

使用自组织增量径向基网络的 在线网络流量分类算法

陈卓, 吕娜, 张彦晖, 缪竞成
(空军工程大学信息与导航学院, 710077, 西安)

摘要: 针对传统流量分类方法存在模型训练速度慢、样本标记成本高和难以满足分类实时性的问题,构建了一种在线的半监督流量分类算法——OSOINN-RBF。首先使用改进的自组织增量神经网络对数据进行在线增量式的无监督学习,获得代表流量数据分布的 SOINN 网络,SOINN 中节点的权重和半径分别作为径向基网络隐藏层节点的中心和半径;径向基网络能够捕获数据中难以发现的规律性,具有良好的泛化能力和学习收敛速度;最后使用少部分标记样本调节径向基网络输出层的权重,提高径向基网络的分类性能。实验结果证明,与主流分类算法相比,OSOINN-RBF 算法的分类性能都为最优,具有最小的时间开销;面对未知流量类别时,OSOINN-RBF 算法相比 SOINN 算法的分类准确率提高了 5%~7%。

关键词: 流量分类;半监督学习;径向基函数;神经网络

中图分类号: TP393 **文献标志码:** A

DOI: 10.7652/xjtuxb202012008 **文章编号:** 0253-987X(2020)12-0062-08



OSID 码

A Classification Algorithm of Online Network Traffic Based on Self-Organizing Incremental Radial Basis Network

CHEN Zhuo, LÜ Na, ZHANG Yanhui, MIAO Jingcheng

(College of Information and Navigation, Air Force Engineering University, Xi'an 710077, China)

Abstract: An online semi-supervised traffic classification algorithm OSOINN-RBF is proposed to solve the problems of slow model training speed, high sample labeling cost and difficulty of real-time classification in traditional traffic classification methods. First, an improved self-organizing incremental neural network is used to perform online incremental unsupervised learning of data and to obtain a SOINN network that represents the distribution of input traffic data, and the weights and radius of the nodes in SOINN are used as the center and the radius of the hidden layer of the radial basis network; Then, the radial basis network captures the regularity that is difficult to find in the data, and has good generalization ability and learning convergence speed; Finally, a small number of labeled samples is used to adjust the weights of the output layer of the radial basis network and to improve the classification performance of the network. Experimental results and comparisons with mainstream classification algorithms show that the classification performance of the proposed algorithm is the best, with minimal time overhead; The classification accuracy of OSOINN-RBF algorithm is also improved by 5% to 7% compared with SOINN algorithm when facing unknown traffic categories.

收稿日期: 2020-05-26。 作者简介: 陈卓(1997—),男,硕士生;吕娜(通信作者),女,教授,博士生导师。 基金项目: 国家自然科学基金资助项目(61703427)。

网络出版时间: 2020-07-29

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1069.T.20200729.1425.002.html>

Keywords: traffic classification; semi-supervised learning; radial basis function; neural network

目前网络规模日益扩大,网络行为日趋复杂,为了更好地进行网络资源调度、降低网络安全隐患、提供更好的服务质量,需要进行网络流量分类^[1]。

网络流量分类主要有端口方法、深度包检测方法、流统计分析方法等。基于端口的分类方法实现简单,但越来越多的应用采用动态端口技术,导致这种方法的分类准确率大大降低。深度包检测方法需要分析数据包中的全部有效载荷来判断应用类型,这种方法的存储开销和计算量较大,不适合在线分类。

近年来,基于流量统计特征的机器学习分类方法受到了越来越多的关注^[2]。机器学习算法通常按训练数据是否标记分为有监督学习^[3]、无监督学习^[4]和半监督学习^[5]3种。有监督学习依赖于大量有标记的训练样本数据,虽然分类准确率高,但难以识别未知流量,而且有监督学习方法还受到标签稀缺问题的严重限制。无监督学习则无需标记训练样本,常用的方法有 K-means、PCA 等,这些方法可以挖掘未标记样本的隐含特征来进行分类,但往往因为缺少标记样本的指导,分类准确率较低^[6]。半监督学习作为这两种方法的折中,使用少量标记样本和大量无标记样本,既能取得较好的分类性能又能减少标记样本的难度^[7]。

