一种新的半监督入侵检测方法

梁 辰 李成海
(空军工程大学防空反导学院 西安 710051)

摘 要 针对基于监督的入侵检测算法在现实网络环境中通常面临的训练样本不足的问题,提出了一种基于纠错输出编码的半监督多类分类入侵检测方法。该方法综合 cop-kmeans 算法的半监督思想,挖掘未标记数据中的隐含关系,扩大有标记正常网络数据的数量。该算法首先采用 SVDD 计算入侵检测各类别的可分程度,从而得到由不同子类构成的二叉树;然后分别对二叉树的各层节点进行编码并形成层次输出编码,得到最终的分类器。实验表明,该算法对各种类型的攻击具有更高的检测率,在现实网络环境中具有较好的实用性。

关键词 入侵检测系统,纠错输出编码,半监督聚类,类间可分性,支持向量数据描述
中图法分类号 TP393.08 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.5.016

Novel Intrusion Detection Method Based on Semi-supervised Clustering

LIANG Chen LI Cheng-hai
(School of Air & Missile Defense, Air Force Engineering University, Xi'an 710051, China)

Abstract A new semi-supervised intrusion detection method based on error-correcting output codes was proposed to solve the difficulties which existing in intrusion detection algorithms based on supervised learning usually face when the training samples are insufficient. This method mines the relationship under the unlabeled data to enlarge the known labeled normal data by introducing the idea of semi-supervised cop-kmeans algorithm. Firstly, the SVDD is used to measure the class separabilty quantitatively. Then the inter-class separability matrix is got gradually. The binary tree is built based on the matrixes from the bottom to the up. Each node of the binary tree is encoded by level to get the final hierarchical error-correcting output codes and classifiter. The experiments in KDD Cup 1999 network data sets prove that the method has better performance in detection accuracy and good adaptability in the real network environment.

Keywords Intrusion detection system, Error-correcting output codes, Semi-supervised clustering, Class separability, SVDD

入侵检测(Intrusion Detection)是一种动态的网络安全技术,它建立在入侵行为与系统行为不同这一假设基础上,通过分析网络流量或系统审计记录等实时发现网络或系统中是否有违反安全策略的攻击行为,对可能危害到系统机密性、完整性和可用性的行为进行响应和拦截[1]。

自 1987 年 D. E. Denning[2] 提出入侵检测的概念至今,入侵检测技术引起了各界广泛关注,各种基于机器学习的入侵检测方法被应用到该领域中,如 Mukkamala[3] 与 Lee[4] 等人提出的基于监督学习的检测方法,以及 Portnoy[5] 与 Depren[6] 等人提出的基于非监督学习的检测方法等。基于监督学习的入侵检测方法具有较高的检测精度,但要求所有的训练样本数据带有先验信息,面临着训练样本不足的问题。基于非监督学习的入侵检测方法根据数据的相似性进行分组,克服了监督学习过程中需要标记数据的不足,但其由于学习的结果只能完成聚类,缺乏训练过程,难以达到有监督检测算法的检测精度。Fiore[7] 与阳时来[8] 等人采用半监督学习方法,利用少量标记的网络数据及数据之间的约束关系,指导海量的非标记原始网络数据的聚类过程,并依据有标记数据对聚类结果进行类型识别,一定程度上克服了监督学习和非监督学习的弊端,解决了训练样本不足的问题,并能够基本适应网络数据随时间变化的特点。已有的基于半监督学习的入侵检测方法的精度和异常数据区分度仍然可以进一步得到提高。

综合上述对基于机器学习的入侵检测方法的研究,提出了一种基于纠错输出编码的半监督多类分类入侵检测方法(Semi-Supervised Error-correcting Output Codes, SS-ECOC)。文中采用基于支持向量数据描述(Support Vector Domain Description, SVDD)的层次纠错输出编码方法[9],利用少量有标记正常网络数据样本协助建立 SVDD 模型以获得类间可分性度量准则[10],找出最优的类别组合并对其标注,据此构建层次编码输出,构造一个基于入侵检测数据样本的分类器。纠错输出编码的应用不但提高了异常数据区分度,能够及时检测出系统或网络遭受的入侵行为,而且提高了算法的检测精度。

