计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):210−217 [doi: 10.15888/j.cnki.csa.007667]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于受限玻尔兹曼机的电力信息系统多源日志综合特征提取①

刘冬兰[1], 孔德秋[2], 常英贤[3], 刘　新[1], 马　雷[1], 王　睿[1]

[1](国网山东省电力公司电力科学研究院, 济南 250003)
[2](国网山东省电力公司经济技术研究院, 济南 250001)
[3](国网山东省电力公司, 济南 250000)
通讯作者: 刘冬兰, E-mail: liudonglan2006@126.com

摘　要: 为了充分利用电力信息系统中的异构数据源挖掘出电网中存在的安全威胁, 本文提出了基于受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 的多源日志综合特征提取方法, 首先采用受限玻尔兹曼机神经网络对各类日志信息进行规范化编码, 随后采用对比散度快速学习方法优化网络权值, 利用随机梯度上升法最大化对数似然函数对 RBM 模型进行训练学习, 通过对规范化编码后的日志信息进行处理, 实现了数据降维并得到融合后的综合特征, 有效解决了日志数据异构性带来的问题. 通过在电力信息系统中搭建大数据威胁预警监测实验环境, 并进行了安全日志综合特征提取及算法验证, 实验结果表明, 本文所提出的基于 RBM 的多源日志综合特征提取方法能用于聚类分析、异常检测等各类安全分析, 在提取电力信息系统中日志特征时有较高的准确率, 进而提高了网络安全态势预测的速度和预测精度.

关键词: 电力信息系统; 受限玻尔兹曼机; 特征提取; 神经网络; 对比散度快速学习; 随机梯度上升法

引用格式: 刘冬兰,孔德秋,常英贤,刘新,马雷,王睿.基于受限玻尔兹曼机的电力信息系统多源日志综合特征提取.计算机系统应用,2020,29(11):210–217. http://www.c-s-a.org.cn/1003-3254/7667.html

Multi-Source Log Comprehensive Feature Extraction Based on Restricted Boltzmann Machine in Power Information System

LIU Dong-Lan[1], KONG De-Qiu[2], CHANG Ying-Xian[3], LIU Xin[1], MA Lei[1], WANG Rui[1]

[1](State Grid Shandong Electric Power Research Institute, Jinan 250003, China)
[2](Economic & Technology Research Institute, State Grid Shandong Electric Power Company, Jinan 250001, China)
[3](State Grid Shandong Electric Power Company, Jinan 250000, China)

Abstract: In order to excavate security threats in power grid by making full use of heterogeneous data sources in power information system, this study proposes a multi-source log comprehensive feature extraction method based on Restricted Boltzmann Machine (RBM). Firstly, the RBM neural network is used to normalize coding all kinds of log information. Then, the contrast divergence fast learning method is used to optimize the network weight, and the stochastic gradient rise method is used to maximize the logarithmic likelihood function for the training and learning of the RBM model. The data dimension reduction is realized by processing the normalized coded log information. At the same time, the comprehensive features are obtained, which can effectively solve the problems caused by the heterogeneity of log data. The big data threat early warning monitoring experimental environment was set up in the power information system, and the comprehensive feature extraction and algorithm verification of the security log were carried out. Experimental results

① 基金项目: 国网山东省电力公司科技项目 (520626200013)
Foundation item: Scientific Research Program of State Grid Shandong Electric Power Company (520626200013)
收稿时间: 2020-04-04; 修改时间: 2020-04-28; 采用时间: 2020-05-10; csa 在线出版时间: 2020-10-29

show that the proposed method can be applied to all kinds of security analysis, such as clustering analysis, anomaly detection, etc., and it has high accuracy in extracting log features in power information system, which improves the speed and accuracy of network security situation prediction.

Key words: power information system; restricted boltzmann machine; feature extraction; neural network; contrast divergence fast learning; stochastic gradient rise method

1 引言

在能源与电力系统领域, 电力大数据具有数据规模大、数据类型多等多个特征, 因此, 对数据进行采集存储分析也比较困难[1]. 大数据挖掘分析及应用的关键要素是保障数据的真实性. 各类网络安全设备的日志信息反映的是用户对网络及业务系统的访问情况, 通过对各类安全日志信息进行深度挖掘, 能够分析出网络中的恶意攻击, 能够对公司安全隐患及时进行隐患消缺.

电力公司各单位为了确保公司网络安全稳定运行, 通过在网络出入口处部署入侵检测系统、入侵防御系统、防火墙等安全防护设备, 从而更好地保障公司内部网络安全性. 在设备运行过程中, 若有网络访问或攻击等行为, 设备都会通过安全日志记录下来, 从而达到实时监控网络攻击的效果. 由于各类网络安全设备功能有相似的地方, 因此不同设备产生的安全日志会存在较高的重复率, 网络管理人员很难找出日志隐藏的关联性, 从而对网络态势进行融合分析就相对比较困难[2]. 但是, IDS、IPS 和防火墙等设备日志间缺乏协同机制, 其语义级别低, 设备日志信息结构多种形式并且数据分散在不同系统, 包含的安全日志信息相互隔离, 形成信息孤岛, 管理人员不能及时发现网络攻击并快速响应. 因此, 需要有一种方法来从各类日志信息中提取综合特征, 从而帮助管理人员从总体上把握信息系统的安全态势. 例如, 基于这些特征可以在宏观层面进行未知威胁检测等工作.