传统的机器学习分类方法为批量离线式,要求所有训练数据必须在本地可用,存在着模型训练慢、计算复杂度高的缺陷^[8],并不适合实时处理在线网络数据。针对网络流量分类的实时性问题,文献[9-10]比较了离线训练和在线训练方法,证明在线训练方法具有较低的时间和空间复杂度,适合网络流量在线分类。文献[11]提出了自组织增量神经网络(SOINN),SOINN 使用变化的网络拓扑结构来反映数据流分布,利用两层神经网络训练得到的分类器实现在线增量分类。文献[12]将 SOINN 与其他神经网络算法进行分析比较,证明了 SOINN 比其他在线半监督学习方法具有更好的分类性能。

针对高速大规模的网络流量分类问题,为了提高分类的准确性和实时性,降低分类模型的计算复杂度,本文提出一种在线的半监督流量分类算法(OSOINN-RBF)。在多个数据集上验证了算法的有效性,相比其他在线学习算法,OSOINN-RBF 具有较高的分类准确性和较低的时间开销,当面对新的未知流量类别时,OSOINN-RBF 算法不仅具有较

高的识别率,整体分类性能也更为稳定。

1 在线流量分类框架及算法

在流量分类的研究中广泛采用 5 元组定义流^[13],即{源 IP,目的 IP,源端口,目的端口,传输层协议}。文献[8]对流中所有数据包进行统计特征分析,该方法只能用于离线的网络流量分类,不能满足实时在线的流量分类。本文提出了一种基于在线学习的网络流量分类框架,如图 1 所示。

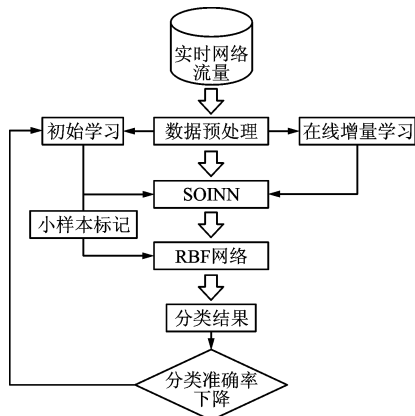


图 1 流量分类框架

图中的数据预处理模块以流的前 5 个数据包形成子流^[13],并计算其到达时间间隔、数据包大小等参数的均值、方差等统计特征,作为流量特征。使用 CICFlowMeter 工具实现子流生成及特征提取功能,这种流量特征提取方法速度快,内存占用小,适合在线分类^[14]。

本文的网络流量分类框架分为 2 个主要阶段:初始学习和在线增量学习。

(1)初始学习:使用开源工具 nDPI 对抽样获得的少部分子流 $S = \{s_i, 0 < i \leq N\}$ 进行标记,得到初始应用类型集为 $Y = \{y_j, 0 < j \leq M\}$,生成 $\{s_i, y_j\}$ 映射关系表,获得初始标记样本数据集,利用 OSOINN-RBF 算法训练模型的隐藏层和输出层,径向基函数(RBF)网络输出层使用标记的监督样本调整权重,得到流量分类器。

(2)在线增量学习:将新到的不含标记的流数据输入模型,使用改进的 SOINN 算法对隐藏层进行在线增量学习,输出层保持不变,得到分类结果。在线增量学习中,分类器会偏向于新输入的类别,因此当模型准确率具有明显的下降趋势时,回到初始学习阶段重新训练输出层,并判断是否产生新类别。

1.1 SOINN 算法及其改进

自组织增量神经网络(SOINN)是一种基于竞争学习的神经网络模型^[11],具有良好的数据压缩特性,用于在线学习时可以有效降低存储和计算开销^[15],因此本文使用 SOINN 算法对数据进行在线增量式的无监督学习。

但是,原始的 SOINN 存在 2 个问题:①采用双层网络进行在线学习,难以确定何时停止学习第一层,以及何时开始学习第 2 层;②无法很好地分离高密度重叠区域。本文使用改进的 SOINN 算法解决这些问题,采用单层网络减少计算难度^[16],通过网络去噪技术获得更好的聚类结果。改进 SOINN 算法采用单层结构来实现在线增量学习,根据 Shen 等人的分析,它可以实现与两层 SOINN 相同的分类性能,并且具有较快的学习速度^[12]。改进 SOINN 算法的训练过程主要分为 3 步:神经元分布、拓扑学习和网络去噪。