到稿日期:2016-01-07 返修日期:2016-02-03 本文受基于 SVM 集成和证据理论的多传感器目标识别技术研究(60975026),基于多特征融合和集成学习的多目标识别技术研究(61273275)资助。

梁 辰(1992—),男,硕士生,主要研究方向为网络信息安全,E-mail:3659442@qq.com;李成海(1966—),男,硕士,教授,主要研究方向为网络信息安全。

1 相关工作

半监督学习方法作为有监督和无监督方法的折中，同时利用有标记数据和无标记数据，只需要较少的人工参与就能获得更精确的学习精度，在相关研究领域受到了越来越多的关注[11]。

目前，半监督思想多用于图像分割及信息检索等领域。根据半监督学习算法的学习方式，大致可将现有的常用的半监督学习算法分为3类。第一类算法以生成式模型为分类器，以样本的标签对于样本的输入特征的条件概率建模，然后利用EM算法进行标签估计和计算模型参数，如Chapelle等人提出的基于核的半监督学习方法[12]。第二类算法是基于图正则化框架的半监督学习算法，如Kulis等人提出的SS-Kernel-Kmeans算法[13]。第三类算法是协同训练(Co-Training)算法，此类算法隐含地利用了聚类假设或流形假设，要求问题有充分冗余的两个视图，在两个视图上分别训练出一个分类器，然后挑选若干置信度高的实例加入对方的训练集中，以提高分类精度，最后选取置信度大的分类器对新的实例进行分类。

文献[14]提出的cop-kmeans算法是一种比较典型的半监督算法，Wagstaff等人根据应用领域中的背景知识提出了将Must-Link与Cannot-Link两类约束作为先验知识，具有Must-Link关系的两个数据应该归为一类，而具有Cannot-Link关系的两个数据不能划归一类，据此选择与待测数据欧氏距离最小的聚类作为其归属。

ECOC子类划分会潜在地将样本划分得不平衡，本文借鉴cop-kmeans算法的半监督思想，将其引入到ECOC编码中，以获得尽可能最优的子类划分，具有Cannot-Link关系的子类之间相关性较小，易于分类；而具有Must-Link关系、相关性较大的原始数据类别将被分为同一子类。基于此类划分构造的基分类器分类难度小，能达到较高的准确率，从而实现了分类效果的整体提高。

同时，基于特征空间几何距离的方法对样本数据的充分性和分布先验知识的要求不高，可以较快地进行子类划分。文中使用的基于距离测度的SVDD学习过程仅仅需要“目标类”样本，很好地解决了分类中样本不足或者难以获得目标样本带来的学习问题[15]。

2 基于纠错输出编码的半监督多类分类入侵检测方法

纠错输出编码继承了纠错码特有的纠错能力，将复杂的多类问题分解为多个简单的二类分类任务，同时对由二类分类器产生的错误具有一定的纠错能力，是一种分而治之的多类解决方案，具有极好的容错性、稳定性和准确性，同时也具有更强的异常数据区分度。本节利用SVDD作为类间可分性度量，融入半监督思想，获得最优子类类别划分。而后根据子类划分自下而上构建二叉树，对二叉树中每层节点进行编码，最终得到基于入侵检测样本的层次编码矩阵，并据此构建分类器。

2.1 基于SVDD的可分性度量

2.1.1 支持向量数据描述

支持向量数据描述(SVDD)的基本思想是在特征空间中设法找到一个半径尽可能小的超球体来包含尽可能多的样本数据。并根据样本数据建立一个目标样本尽可能地被该超球体覆盖，而非目标样本点尽可能在此超球体之外的区域，从而实现两类分类。