目前, 在网络设备日志融合方面主要有基于逻辑关系的算法和基于规则推理的算法等[2–4]. 基于逻辑关系的算法核心思想是通过采用常规经验来设计逻辑规则, 进而对日志信息内在的逻辑关系进行融合处理[3]. 基于规则推理的算法核心思想是通过量化评估多源日志信息的不确定性, 进而利用基于规则推理的思想去预测威胁[3]. 这些算法都需要一定程度的先验知识和领域专家知识. 在面向大量异构设备和不断演化的网络环境时, 适配性问题较为突出.

酆勇等[5]提出了一种基于受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 来表达输入语音超向量的说话人信息. 陈龙等[6]提出了一种融合多源日志的基于事件“前提/结果”因果关系的事件场景关联方法. 江雨燕等[7]提出了基于 RBM 的分布式主题特征提取模型可以更好地使用文档中的多标记信息. 程乐峰等[8]总结了深度学习、对抗学习和集成学习等 7 种代表性机器学习在电力系统调度优化和控制决策等方面的应用.

人工神经网络具有较强的非线性映射能力, 特别是受限玻尔兹曼机 (RBM) 神经网络具有较强的自编码能力[9–17]. 前期有学者对文本特征提取[18–20]进行了相关研究. 本文提出了基于受限玻尔兹曼机的多源日志信息综合特征提取方法, 采用受限玻尔兹曼机神经网络对各类日志信息进行融合处理, 有效解决了日志数据的异构性等问题, 提升了电力信息系统安全态势预测速度和精度.

2 基于 RBM 的多源日志综合特征提取方法

2.1 电力信息系统数据采集预处理

电力信息系统数据采集预处理, 通过采集电力信息系统中各个设备包含的历史数据和实时数据的日志信息, 日志信息中包括设备状态信息、动态传输数据信息、安全防护信息及故障信息.

首先获取电力信息系统中各类设备, 例如安全设备、网络设备、主机及其他安全防护系统产生的日志信息, 并对采集到的原始数据进行实时的预处理和分析, 对原始数据的预处理包括数据去重、数据噪声去除等. 数据去重是确保所采集的数据是可信数据, 将源数据中的无关数据和噪声数据去除. 经过预处理的数据进行分布式存储, 对所有存储的数据创建数据索引, 以便后续查询追溯使用.

2.2 多源日志综合特征提取方法的思想

对于每类日志信息 r, 构建初始化受限波尔兹曼机神经网络 RBM_r, 其中 r 为 1 与 t 之间的正整数; t 为日

志信息的类数, 记为{ 日志 1, 日志 2, ⋯, 日志 t }, 其数据维度分别为{M_1, M_2, ⋯, M_t}. 其中, 日志信息的类别是按照设备划分的, 不同的设备是不同的类别; 维度表示日志数据的字段数.

如图 1 所示, 初始化受限波尔兹曼机神经网络 RBM_r 为具有可见层和隐藏层的两层网络, 将不同长度的日志信息数据输入后将其编码为长度为 N 的数据并输出, 可见层节点数与输入数据的维度相对应为 M_r 个, 隐藏层节点数为 N 个, 随机产生 [0,1] 之间的随机数作为 RBM 可见层节点和隐藏层节点之间的连接权值.

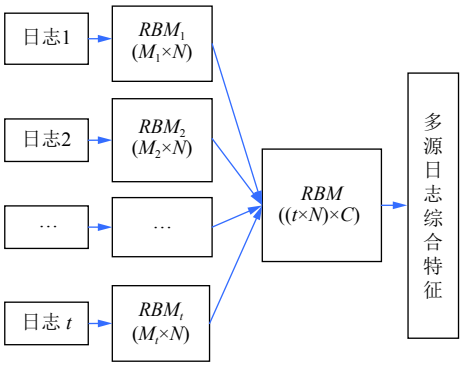

图 1 基于 RBM 的多源日志信息综合特征提取方法示意图

受限玻尔兹曼机 (RBM) 是一个随机神经网络,它包含一层可见层和一层隐藏层. 在 RBM 神经网络中通常事先设定隐藏单元数, 可见单元数赋值为训练数据的特征维数. 隐藏单元数目的设定, 通常采用训练集乘以单个数据的比特数, 进而采用低一个数量级的值设定为隐藏单元的数量[21]. 由于电力信息系统中数据冗余度较高, 因此可以使用更少一些的隐藏单元. 在本文方案中, N 小于 M_r 的 1/2 以上. N 的大小上限是与数据维度相关, 本文设定 N 的取值小于所有 M_r 的一半.