(1)神经元分布。假设输入的数据为 x_t ,初始神经元集合为 N ,在集合 N 中寻找与 x_t 最近的神元分别作为获胜者 w_1 和第二获胜者 w_2 ,具体计算公式如下

$$\left. \begin{aligned} w_1 &= \arg \min_{i \in N(t)} \|W_i - x_t\| \\ w_2 &= \arg \min_{i \in N \setminus \{t, w_1\}} \|W_i - x_t\| \end{aligned} \right\} \quad (1)$$

然后计算 w_1 和 w_2 的相似度阈值 σ_1 和 σ_2 ,相似度阈值即神经元的半径,定义为 $\sigma_i = \max_{j \in N_i} \|W_i - W_j\|$,其中 N_i 为节点 i 的邻居节点集合,若 N_i 为空则 $\sigma_i = \min_{j \in N \setminus \{i\}} \|W_i - W_j\|$ 。

通过输入样本 x_t 与获胜者 w_1 和 w_2 的距离来判断是否生成新神经元,若 $\|x_t - w_1\| > \sigma_1$ 或者 $\|x_t - w_2\| > \sigma_2$,则将 x_t 作为新生神经元节点插入网络,如图 2a 所示。

图 2b 所示的第 2 种情况是输入样本位于获胜神经元的半径内时,将建立这 2 个单元之间的连接,

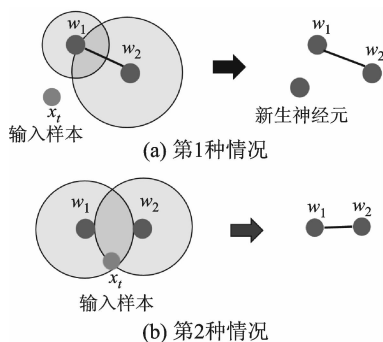


图2 神经元学习的 2 种情况

不再生成新的神经元。

(2)拓扑学习。假设神经元连接关系集合 $C \subset N \times N$,当 x_t 成为新生神经元,则更新神经元集合为 $N = N \cup \{i\}$,若 w_1 和 w_2 不存在连接,获胜者和第二获胜者将建立连接,并将连接添加到 C 中, $C = C \cup \{(w_1, w_2)\}$ 。新添加的边年龄设置为 0,与获胜者相连的所有边的年龄将增加 1。 $A_{\text{age}, w_1, a} = A_{\text{age}, w_1, a} + 1$,其中 a 为 w_1 周围的邻居神经元。

如果输入样本被判断为获胜者 w_1 或第二获胜者 w_2 的同一集群,则获胜者及其相邻节点的权重向量将更新

$$\left. \begin{aligned} w_1 &= w_1 + \epsilon(t)(x_t - w_1) \\ w_2 &= w_2 + \epsilon(t)\lambda(x_t - w_2) \end{aligned} \right\} \quad (2)$$

式中: $\epsilon(t)$ 是神经元获胜时间的倒数; λ 通常为 0.01。

(3)网络去噪。在实际网络中,由于输入数据存在噪声,导致 SOINN 学习生成不必要的节点,从而导致难以分离高密度重叠区域。因此在算法运行过程中,周期性地查找并删除处于密度较低区域的部分节点,可以降低噪声数据的影响。此外,在 2 个距离比较近的聚类交界处,数据点的密度也会明显低于聚类的中心区域,因此删除这部分的节点以及与其相关的连接,有助于算法够稳定地把 2 个相距较近的聚类区别开来。本文定义节点 i 的密度 $D(i)$ 为

$$D(i) = \frac{1}{(1 + d_i)^2} \quad (3)$$

式中: $d_i = \frac{\sum_{j \in N_i} \|W_i - W_j\|}{|N_i|}$ 表示节点 i 及其邻居之间距离的平均值, $|N_i|$ 代表邻居节点数量。

2 在线半监督学习的流量分数算法

本文结合径向基函数(RBF)网络和无监督学习算法(SOINN),提出一种基于半监督学习的流量分类算法(OSOINN-RBF),使用 SOINN 算法获取 RBF 网络隐藏层节点的中心和半径,使用少部分监督样本调节 RBF 网络输出层的权重。本文 OSOINN-RBF 算法原理如图 3 所示。本文算法通过隐藏层将原始输入映射到更高维的空间,能够捕获数据中难以发现的规律性,具有良好的泛化能力和学习收敛速度。