为了使优化区域更加紧凑，采用满足Mercer条件的核函数$K(x,y)$替代高维空间中的内积运算，使得$K(x,y)=\langle\Phi(x),\Phi(y)\rangle$，引入松弛变量$\xi_i$，得到约束条件：

$$\min_r r^2+C\sum_i\xi_i$$

$$\text{s.t. } \|\boldsymbol{x}_i-\boldsymbol{o}\|^2\leqslant r^2+\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,N$$

采用Lagrange乘子，将问题转化为对偶问题：

$$\max\sum_i\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i,j}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$s.t.\ 0\leqslant\alpha_i\leqslant C,\sum_i\alpha_i=1,j=1,2,\cdots,N$$

其中使$0<\alpha_i\leqslant C$的样本点被称为支持向量。

对于未知样本x而言，设

$$f(x)=\|\boldsymbol{x}-\boldsymbol{o}\|^2$$

$$=K(\boldsymbol{x},\boldsymbol{x})-2\sum_{i=1}^{N}\alpha_iK(\boldsymbol{x},\boldsymbol{x}_i)+\sum_i^N\sum_j^N\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

当它满足到超球体中心的距离小于或等于r时，即$\|\boldsymbol{x}-\boldsymbol{o}\|^2\leqslant r^2$时，则未知样本被判为目标类，否则为非目标类。

2.1.2 类间可分性度量

假设入侵检测是一个k类分类问题，对于训练样本集$\{X_1,\cdots,X_k\}$，$X_i=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N_i}\}$，$i=1,\cdots,k$，采用核函数，分别用不同类别的少量有标记网络数据构造SVDD模型，同时对大量无标记网络数据进行聚类，利用cop-kmeans方法挖掘无标记样本中的隐含关系，不断对超球面进行调整，将无标记样本中被识别出的同类别网络数据和初始有标记网络数据一起作为新的训练集，得到球面集合：$S=\{S_1,\cdots,S_k\}=\{(r_1,\boldsymbol{c}_1),(r_2,\boldsymbol{c}_2),\cdots,(r_k,\boldsymbol{c}_k)\}$，其中$(r_i,\boldsymbol{c}_i)$表示第$i$个超球面的半径和球心。计算不同类别超球体两两之间的球心距离作为类间可分性准则：

$$D=\begin{bmatrix}d_{11}(t) & \cdots & d_{1k}(t)\\ \vdots & \ddots & \vdots\\ d_{k1}(t) & \cdots & d_{kk}(t)\end{bmatrix}$$

$$d_{ij}=\frac{\|\boldsymbol{c}_i-\boldsymbol{c}_j\|}{r_i+r_j}$$

其中，对类间可分性度量矩阵D描述：(1)对角线元素为0；(2)$d_{ij}=d_{ji}$，$i,j=1,\cdots,k$；(3)当$d_{ij}\geqslant 1$时，说明该元素对应的两个类别在特征空间没有交集(d_{ij}越大，区分度越好)。当$0<d_{ij}<1$时，说明该元素对应的两个类别在特征空间的距离定义上相交(d_{ij}越大，区分度越差，即可分性越差)。当$d_{ij}=0$时，说明该元素对应的两个超球体在特征空间中完全重叠。

2.2 基于SVDD的层次编码矩阵构造

根据2.1节所得类间可分性度量准则对多类入侵检测数据进行分类，构造层次编码矩阵，其步骤如下：首先，将每个目标类视为一个子类类群，然后利用SVDD综合cop-kmeans半监督思想，计算类间可分性度量矩阵D，将相交程度最高的两个子类(即除去同类间距离度量值d_{ii}后，d_{ij}的最小值所对应的两类)合并成一个子类，重新计算两个子类合并后的子类与剩余其他类别之间的可分性度量矩阵，再将相交程度最高的两个子类进行合并，直到得到与其他子类在特征空间没有交集的若干类别。