对于每类日志 r, 训练相应的受限玻尔兹曼机神经网络 RBM_r. 训练的输入数据为日志 r 的数据, 从受限玻尔兹曼机神经网络 RBM_r 的可见层神经元输入数据, 根据对比散度快速学习方法优化网络权值, 由此得到稳定的 RBM_r. 神经网络系统概率分布越集中, 则系统的能量越小. 能量函数的最小值, 对应着系统的最稳定状态. 通过调整网络的权值和偏置值使得网络对该输入数据的能量最低. 稳定状态是指当前的神经网络具有最小的能量.

2.3 训练学习 RBM 模型

在 RBM 模型训练学习中, 采用对比散度快速学习方法和随机梯度上升法最大化对数似然函数对 RBM 模型进行训练学习, 通过对规范化编码后的日志信息进行处理. 对比散度快速学习方法优化网络权值的过程为: 可见层 v 和隐藏层 h 的神经元数目分别设为 n 和 m, a 和 b 分别为可见层和隐藏层的偏置向量, W 为 v 和 h 之间的权值矩阵[22,23]. RBM 对应的图是一个二分图, 层与层之间全部相连, 但层内的神经元之间没有连接线. 原始特征向量作为最底层的神经元的输入, 进而向 RBM 网络从下往上传递, 最后将提取到的特征向量转化成抽象的特征向量并对数据进行降维处理[23].

对于可见层 v 和隐藏层 h, v_i 表示第 i 个可见单元的状态, h_j 表示第 j 个隐藏单元的状态; 从受限玻尔兹曼机神经网络可见层神经元输入数据, 根据神经元 v_i 更新隐藏层神经元 h_j 的状态; 再由隐藏层神经元 h_j 重构出可见层神经元 v_i 的状态, 接着再由重构出的可见层神经元 v_i 的状态再重构出隐藏层神经元 h_j 的状态, 完成一次受限玻尔兹曼机神经网络训练学习过程, 直到神经网络具有最小的能量值.

RBM 是一种基于能量的模型, 所以我们可以使用能量函数来描述. 对于给定的状态 (v, h), RBM 具备的能量为:

$$E(v,h|\theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i W_{ij} h_j \quad (1)$$

其中, $\theta = \{W_{ij}, a_i, b_j\}$是 RBM 的参数, 均为实数, 为把 W, a, b 的所有分量拼起来得到的长向量, W_{ij}是可见单元 i 与隐藏单元 j 之间的连接权重, a_i是可见单元 i 的偏置, b_j是隐藏单元 j 的偏置. 基于能量函数可以得到 (v, h) 的联合概率分布:

$$P(v,h \mid \theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \quad (2)$$

其中, $Z(\theta) = \sum_{v,h} e^{-E(v,h|\theta)}$为归一化因子.

由于 RBM 不同层的单元之间有连接, 而层内单元之间无连接. 因此, 当对可见单元的状态赋予确定的数值时, 各隐藏单元之间的激活状态相互独立[24–28]. 因此, 第 j 个隐藏单元的激活概率为:

$$P(h_j = 1 \mid v,\theta) = \omega\left(b_j + \sum_i v_i W_{ij}\right) \quad (3)$$

当对隐藏单元的状态赋予确定的数值时, 各可见

单元的激活状态也相对独立, 则第 i 个可见单元的激活概率为:

$$P(v_i = 1 \mid h,\theta) = \omega\left(a_i + \sum_j h_j W_{ij}\right) \tag{4}$$

其中, $\omega(x) = \dfrac{1}{1+\exp(-x)}$是 Sigmoid 激活函数.

训练学习 RBM 的任务是计算参数θ值, 进而模拟出给定的训练数据, 保持能量$E(v,h|\theta)$守恒. 通过采用最大化 RBM 在训练集上的对数似然函数可计算出参数θ, 如下:

$$\theta^* = \arg\max_\theta \zeta(\theta) = \arg\max_\theta \sum_{t=1}^{T} \log P\left(v^{(t)} \mid \theta\right) \tag{5}$$

其中, T 为包含的样本数.

为得到参数的最优值, 计算θ^*的最大值$\zeta(\theta)$采用随机梯度上升法进行计算:

$$\zeta(\theta) = \sum_{t=1}^{T} \log P\left(v^{(t)} \mid \theta\right) \tag{6}$$

由于电力信息系统获得的观测数据 (即训练样本数据) 的特征维度通常较高, 因此对 RBM 的训练效率有更高的要求. 对比散度方法是一种快速学习方法, 方法处理开始时, 初始化一个训练样本并作为可见单元的状态输入, 随后再根据上述式 (3) $P\left(h_j = 1 \mid v,\theta\right) = \omega\left(b_j + \sum_i v_i W_{ij}\right)$计算所有隐藏单元的二值状态[24]. 当隐藏单元的所有状态值都确定后, 再根据式 (4) $P(v_i = 1 \mid h,\theta) = \omega\left(a_i + \sum_j h_j W_{ij}\right)$计算第 i 个可见单元 v_i 等于 1 的概率, 生成可见层的重构. 此时训练数据值时我们采用随机梯度上升法最大化对数似然函数, 可见层和隐藏层的权重调整方式及噪声控制参数a_i、b_j的调整方式为:

$$\Delta W_{ij} = \varepsilon\left(\left\langle v_i h_j \right\rangle_{\text{data}} - \left\langle v_i h_j \right\rangle_{\text{recon}}\right) \tag{7}$$

$$\Delta a_i = \varepsilon\left(\langle v_i \rangle_{\text{data}} - \langle v_i \rangle_{\text{recon}}\right) \tag{8}$$

$$\Delta b_j = \varepsilon\left(\left\langle h_j \right\rangle_{\text{data}} - \left\langle h_j \right\rangle_{\text{recon}}\right) \tag{9}$$