RBF 网络可以表示为

$$f(x) = \sum_{k=1}^K w_k \varphi_k(x) \quad (4)$$

式中: w_k 为隐藏单元的权重;核函数 $\varphi_k(x)$ 用来衡

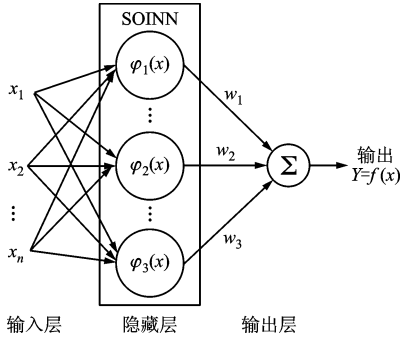


图 3 OSOINN-RBF 算法

量样本和样本之间的相似度。本文选择高斯核作为径向函数

$$\varphi(x, \mu_k, \sigma_k) = \exp\left(-\frac{1}{2\sigma_k^2} \|x - \mu_k\|^2\right) \quad (5)$$

s. t. $k = 1, 2, \dots, K$

其中 μ_k 和 σ_k 是高斯核的超参数, 分别代表隐藏单元 k 的中心与半径。

RBF 网络的主要参数在于隐藏单元的中心, 半径和输出层权重。本文使用 SOINN 算法接受新数据来更新隐藏层的网络单元分布及其连接, 最终获得代表流量数据分布的 SOINN 网络, SOINN 中节点的权重和半径分别代表隐藏单元 k 的中心与半径。网络的隐藏层到输出层的映射可以写成

$$RW = Y$$

$$\text{s. t. } \begin{cases} \mathbf{R} = [\Phi(x_1), \Phi(x_2), \dots, \Phi(x_n)]^T \\ \Phi(x_i) = [\varphi(x_i, \mu_1, \sigma_1), \dots, \varphi(x_i, \mu_K, \sigma_K)] \\ \mathbf{W} = [w_1, w_2, \dots, w_k]^T \\ \mathbf{Y} = [y_1, y_2, \dots, y_n]^T \end{cases} \quad (6)$$

式中: $\mathbf{R}$ 是隐藏层的输出; $\mathbf{W}$ 是输出层的权重参数; $\mathbf{Y}$ 是数据样本真实标签, 使用数据的真实标签来调整 $\mathbf{W}$ 。利用最小二乘法可以得到权重的计算公式

$$\mathbf{W} = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{Y} \quad (7)$$

2.1 模型优化算法

当使用最小二乘法进行矩阵求逆时, 计算复杂度是 $O(d^3)$, 其中 d 为特征数目, 不适合在线计算。使用较多的在线优化算法是随机梯度下降 (SGD), 利用每个样本来迭代更新模型, 计算复杂度是 $O(d)$, 其参数更新为 $w_{t+1} = w_t - \eta_t g_t$, 其中 η_t 为 t 时刻的学习率, g_t 为 t 时刻模型预测损失的梯度。这种方法最大的优点就是训练速度快, 但是缺点在于这种方法训练得到的模型不够稀疏, 占用的存储空间较大。一种能够让模型稀疏的方法是在损失函

数中加入 L_1 惩罚项, 对模型作用较小的特征的权重置为零, 保留与目标变量最相关的特征变量。

本文使用 Google 提出的 FTRL (Follow The Regularized Leader)^[17] 模型进一步优化 SGD 算法来调整权重, 其参数更新公式为

$$w_{t+1} = \arg \min_w \left(g_{1:t} w + \frac{1}{2} \sum_{s=1}^t \eta_s \|w - w_s\|_2^2 + \lambda_1 \|w\|_1 \right) \quad (8)$$

式中: $g_{1:t}$ 为从初始时刻到 t 的所有梯度的和; η_s 为 s 时刻的学习率。该优化函数保证了新产生的权重与历史权重不会相差太大, 第 2 项为利用 L_1 正则进行稀疏性约束, 第 3 项为利用 L_2 正则来防止过拟合。FTRL 的核心思想是利用之前的模型更新信息对当前步的模型进行更新, 在保证模型准确度的情况下使模型尽可能稀疏, 降低模型复杂度。

