督学习算法 Co-S3OM 应用到入侵检测领域，提出具体的半监督网络入侵分类方案。根据不同的攻击类型选取不同的标记样本作为训练集，完成三个 S3OM 分类器的初始化，通过三个分类器协同投票的方法扩充已标记样本数据。详细地给出使用 KDD Cup 99 数据集进行半监督分类的实验构造过程。实验表明，基于 Co-S3OM 的入侵分类模型具有较高的数据样本标记率和较高入侵分类率。

参考文献：

[1] WU Qing-tao, SHAO Zhi-qing. Survey on intrusion detection techniques[J]. Application Research of Computers, 2005,22(12): 11-44.

[2] 夏战国，万玲，蔡世玉，等. 一种面向入侵检测的半监督聚类算法[J]. 山东大学学报：工学版，2012，42(6)：1-6.

[3] ZHU Xiao-jing. Semi-supervised learning literature survey[R]. Madison: University of Wisconsin, 2008.

[4] 李昆仑，曹铮，曹丽苹，等. 半监督聚类的若干新进展[J]. 模式识别与人工智能，2009，22(5)：735-742.

[5] CHAPELLE O, ZIEN A. Semi-supervised classification by low density separation [C] // Proc of the 10th International Workshop on Artificial Intelligence and Statistics. 2005: 57- 64.

[6] 赵建华，李伟华. 一种协同半监督分类算法 Co-S3OM[J]. 计算机应用研究，2013，30(11)：3237-3239.

[7] ZHOU Zhi-hua, LI Ming. Tri-training: exploiting unlabeled data using three classifiers [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(11): 1529-1542.

[8] 于重重，商利利，谭励，等. 一种增强差异性的半监督协同分类方法[J]. 电子学报，2013，41(1)：36- 41.

[9] ZHANG Min-ling, ZHOU Zhi-hua. CoTrade confident co-training with data editing [J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(6): 1612-1626.

[10] MENG Jun, WU Li-xia, WANG Xiu-kun. Granulation-based symbolic representation of time series and semi-supervised classification [J]. Computers and Mathematics with Applications, 2011, 62 (9): 3581-3590.

[11] PACHGHARE V K, KHATAVKAR V K, KULKARNIP. Performance analysis of semi-supervised intrusion detection system [J]. International Journal of Computer Applications, 2011, NSC(4): 15-19.

[12] LI Yong-zhong, LI Zheng-jie, WANG Ru-shang. Intrusion detection algorithm based on semi-supervised learning[C]//Proc of International Conference of Informat ion Technology, Computer Engineering and Management Sciences. [S. l.]: IEEE Press, 2011: 153-156.

[13] HAGAN M T, DEMUTH H B, BEALE M H. Neural network design [M]. Beijing: China Machine Press, 2002.

[14] 赵建华，李伟华. 有监督 SOM 神经网络在入侵检测中的应用[J]. 计算机工程，2012，38(12)：110-111.