其中, ε是学习率, $\langle\cdot\rangle_{\text{recon}}$是表示进一步重构后模型定义的分布上的数学期望, $\langle\cdot\rangle_{\text{data}}$是训练数据集所定义的分布上的数学期望, $\left\langle v_i h_j \right\rangle_{\text{data}}$是可见层神经元与隐藏层神经元在输入数据下的二进制状态乘积, $\left\langle v_i h_j \right\rangle_{\text{recon}}$是可见层神经元与隐藏层神经元在重构数据下的二进制状态乘积.

当输入 v 时, 利用 $p(h|v)$ 能计算出隐藏层 h; 当计算出 h 时, 采用 $p(v|h)$ 又能计算出可视层, 通过不断调整参数, 使从隐藏层计算出的可视层 v_1 与最初的可视层 v 相同, 则计算出的隐藏层即是可视层另外一种描述, 所以可以把隐藏层当作可视层输入数据的特征.

2.4 对比散度快速学习方法的优势

本文在 RBM 模型训练学习中, 采用对比散度快速学习方法优化网络权值, 相对于其他典型网络权值优化方法[29–35] 来说, 本文方法具有如下优势.

(1) 对比散度快速学习方法克服了传统误差反向传播算法易陷于局部极值的问题, 是一种不依赖标签的无监督学习方法, 能够在无监督方法中自动从原始数据中学习特征.

(2) 对比散度快速学习算法可以高效训练结构简单的马尔可夫随机模型 (包括 RBM), 每次训练只需要进行 k 次 (一般就是 1 次) 状态转移, 从而极大提升了训练效率.

(3) 在对比散度快速学习中, 重复训练过程通过不断更新参数, 最后就能高效率地完成模型训练.

(4) 对比散度快速学习算法使推荐模型能够包括异构内容信息, 例如文本、图像、音频甚至视频, 具有从多种来源学习的表现形式的潜力.

(5) 对比散度快速学习技术具有很高的灵活性, 特别是随着许多流行的深度学习框架的出现, 如 Tensorflow, Keras, Caffe, MXnet, DeepLearning4j, PyTorch, Theano 等. 这些工具大多以模块化方式开发并具有活跃的社区和专业的支持, 良好的模块化使开发更有效率.

(6) 对比散度快速学习算法可以很容易地将不同的神经网络结构组合起来以形成强大的混合模型, 因此, 我们可以轻松地构建混合和复合推荐模型, 以同时捕获不同的特征和因素.

2.5 数据降维方法及其优势

本文 RBM 模型训练过程中, 通过利用随机梯度上升法最大化对数似然函数对 RBM 模型进行训练学习, 通过对规范化编码后的日志信息进行处理, 实现了数据降维并得到融合后的综合特征, 可以有效解决日志数据异构性带来的问题. 本文研究中采集的数据流量是公司网络出口处的全部数据流量, 包括各种安全日志、系统日志等信息, 需要处理的数据是多维的, 算法的时间复杂度与维数成指数级增加, 因此就需要进行降维处理[36–38].

在特征降维技术中主成分分析方法 (Principal Component Analysis, PCA) 是最为经典和实用的特征降维技术, 主成分分析的基本思想是通过构造原变量的一系列线性组合形成几个综合指标, 以去除数据的相关性, 并使低维数据最大程度保持原始高维数据的方差信息[39–41]. 我们通过对综合日志特征进行降维, 实现了数据降维, 从而使数据集更容易使用, 降低算法的计算开销, 去除数据噪声, 减轻过拟合, 也就更容易获取有价值的信息. 我们通过对特征加权, 特征越重要, 所赋予的权值就越大, 而不太重要的特征赋予较小的权值, 模型中对每一个特征都赋予了一个权值, 从而有效保障了日志信息的完整性.

数据降维分为特征选择和特征提取两种方法, 文中采用的是特征提取方法, 即经已有特征的某种变换获取简约特征. 通过采用变换 (映射) 的方法, 把原始特征变换为较少的新特征, 由原始数据创建新的特征集, 从而有效提取出网络攻击等特征信息.

3 基于 RBM 的多源日志信息综合特征提取系统

利用上述 RBM 的训练学习方法, 构建用于提取综合特征的受限玻尔兹曼机神经网络 RBM_{com}, RBM_{com} 为两层网络, 可见层节点数为 $t\times N$, 隐藏层节点数为 C, 随机产生 [0,1] 之间的随机数作为神经网络的连接权值, 初始化 RBM_{com}. 基于 RBM 的多源日志信息综合特征提取系统构建如图 2 所示.