本文所提 OSOINN-RBF 算法描述如下。

步骤 1 初始化神经元集合 $\mathbf{N} = \{n_1, n_2\}$, 初始化边集合 $\mathbf{C}$ 为空集;

步骤 2 输入数据样本 x_t ;

步骤 3 寻找 $\mathbf{N}$ 中与 $\mathbf{W}$ 最近的 2 个神经元作为获胜者 w_1 和第二获胜者 w_2 ;

步骤 4 计算自适应阈值 σ_1 和 σ_2 , 若满足 $\|x_t - w_1\| > \sigma_1$ 或者 $\|x_t - w_2\| > \sigma_2$, 则添加权重为 x_t 的新生神经元 n_s , $\mathbf{N} = \mathbf{N} \cup n_s$;

步骤 5 如果 n_1, n_2 间不存在连接, 则建立连接, $\mathbf{C} = \mathbf{C} \cup \{n_1, n_2\}$, $\text{age}(n_1, n_2) = 0$;

步骤 6 如果没有建立新的节点, 则使用式 (2) 更新 2 个胜者节点的值;

步骤 7 与获胜者节点 w_1 相连的所有边的参数 $\text{age}(w_1, i)$ 加 1, 即 $\text{age}(w_1, i) = \text{age}(w_1, i) + 1$, $\forall i \in N_i$;

步骤 8 检查 $\mathbf{C}$ 中所有连接的年龄参数, 若年龄参数 $\text{age}(i, j) > T_{\text{th, max}}$, 就从 $\mathbf{C}$ 移除该连接;

步骤 9 如果当前输入的数据样本总数是 λ 的整数倍, 检查整个 SOINN 网络执行去噪过程;

步骤 10 步骤利用 SOINN 得到的隐藏单元 k 的中心与半径以及样本标签 $\mathbf{Y}$, 通过式 (8) 得到隐藏单元的权重 $\mathbf{W}$ 。

2.2 算法复杂度分析

对于在线算法而言, 计算复杂度和存储开销是最为重要的 2 个指标^[18]。

(1) 计算复杂度: 本文提出的 OSOINN-RBF 算法的计算复杂度主要包括两部分, 分别是隐藏层和

输出层。将学习到的神经元的数量表示为 s , 输出层的计算复杂度为 $O(s)$ 。隐藏层的神经元分布阶段, 计算复杂度为 $O(nds)$, 其中 n 是输入样本数, d 是特征数量。网络去噪过程需要将所有节点聚类并删除有噪声的节点, 这两个操作需要扫描所有神经元节点, 因此计算复杂度为 $O(s)$ 。OSOINN-RBF 算法的总体计算复杂度为 $O(nds+s)=O(nds)$, 每次输入一个样本的计算复杂度为 $O(ds)$ 。

(2) 存储开销: OSOINN-RBF 在线算法需要存储模型各层参数: 输入层 $x_i \in \mathbf{R}^{d \times n}$, 隐藏层 $\Phi(x_i) \in \mathbf{R}^{d \times k}$, 输出层 $\mathbf{W} \in \mathbf{R}^{k \times 1}$, 其存储开销为 $dn+dk+k$, 在线分类算法需要存储 p 个初始的神经元中心及其权重, 其存储开销为 $2p$ 。因此, 总的存储开销为 $dn+dk+k+2p$ 。

在上述的计算复杂度和存储开销中, 由于 RBF 网络中高斯核函数的局部逼近特性, 只有少数样本特征得到激活, 类似于 L_1 惩罚项的作用, 产生稀疏权重矩阵。因此实际计算的样本特征数量远小于 d , 实现了模型的稀疏性, 进一步降低了模型的计算复杂度。

3 实验测试与分析

为了验证算法的可行性, 本文在一台内存为 16 GB, 搭载主频为 1.8 GHz 处理器的 64 位 PC 上使用 Anaconda 3 软件进行实验。

3.1 实验数据

为了准确、全面评估分类模型, 实验使用了 2 个公开的网络流量数据集: Moore 数据集^[19]; ISCX-VPN 数据集^[20]。

(1) Moore 数据集中不同应用程序产生的流量类别的样本数量的统计信息如表 1 所示。Moore 数

表 1 Moore 数据集中不同应用类别的样本数量	
应用类别	样本数
WWW	328 092
MAIL	28 567
BULK	11 539
DATABASE	2 648
SERVICE	2 099
P2P	2 094
ATTACK	1 793
MEDIA	1 152
INT	110
GAME	8
总计	377 526