对于k类入侵检测问题而言，上述过程构成了一个倒二

叉树 T。最后，对二叉树中每层节点进行编码，以得到一个基于 k 类入侵检测样本的层次编码矩阵：

$$M(i,j)=\begin{cases}1, & X_i\in T^j_{left}\\ -1, & X_i\in T^j_{right}\\ 0, & X_i\notin T^j\end{cases} \qquad (*)$$

其中，$M(i,j)$表示编码矩阵的第 i 行第 j 列码值，T^j_{left} 和 T^j_{right} 分别为二叉树 T 中第 j 层节点的左右树枝。

2.3 基于 SS-ECOC 的入侵检测算法

基于 SS-ECOC 的入侵检测算法的描述如下。

输入：少量的有标记正常网络数据样本集 L 和大量的无标记网络数据样本集 U，$L=L_1\cup L_2\cup\cdots\cup L_k$。其中网络数据 x 满足 $x\in L\cup U$。

输出：网络数据 x 的类型。

(1)对 $L_i(i=1,2,\cdots,k)$进行聚类，创建 k 个簇，使得同一个簇 $C_h(h=1,2,\cdots,k)$中只包含相同类型的数据；

(2)利用(1)中类簇在对应空间中求解包含所有样本的最小覆盖球 S_i；

(3)随机取无标记样本集 U 中的网络数据 x^*。$\forall x^*\in U$，当它到超球体 S_m 中心的距离满足小于或等于 r_m 时，即 $\|x^*-o_m\|^2\leqslant r_m^2$ 时，则未知样本 x^* 被判为目标类 m，否则判为非目标类；

(4)计算类间可分性度量矩阵 D，将相交程度最高的两个子类簇合并成一个类簇；

重复(4)直到得到 k 个与其他类簇在特征空间没有交集的类簇 C_h^*，根据其中包含的标记数据确定该簇的类型；

(5)自上而下构造二叉树，利用式(*)对其进行编码，得到最终的层次输出编码矩阵。构造一个基于 k 类入侵检测样本的分类器，并输出 x^* 的类型。

3 实验

入侵检测实验采用 KDD CUP 1999 数据集[16]在 Matlab R2014a 平台上进行。实验中采用的解码策略为 Hamming 距离解码，采用支持向量机(多项式核函数，$C=2$)作为基分类器。

3.1 数据描述

由于 KDD 数据集过于庞大，实验中选用 8 类共 32000 条数据组成训练数据集，其中正常数据 20000 条，7 类攻击数据共计 12000 条，每类数据中有标记数据所占比例为 3%，另选用包含训练数据集中 8 种数据类型的 19000 条数据作为测试数据集。所用数据集如表 1 所列。

表 1 KDD CUP99 10%数据集攻击行为及分布

数据类型 Data Type	数量	
	训练数据集 Training Data	测试数据集 Test Data
normal(0)	20000	10000
smurf(1)	3000	2000
satan(2)	1500	1000
portsweep(3)	1000	1000
neptune(4)	3000	2000
back(5)	1500	1000
ipsweep(6)	1000	1000
warezclient(7)	1000	1000
Total	32000	19000

KDD 数据集中每个网络连接实例用 41 个特征进行描述，但并不是所有的属性都适用于入侵检测，且其中部分属性包含冗余信息。对网络数据的 41 个属性进行主成分分析(PCA)，得到 28 个可以完全描述原数据的属性，如表 2 所列。