对第一层受限玻尔兹曼机神经网络 RBM_r 隐藏层输出的数据进行拼接组成数据序列, 将所述数据序列作为训练输入数据对受限玻尔兹曼机神经网络 RBM_{com} 进行训练, 所述数据序列的维度为 $t*N$, 根据对比散度快速学习方法优化网络权值, 由此得到稳定的 RBM_{com}. 基于训练好的受限玻尔兹曼机神经网络集合{ RBM_1, RBM_2, ⋯, RBM_t, RBM_{com} }, 构建多源日志综合特征提取系统. 基于 RBM 的多源日志信息综合特征提取算法详细过程如算法 1 所示.

在算法 1 中, 通过将从电力信息系统采集的各类日志数据输入相应的受限玻尔兹曼机神经网络, 经过不断训练学习, 构建多源日志综合特征提取系统, 再将各类日志数据输入 RBM 神经网络训练即可获得维度为 C 的综合特征数据.

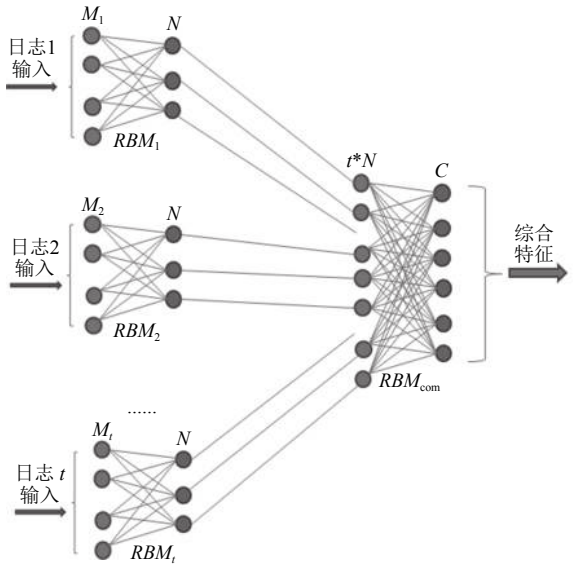

图 2 基于 RBM 的多源日志信息综合特征提取系统

算法1. 基于RBM的多源日志信息综合特征提取算法

输入: 各个设备的日志信息{日志1, 日志, ⋯, 日志t};

输出: 维度为C的综合特征数据.

1. 对于每类日志i, 构建受限玻尔兹曼机RBM_i.
2. 若输入维度为M_i, 构建可见层为M_i, 隐藏层为N的RBM_i.
3. 对于每类日志分别训练RBM_i.
4. 若共有t类日志, 构建可见层为$t*N$, 隐藏层为C的RBM_{com}.
5. 训练受限玻尔兹曼机神经网络RBM_{com}.
6. 构建多源日志综合特征提取系统.
7. 将各类日志数据输入相应的受限玻尔兹曼机神经网络, 即可获得维度为C的综合特征数据.
8. 结束.

4 实验结果与分析

为了验证基于 RBM 的多源日志综合特征提取方法的有效性, 本文以某电力公司的网络安全日志数据为例, 测试本文算法的有效性. 实验采用的日志数据主要包括公司内外网 IDS、IPS、防火墙、防病毒系统和数据库访问日志等.

我们基于前期研究搭建的电力系统中基于大数据的网络安全态势感知预警平台[42], 目前接入公司信息内外网共计 43 套信息系统重点资产作为监控对象. 平台已采集公司威胁事件数据 30TG, 记录攻击威胁条数 8160 兆条. 采集的日志信息样本集包括: 攻击触发时间、攻击威胁类型、危险等级、安全设备、客户单位、攻击源 IP、攻击目标 IP. 采集的日志信息样本类型如表 1 所示.

表 1 日志信息样例

攻击触发时间	攻击威胁类型	危险等级	安全设备	客户单位	源IP	目标IP
2019-10-14 09:34:49	特征值警报	高危	TSA	本部	172.29.21.24:56 104	42.96.207.142:80
2019-10-14 09:33:03	可疑域名	高危	TSA	本部	58.56.177.203:56 504	173.16.1.171:80
2019-10-14 09:31:42	IPS	高危	TSA	本部	219.146.247.136:54 181	172.168.9.46:80
2019-10-14 09:30:21	特征值警报	高危	TSA	本部	172.29.31.24:62 802	172.168.5.13:80
2019-10-14 09:29:44	Web攻击	高危	TSA	本部	172.29.21.44:56 104	172.3.4.242:80
2019-10-14 09:27:25	WEB攻击	高危	TSA	本部	162.159.208.73:56 504	172.169.9.72:80
2019-10-14 09:26:39	黑IP警报	高危	TSA	本部	68.64.174.82:56 504	172.168.8.28:80

通过利用基于 RBM 的多源日志综合特征提取方法对采集的安全日志信息进行训练学习, 提取的攻击类型包括: Web 攻击、IPS、特征值警报、可疑域名、黑 IP 警报、邮件敏感字、攻击事件-SCAN、攻击事件-DDOS 等. 通过对提取的攻击特征进行分析, 若有攻击则通过预警平台前端页面进行实时展示.