据集的流量中约有 86% 属于网络浏览器(WWW), 还有 7.5% 属于电子邮件(MAIL), 仅这两个类别占所有流量类别的 95%, 可以看出网络流量类别分布极其不均匀。

(2) ISCX-VPN 数据集由 15 种不同应用程序的实际流量组成。本文将应用程序标记为以下大类: 多媒体应用、VoIP、文件传输和即时消息应用。表 2 为实验中选择的网络流构成的小型数据集, 由表 2 可见, 数据集不同流量类别分别较为均匀。

表 2 ISC-XVPN 数据集中不同应用类别的样本数量	
应用类别	样本数
多媒体应用	452
VoIP	645
文件传输	740
即时消息应用	420

3.2 实验与结果分析

实验仿真中, 分别选取半监督和完全监督算法中应用广泛的分类算法: 增量支持向量机 (IS-VM)^[21]、增量 K-means^[22]、在线随机森林 (ORF)^[23] 和 CNN-LSTM^[24] 进行对比, 其中, CNN-LSTM 采用离线训练的方法, 其他均为在线训练, 算法的参数按照原文文中的设置。

3.2.1 模型整体性能和时间开销对比 Moore 数据集用来模拟大样本场景, ISCX-VPN 数据集模拟小样本场景, 测试模型多分类的能力。随机选择 80% 样本作为训练集, 20% 样本为测试集。表 3 给出了算法的整体分类性能对比, 可以看出 OSOINN-RBF 的平均准确率最高, 为 99.6%, 对比 SOINN 算法验证了本文算法的设计有利于提高分类性能。CNN-LSTM 采用完全监督的方法, 平均准确率为 99.2%, 性能仍弱于本文的 OSOINN-RBF 算法。

表 3 不同算法在 2 个数据集上的整体性能对比		
算法	平均准确率	
	Moore	ISCX-VPN
ISVM	0.934	0.923
K-means	0.956	0.894
ORF	0.962	0.914
CNN-LSTM	0.992	0.973
SOINN	0.991	0.956
OSOINN-RBF	0.996	0.997

两个数据集的的分类结果对比,可以看出 ISVM、K-means、ORF 和 CNN-LSTM 在小样本场景中的表现较差,明显不如大样本场景中的分类性能。而 OSOINN-RBF 算法无论大样本场景还是小样本场景,分类性能都为最优且非常稳定。

表 4 为各算法在 Moore 数据集的训练集上的模型复杂度对比,选择的指标是模型参数数量(Params)以及浮点运算次数(FLOPs)。可以看出 OSOINN-RBF 具有最小的模型运算复杂度,因此计算量最小,学习速度最快。

表 4 不同算法的复杂度对比

算法	复杂度/ 10^6	
	Params	FLOPs
ISVM	534.2	30365.4
K-means	247.5	8465.2
ORF	0.751	7.625
CNN-LSTM	7.529	16.004
SOINN	0.358	4.121
OSOINN-RBF	0.137	1.612

对比各种算法在不同数量的测试集下的时间开销,选择 Moore 数据集的测试集,定义 N 为实验中在线测试时所选择的测试集占总测试集的比例,表 5 为各算法的时间开销对比。由表 5 可见:由于 K-means 和 ISVM 的计算复杂度较高,时间开销随着数据量的增长而急剧增加;OSOINN-RBF 算法相比 OSOINN 算法运行速度更快,这是由于算法的设计使其具有更小的模型运算复杂度,因此速度更快;OSOINN-RBF 的线测试时间开销最小,不超过 0.14 s,满足实时分类要求。

表 5 不同算法的在线测试时间开销对比

算法	测试时间/s		
	$N=10\%$	$N=50\%$	$N=90\%$
ISVM	0.173	0.625	3.121
K-means	0.024	1.414	4.656
ORF	0.057	0.605	1.212
SOINN	0.011	0.091	0.221
OSOINN-RBF	0.010	0.061	0.134