表 2 实验中使用的网络数据的属性

No.	Feature name	Description	Data type
1	duration	Length(number of seconds) of the connection	continuous
2	protocol_type	Type of the protocol, e. g. , tcp or udp	discrete
3	service	Network service on the destination, e. g. , http or telnet	discrete
4	flag	Normal or error status of the connection	discrete
5	src_bytes	Number of data bytes from source to destination	continuous
6	dst_bytes	Number of data bytes from destination to source	continuous
7	land	1 if connection is from/to the same host/port; 0 otherwise	discrete
8	wrong_fragment	Number of "wrong" fragments	continuous
9	urgent	Number of urgent packets	continuous
10	count	Number of connections to the same host as the current connection in the past two seconds	continuous
11	srv_count	Number of connections to the same service as the current connection in the past two seconds	continuous
12	serror_rate	% of same-host connections that have "SYN" errors	continuous
13	srv_serror_rate	% of same-service connections that have "SYN" errors	continuous
14	rerror_rate	% of same-host connections that have "REJ" errors	continuous
15	srv_rerror_rate	% of same-service connections that have "REJ" errors	continuous
16	same_srv_rate	% of same-host connections to the same service	continuous
17	diff_srv_rate	% of same-host connections to different services	continuous
18	srv_diff_host_rate	% of same-service connections to different hosts	continuous
19	dst_host_count	Count of connections having the same destination host	continuous
20	dst_host_srv_count	Count of connections having the same destination host and using the same service	continuous
21	dst_host_same_srv_rate	% of connections having the same destination host and using the same service	continuous
22	dst_host_diff_srv_rate	% of different services on the current host	continuous
23	dst_host_same_src_port_rate	% of connections to the current host having the same src port	continuous
24	dst_host_srv_diff_host_rate	% of connections to the same service coming from different hosts	continuous
25	dst_host_serror_rate	% of connections to the current host that have an S0 error	continuous
26	dst_host_srv_serror_rate	% of connections to the current host and specified service that have an S0 error	continuous
27	dst_host_rerror_rate	% of connections to the current host that have an RST error	continuous
28	dst_host_srv_rerror_rate	% of connections to the current host and specified service that have an RST error	continuous

3.2 有标记数据集扩展的实验与分析

首先，分别测试 SS-ECOC、SVM 和 cop-kmeans 算法在有标记样本数不同时的入侵检测性能，以评价半监督算法是否有助于提高入侵检测精度。实验中，从 corrected 数据集中随机抽取 20000 条连接记录组成测试集，使其包含 3.1 节中所述训练测试集所有的 8 种类型的网络数据。分别对 0.05%、0.1%、0.5%、1%、10%、30%及 50%的测试样本进行标记，取 10 次运行算法的平均值作为结果，如图 1 所示。

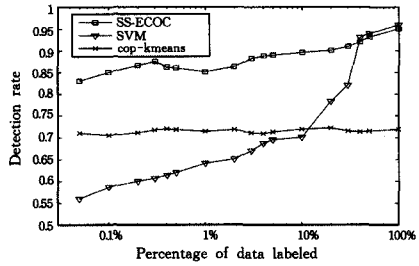

图 1 给定不同比例的标记数据时 SS-ECOC、SVM 及 cop-kmeans 算法的比较结果

从图 1 可以看出，当标记数据很少时，SS-ECOC 算法明显优于 SVM 算法和 cop-kmeans 算法。其中，当标记样本的比例为 0.05%时，SS-ECOC 算法的检测精度要高于 SVM 算法 27%，高于 cop-kmeans 算法 12%。只有当有标记数据的比例达到 30%时，SVM 算法才接近于 SS-ECOC 算法的检测性能。而在标记数据的比例超过 50%时，SVM 算法的检测率稍高于 SS-ECOC 算法的。这是因为当标记数据很少时，训练数据不足，SVM 算法分类精度降低，而 SS-ECOC 算法利用半监督思想对有标记数据集进行了扩展，提高了分类模型的泛化精度，所以其检测率明显高于 SVM 的。当标记数据很多时，由于 SVM 算法的训练集中近乎均为正确标记的训练样本，而 SS-ECOC 算法的训练集则是由半监督聚类方法生成，包含部分错误标记数据，导致算法精度低于 SVM 算法精度。