本实验中, 通过对公司内外网重点资产的攻击日志进行实时监测分析, 当监控到有 IP 地址在对公司系统进行持续攻击并多次触发高危报警, 便可以对该 IP 的攻击行为进行回溯取证, 根据提取的日志特征研判攻击类型, 从而预测预警公司的网络安全态势. 通过定位到该 IP 地址攻击行为的所有数据包, 然后下载该攻击数据包, 对攻击过程进行还原. 实验环境下监测到 2019 年 11 月的 Top IP 前 10 个主机总流量如图 3 所示, 攻击数据包的信息如图 4 所示.

图 3 Top IP 主机总流量

节点1->	端口1->	<-节点2	<-端口2	数据包	字节数	协议
114.254.131.230	41731	123.232.5.62	80	4	604.00 B	HTTP
123.232.5.126	53711	fund.eastmoney....	80	305	187.10 ...	HTTP
114.254.131.230	64242	123.232.5.62	80	3	1.16 KB	HTTP
114.254.131.230	17952	123.232.5.62	80	7	2.03 KB	HTTP
123.232.5.126	53808	fund.eastmoney....	80	176	101.42 ...	HTTP
114.254.131.230	19920	123.232.5.62	80	6	1.34 KB	HTTP
114.254.131.230	23334	123.232.5.62	80	1	468.00 B	HTTP
114.254.131.230	30697	123.232.5.62	80	5	1022.0...	HTTP
101.36.76.70	51464	58.56.115.224	443	14	10.83 KB	HTTP
123.232.5.4	65324	112.124.34.135	80	2	679.00 B	HTTP
123.232.5.15	65462	60.205.90.101	80	36	16.06 KB	HTTP
123.232.5.126	56942	fund.eastmoney....	80	279	177.37 ...	HTTP
123.232.5.6	31784	push.tianrow.com	80	8	2.23 KB	HTTP
202.110.201.209	60397	www.400800000...	80	230	155.63 ...	HTTP
58.56.115.207	4173	img.meijutt.com	808	96	82.09 KB	HTTP

图 4 攻击数据包信息

通过对攻击数据包的通信协议进行分析研判发现, IP 地址 114.254.131.230, TCP 端口号为 41 731 的主机对 IP 地址为: 123.232.5.62, TCP 端口号为 80 的攻击是一主机被植入了木马. 通过查看详细的 TCP 信息发现, 可进一步提取该主机的 passwd 敏感路径信息, 如图 5 所示. 通过对受感染的终端进行具体定位, 发现为某用户上网时不慎导致木马感染, 之后立即对该木马进行清除, 并在全公司范围内下发安全预警进行排查, 对发现的隐患进行全面消缺.

```
节点 1: IP 地址 = 114.254.131.230, TCP 端口 = 41731
节点 2: IP 地址 = 123.232.5.62, TCP 端口 = 80

GET /mrtg.cgi?cfg=/../../../../../../../../../etc/passwd HTTP/1.1
Connection: Keep-Alive
Host: 123.232.5.62
Pragma: no-cache
User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0)
Accept: image/gif, image/x-xbitmap, image/jpeg, image/pjpeg, image/png, */*
Accept-Language: en
Accept-Charset: iso-8859-1,*,utf-8
```

图 5 提取攻击特征找到被植入木马的主机

实验环境中, 通过威胁预警平台对安全设备日志特征进行提取分析, 以实验平台中 2019 年 10 月和 11 月的数据为例, 各类威胁次数排名 Top5 的对比分析如图 6 所示.

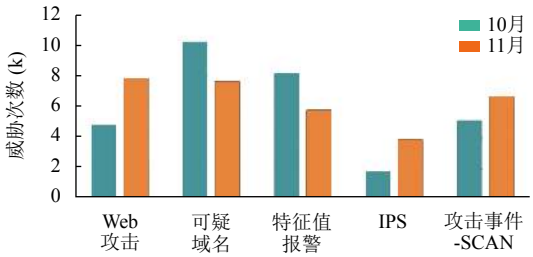

图 6 威胁类型次数排名 Top5 的对比分析

图 6 中, 竖轴中的 k 表示 1000 次. 图 7 展示了近 30 天内被攻击的 Top1 资产系统, 统计了月度内的威

胁事件总数和攻击源 IP 等信息. 从公司重点资产安全态势中能够清晰地展示各业务系统面临的威胁, 进而定位被攻击资产并进行隐患消缺.