3.3 SS-ECOC 算法的实验与分析

采用 3.1 节所述训练数据集和测试数据集，对经典编码方法与本文方法进行比较，讨论编码的分类效果和有效性。在选择两种随机编码方法时，分别从已产生的 2000 个密集及稀疏随机编码阵集（对应各码元概率分别为：$p(-1)=0.5$，$p(+1)=0.5$；$p(-1)=1/3$，$p(0)=1/3$，$p(+1)=1/3$）中随机选择所需要的编码阵。最后将基于无监督的 SVM 方法与本文方法以及文献[8]中基于半监督 GHSOM 的入侵检测方法在上述相同实验条件下进行比较，讨论算法在分类器选择上的优劣。两组实验均取 10 次运行算法的平均值作为结果。

其中，SVDD 的核函数采用高斯核函数 $K(\boldsymbol{x},\boldsymbol{y})=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\sigma^2)$，通过交叉验证法选择惩罚系数 C 和高斯核函数的自由参数 σ^2。在估计检测率时为保证估计的准确性，样本数据个数大于 500 时采用 10 重交叉验证，小于 500 时采用 5 重交叉验证来进行，并利用双边估计 t 检验法来计算置信水平为 0.95 的分类错误率置信区间作为最终结果，计算公式如下：

$$\frac{|\bar{x}-\mu|}{\sigma/\sqrt{n}}\geqslant t_{0.025}(n-1)$$

其中，μ,σ 分别表示 n 重交叉验证的均值和标准差。

3.3.1 分类结果比较

表 3 列出了在 8 种不同编码方法下以 SVM 作为基分类器、采用 Hamming 距离解码方法的分类结果。

表 3 基于 SVM 和 Hamming 距离解码的检测率及置信水平为 0.95 的置信区间(%)

	one vs all	one vs one	dense	sparse	DECOC	SECOC	SS-ECOC
KDD CUP99	77.14±11.08	75.65±6.04	79.22±5.09	81.64±9.67	78.49±18.20	73.20±7.43	88.20±6.38
	8×8	8×28	8×10	8×10	8×7	12×15	8×7

从表 3 的结果来看，由 SS-ECOC 获得的编码矩阵在大部分情况下的分类精度要优于其他经典的事前编码或部分基于数据编码方法。同时，SS-ECOC 编码矩阵长度要比其他编码方法的较短，这是因为从初始数据集开始，进行了类间可分性比较，在获得可分性矩阵的前提下，对二叉树从上至下进行了编码，使得对于 N 类样本数据，其编码矩阵为 $N\times N-1$，这与 DECOC 编码的码字长度类似，但分类精度表现更好；同时针对其他的编码方法，尤其是事前编码，SS-ECOC 不仅能在促进多类分类的实际效果的情况下提高编码的纠错性能，而且能获得更加紧凑的编码，大大缩减了训练和测试时间，提高了编码解码的速度。

3.3.2 各算法检测率比较

表 4 列出了在相同实验条件下 SVM 算法、半监督 GHSOM 算法及 SS-ECOC 算法的检测率情况。通过对比可以看出，在数据集有标记数据所占比例为 3%的入侵检测实验中，相对于 SVM 算法，采用 SS-ECOC 算法后，总体检测率由 68.22%提高到 88.2%。多数类型（normal，satan，ipsweep，portsweep，back）的检测率均有不同程度的提高，其中对 ipsweep、satan 类型的检测率提高较大；少数类型（smurf，neptune）的检测率变化不明显，原因是这些类型在无监督条件下训练精度已经很高，标记数据的引导作用不明显；个别类型（warezclient）的检测率有所下降，原因是 SVM 算法对该攻击类型数据的检测率本来就很低（0.564%），在训练过程中，一部分数据被错误聚类到 normal 类型中，致使部分该类型数据在检测中被误判为 normal 类型，导致检测率下降。从以上分析可以得知，SS-ECOC 算法对那些使用 SVM 算法获得一定检测率且有较大改进空间的攻击类型具有较好的改进效果；对那些原本已具有较高检测率的攻击类型，改进效果不明显；对原有检测率较低的个别攻击类型，可能会有一定的负面影响。总体上，相对于有监督的 SVM 算法，SS-ECOC 算法明显具有更好的检测性能。