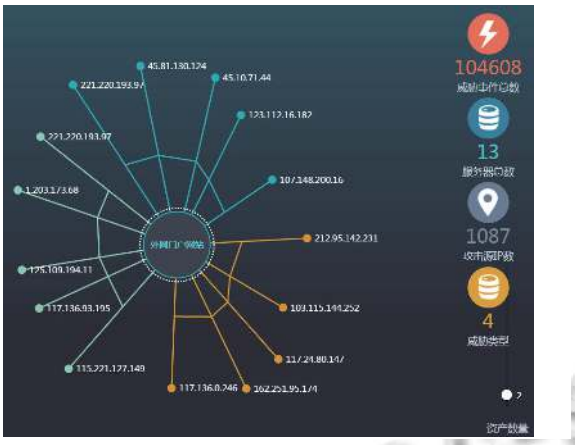

图 7 近 30 天攻击 Top1 及威胁事件总数

5 结论

随着“大云物移智链”信息新技术与电网业务的深度融合应用, 网络攻击威胁越来越多. 本文利用电网信息系统中安全日志信息中的异构数据源, 研究了基于受限玻尔兹曼机的电力大数据多源日志综合特征提取方法. 通过采用受限玻尔兹曼机神经网络对各类日志信息进行规范化编码, 有助于解决日志数据异构性带来的问题; 进而再用受限玻尔兹曼机神经网络对规范化编码后的日志信息进行处理, 可以实现降维并得到融合后的综合特征. 最后通过在电力信息系统中搭建大数据预警监测实验环境, 验证本文的日志综合特征提取方法对威胁检测的效果. 实验结果表明, 本文提出的方法在提取电力信息系统中日志特征时具有较高的准确率, 进而提高了网络安全态势预测的速度和预测精度.

参考文献

1 薛禹胜, 赖业宁. 大能源思维与大数据思维的融合 (一) 大数据与电力大数据. 电力系统自动化, 2016, 40(1): 1–8. [doi: 10.7500/AEPS20151208005]

2 赖特. 网络安全设备日志融合技术研究 [硕士学位论文]. 成都: 电子科技大学, 2015.

3 黄新平. 政府网站信息资源多维语义知识融合研究 [博士学位论文]. 长春: 吉林大学, 2017.

4 王秀锋. 网络环境下异构日志信息获取和预处理研究 [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2010.

5 酆勇, 熊庆宇, 石为人, 等. 一种基于受限玻尔兹曼机的说话人特征提取算法. 仪器仪表学报, 2016, 37(2): 256–262. [doi: 10.19650/j.cnki.cjsi.2016.02.003]

6 陈龙, 周剑, 王国胤. 融合多源日志辅助取证的事件场景关联方法. 重庆邮电大学学报 (自然科学版), 2007, 19(5): 584–589.

7 江雨燕, 桂伟. 基于受限玻尔兹曼机的分布式主题特征提取. 计算机工程与应用, 2017, 53(23): 108–112. [doi: 10.3778/j.issn.1002-8331.1706-0214]

8 程乐峰, 余涛, 张孝顺, 等. 机器学习在能源与电力系统领域的应用和展望. 电力系统自动化, 2019, 43(1): 15–31.

9 刘凯, 张立民, 周立军. 随机受限玻尔兹曼机组设计. 上海交通大学学报, 2017, 51(10): 1235–1240. [doi: 10.16183/j.cnki.jsjtu.2017.10.013]

10 杨杰, 孙亚东, 张良俊, 等. 基于弱监督学习的去噪受限玻尔兹曼机特征提取算法. 电子学报, 2014, 42(12): 2365–2370. [doi: 10.3969/j.issn.0372-2112.2014.12.005]

11 沈卉卉, 李宏伟. 基于动量方法的受限玻尔兹曼机的一种有效算法. 电子学报, 2019, 47(1): 176–182. [doi: 10.3969/j.issn.0372-2112.2019.01.023]

12 康丽萍, 许光銮, 孙显. 受限玻尔兹曼机的稀疏化特征学习. 计算机科学, 2016, 43(12): 91–96. [doi: 10.11896/j.issn.1002-137X.2016.12.016]

13 罗恒. 基于协同过滤视角的受限玻尔兹曼机研究 [博士学位论文]. 上海: 上海交通大学, 2011.

14 何洁月, 马贝. 利用社交关系的实值条件受限玻尔兹曼机协同过滤推荐算法. 计算机学报, 2016, 39(1): 183–195. [doi: 10.11897/SP.J.1016.2016.00183]

15 高琰, 陈白帆, 晁绪耀, 等. 基于对比散度-受限玻尔兹曼机深度学习的产品评论情感分析. 计算机应用, 2016, 36(4): 1045–1049. [doi: 10.11772/j.issn.1001-9081.2016.04.1045]

16 郑志蕴, 李步源, 李伦, 等. 基于云计算的受限玻尔兹曼机推荐算法研究. 计算机科学, 2013, 40(12): 259–263. [doi: 10.3969/j.issn.1002-137X.2013.12.056]

17 张俊玲, 陈志刚, 许旭, 等. 基于改进卷积受限玻尔兹曼机的滚动轴承故障诊断. 组合机床与自动化加工技术, 2019, (5): 73–76. [doi: 10.13462/j.cnki.mmtamt.2019.05.018]

18 张立民, 刘凯. 基于深度玻尔兹曼机的文本特征提取研究. 微电子学与计算机, 2015, 32(2): 142–147. [doi: 10.19304/j.cnki.issn1000-7180.2015.02.033]

19 谈建慧, 景新幸, 杨海燕. 深度信念网络的 Bottleneck 特征提取方法. 桂林电子科技大学学报, 2016, 36(2): 118–122. [doi: 10.16725/j.cnki.cn45-1351/tn.2016.02.007]

20 闫琰. 基于深度学习的文本表示与分类方法研究 [博士学位论文]. 北京: 北京科技大学, 2016.

21 米龙. 自适应深度学习算法在目标分类问题中的应用 [硕

士学位论文]. 沈阳: 东北大学, 2014.
22 孙劲光, 蒋金叶, 孟祥福, 等. 深度置信网络在垃圾邮件过滤中的应用. 计算机应用, 2014, 34(4): 1122–1125.
23 毛冬. 大数据下风电机组发电机故障预警方法的研究 [硕士学位论文]. 北京: 华北电力大学, 2016.
24 张春霞, 姬楠楠, 王冠伟. 受限波尔兹曼机. 工程数学学报, 2015, 32(2): 159–173. [doi: 10.3969/j.issn.1005-3085.2015.02.001]
25 吴志宇. 基于深度学习的稀疏化电能质量扰动识别方法研究 [硕士学位论文]. 成都: 西南交通大学, 2019.
26 周博曦, 秦晋, 王金亮, 等. 基于人工神经网络与有限状态机的变电站告警处理系统. 山东电力技术, 2020, 47(1): 6–13.
27 张廷忠, 张庆辉, 邢强, 等. 基于 LMD 与 GA-BP 神经网络组合的短期风速滚动预测方法. 山东电力技术, 2019, 46(11): 13–20. [doi: 10.3969/j.issn.1007-9904.2019.11.003]
28 田怀源, 张峰, 王新库, 等. 基于灰色关联度和 BP 神经网络的最大负荷同时率预测方法研究. 山东电力技术, 2017, 44(4): 11–15, 21. [doi: 10.3969/j.issn.1007-9904.2017.04.003]
29 Chen R, Ren ZW, Meng ZH, *et al*. Oblique-incidence reflectivity difference method combined with deep learning for predicting anisotropy of invisible-bedding shale. Energy Reports, 2020, 6: 795–801.
30 Velasco JA, Amaris H, Alonso M. Deep Learning loss model for large-scale low voltage smart grids. International Journal of Electrical Power & Energy Systems, 2020, 121: 106054. [doi: 10.1016/j.ijepes.2020.106054]
31 Zheng WY, Huang WJ, Hill DJ. A deep learning-based general robust method for network reconfiguration in three-phase unbalanced active distribution networks. International Journal of Electrical Power & Energy Systems, 2020, 120: 105982. [doi: 10.1016/j.ijepes.2020.105982]
32 Krishnan R, Jagannathan S, Samaranayake VA. Direct error-driven learning for deep neural networks with applications to big data. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(5): 1763–1770. [doi: 10.1109/TNNLS.2019.2920964]
33 Li HL, Weng J, Shi YJ, *et al*. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. Scientific Reports, 2018, 8(1): 6600. [doi: 10.1038/s41598-018-25005-7]
34 Chen CLP, Feng S. Generative and discriminative fuzzy restricted Boltzmann machine learning for text and image classification. IEEE Transactions on Cybernetics, 2020, 50(5): 2237–2248. [doi: 10.1109/TCYB.2018.2869902]
35 Chu JL, Wang HJ, Meng H, *et al*. Restricted boltzmann machines with Gaussian visible units guided by pairwise constraints. IEEE Transactions on Cybernetics, 2019, 49(12): 4321–4334. [doi: 10.1109/TCYB.2018.2863601]
36 Park CH, Lee GH. Comparison of incremental linear dimension reduction methods for streaming data. Pattern Recognition Letters, 2020, 135: 15–21. [doi: 10.1016/j.patrec.2020.03.028]
37 Washington P, Paskov KM, Kalantarian H, *et al*. Feature selection and dimension reduction of social autism data. Pacific Symposium on Biocomputing, 2020, 25: 707–718.
38 Choi JY, Kyung M, Hwang H, *et al*. Bayesian extended redundancy analysis: A Bayesian approach to component-based regression with dimension reduction. Multivariate Behavioral Research, 2020, 55(1): 30–48. [doi: 10.1080/00273171.2019.1598837]
39 Xu C, Yang M, Zhang J. Fast deflation sparse principal component analysis via subspace projections. Journal of Statistical Computation and Simulation, 2020, 90(8): 1399–1412. [doi: 10.1080/00949655.2020.1728761]
40 Sanchez J, Denis F, Coeurjolly D, *et al*. Robust normal vector estimation in 3D point clouds through iterative principal component analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 163: 18–35. [doi: 10.1016/j.isprsjprs.2020.02.018]
41 Zhang X, He L, Zhang J, *et al*. Determination of key canopy parameters for mass mechanical apple harvesting using supervised machine learning and Principal Component Analysis (PCA). Biosystems Engineering, 2020, 193: 247–263. [doi: 10.1016/j.biosystemseng.2020.03.006]
42 刘冬兰, 刘新, 张昊, 等. 基于大数据的网络安全态势感知及主动防御技术研究与应用. 计算机测量与控制, 2019, 27(10): 229–233. [doi: 10.16526/j.cnki.11-4762/tp.2019.10.047]