表 4 各算法分类准确率对比

Data Type	Correctedly Detected Data Number			Detection Rate		
	SVM	Semi-Supervised GHSOM	Semi-Supervised ECOC	SVM	Semi-Supervised GHSOM	Semi-Supervised ECOC
normal	6395	9261	9291	0.6395	0.9261	0.9291
smurf	2000	2000	2000	1	1	1
satan	571	832	840	0.571	0.832	0.840
portsweep	713	787	810	0.713	0.787	0.810
Neptune	1583	1734	1732	0.7915	0.8670	0.8660
back	698	833	868	0.698	0.833	0.868
ipsweep	437	641	675	0.437	0.641	0.675
warezclient	564	423	542	0.564	0.423	0.542
Total	12961	16511	16758	0.6822	0.8690	0.8820

（下转第 121 页）

sessment Model Under Current E-commerce [D]. Hefei: Hfei University of Technology, 2012: 19-36(in Chinese)
卫志诚. 当前电子商务中风险类型分析及评估模型研究[D]. 合肥:合肥工业大学, 2012: 19-36

[13] Zhang Jie, Cohen R. A Framework for Trust Modeling in Multiagent Electronic Marketplaces with Buying Advisors to Consider Varying Seller Behavior and the Limiting of Seller Bids [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(2): 55-73

[14] Liu Dao-qun, Liu Jun. Dynamic Trust Model for P2P Network Security[J]. Journal of Chongqing Institute of Technology(Natural Science), 2009, 23(11): 90-94(in Chinese)
刘道群, 刘军. 一种 P2P 网络安全动态信任模型[J]. 重庆工学院学报(自然科学), 2009, 23(11): 90-94

(上接第 90 页)

相对于半监督的 GHSOM 算法,SS-ECOC 算法在入侵检测中检测率略有提高。其中,对大部分类型的检测率均有提升,在 warezc-lient类型的检测率上提升较为明显。原因是在半监督 GHSOM 算法中,前期落在某些神经元上的 warezc-lient类型的数据数量较 SVM 算法增多,后期大量与 warezc-lient特征较为类似的 normal 类型数据落在这些神经元上将 warezclient 类型的数据淹没掉,导致该神经元所代表的类型被认为是 normal 类型,从而使得检测过程中会有更多的 warezclient 类型的数据被误判为 normal 类型,而在这一过程中,半监督 GHSOM 算法无法自发调整获得神经元标记类型,导致该类型检测率较 SVM 算法大幅下降。SS-ECOC 算法则在对两类相似子类进行合并的过程中,利用 SVDD 作为相关性度量依据,有一个小幅度调整的过程,所以在 warezc-lient类型的检测率上较有监督的 SVM 算法有所下降,但优于半监督 GHSOM 算法。同时,SS-ECOC 算法作为多类分类算法,在异常数据区分度上优于其他大部分半监督的入侵检测方法,具有更低的算法复杂度。

结束语 本文针对多类分类网络入侵检测问题,提出了一种基于半监督纠错输出编码方法的解决办法。借鉴了 cop-kmeans 算法中的半监督思想,在训练数据中引入了有标记数据。在训练过程中,从如何构造基于入侵检测数据的编码矩阵入手,从类间可分性准则出发利用 SVDD 作为其度量依据,首先利用该准则综合半监督思想找出最相似的两类进行合并,从而使相关性较大的子类划分在一起,依此类推,直到所有子类合并为一个类。然后从下至上建立二叉树,对二叉树的每层节点进行编码,从而获得层次纠错输出编码,最终构建一个基于入侵检测数据样本的分类器。在实验中发现,SS-ECOC 算法在有效提高分类准确率的同时,在获得标记困难的网络环境下,相对于传统无监督算法有更优的处理小样本标记数据的能力,具有很好的实用性。

参 考 文 献

[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey [J]. ACM Computing Surveys, 2009, 41(3): 75-79

[2] Denning D E. An intrusion detection model[J]. IEEE Transactions on Software Engineering, 1987, SE-13(2): 222-232

[3] Mukkamala S, Sung A H. Feature ranking and selection for intrusion detection systems[C]// Proc of the 11th Int'l Conf. on Information and Knowledge Engineering. Las Vegas: CSREA Press, 2002: 503-509

[4] Lee W K, Stolfo S J. A framework for constructing features and models for intrusion detection systems[J]. ACM Transactions on Information and System Security, 2000, 3(4): 227-261

[5] Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering[C]// Proc of ACM CSS Workshop on Data Mining Applied to Security. New York: ACM Press, 2001: 51-62

[6] Depren O, Topallar M, Anarim E, et al. An intelligent intrusion detection system(IDS) for anomaly and misuse detection in computer networks[J]. Expert Systems with Applications, 2005, 29: 713-722

[7] Fiore U, Palmieri F, Castiglione A, et al. Network anomaly detection with the restricted boltzmann machine[J]. Neurocomputing, 2013, 122: 13-23

[8] Yang Shi-lai, Yang Ya-hui, Shen Qing-ni, et al. A Method of Intrusion Detection Based on Semi-Supervised GHSOM[J]. Journal of Computer Research and Development, 2013, 50(11): 2375-2382(in Chinese)
阳时来, 杨雅辉, 沈晴霓, 等. 一种基于半监督 GHSOM 的入侵检测方法[J]. 计算机研究与发展, 2013, 20(11): 2375-2382

[9] Lei Lei, Wang Xiao-dan, Luo Xi, et al. Hierarchical error-correcting output codes based on SVDD[J]. Systems Engineering and Electronics, 2015, 37(8): 1916-1921(in Chinese)
雷蕾, 王晓丹, 罗玺, 等. 基于 SVDD 的层次纠错输出编码研究[J]. 系统工程与电子技术, 2015, 37(8): 1916-1921

[10] Li Jun-li, Li Wei-hua. Semi-supervised SVDD-KFCM Algorithm and its Application in Bearing Fault Detection[J]. Computer Science, 2015, 42(6A): 134-137(in Chinese)
李军利, 李巍华. 一种半监督 SVDD-KFCM 算法及其在轴承故障检测中的应用[J]. 计算机科学, 2015, 42(6A): 134-137

[11] Chen Shi-guo, Zhang Dao-qiang. Experimental Comparisons of Semi-Supervised Dimensional Reduction Methods[J]. Journal of Software, 2011, 22(1): 28-43(in Chinese)
陈诗国, 张道强. 半监督降维方法的实验比较[J]. 软件学报, 2011, 22(1): 28-43

[12] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]// Proc of the 10th International Workshop on Artificial Intelligence and Statistic. Barbados. 2005: 19-26

[13] Kulis B, Basu S, Dhillon I, et al. Semi-supervised graph clustering: a kernel approach[J]. Machine Learning, 2009, 74: 1-22

[14] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge[C]// Proc of ICML'01. San Francisco: Morgan Kaufmann Publishers. 2001: 577-584

[15] Zhu Xiao-kai, Yang De-gui. Multi-Class Support Vector Domain Description for Pattern Recognition Based on a Measure of Expansibility[J]. Acta Electronica Sinic, 2009, 37(3): 464-469(in Chinese)
朱孝开, 杨德贵. 基于推广能力测度的多类 SVDD 模式识别方法[J]. 电子学报, 2009, 37(3): 464-469

[16] The UCI KDD Archive. KDD99 Cup Dataset[DB/OL]. (1999-10-28). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html