3.2.2 在线学习能力对比 为测试算法的在线学习能力与分类性能稳定性,图 4、图 5 给出了不同算法的在线学习能力对比,其中图 4 为算法在训练样

本标记比例为 10% 下的分类准确率,图 5 为算法在训练样本标记比例为 20% 下的分类准确率。可以看出,在样本标记比例 10% 的情况下,训练初期样本数较少时,K-means 分类准确率相对较高,但随着数据量的增加,准确率时有下降,分类性能并不稳定,而 OSOINN-RBF 是在线学习算法中,分类准确率提升最快和最为稳定的。

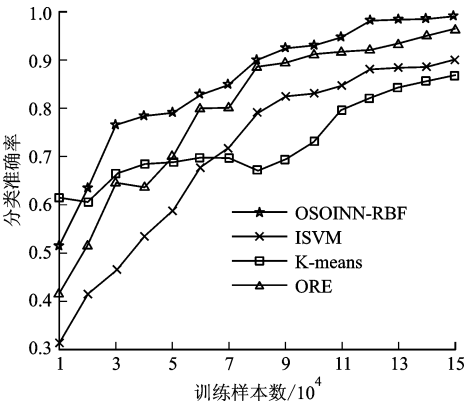


图 4 训练样本标记比例 10% 下的分类性能

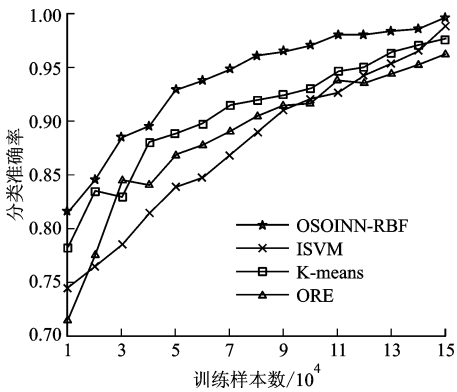


图 5 训练样本标记比例 20% 下的分类性能

对比 OSOINN-RBF 算法在不同标记比例训练样本下的分类准确率,可以看出,训练初期样本数量较少时,标记比例越高,算法的分类准确率越高;由于 OSOINN-RBF 算法具有在线增量学习能力,当样本数量达到 10 000 左右时,分类性能基本持平,不再受样本标记比例影响。实验结果表明,OSOINN-RBF 算法在样本标记数量较少的情况下,通过在线学习依然可以取得很好的分类性能,这也降低了样本标记的成本,适合半监督学习算法的实现。

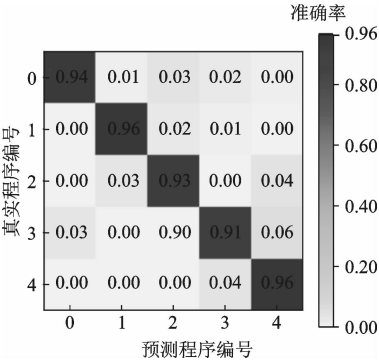
3.2.3 检测未知分类的能力 为测试网络中出现新的流量类别时算法的分类性能,在 Moore 数据集选择部分样本数据组成 Moore_sets 数据集,设置初始数据集含有 5 种类型的应用类别,在线检测数据集中新增两种未知应用类别,数据集具体描述见表 6。

表 6 Moore_sets 数据集的不同应用类别在初始训练集和在线测试集上的样本数量

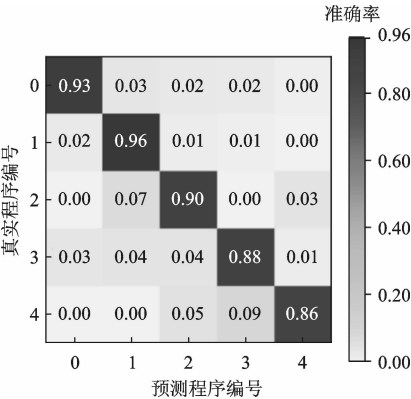
应用类别	样本数	
	初始数据集	在线检测数据集
WWW(0)	20 000	10 000
MAIL(1)	10 000	10 000
BULK(2)	5 000	5 000
DATABASE(3)	1 500	500
SERVICE(4)	1 000	1 000
P2P(5)	0	1 000
ATTACK(6)	0	500

注:(0)、(1)~(6)中数字为对该应用类别的编号。

对比在线学习的 OSOINN-RBF 算法和 SOINN 算法,模型分类结果的混淆矩阵见图 6、图 7(颜色越深表示准确率越高)。



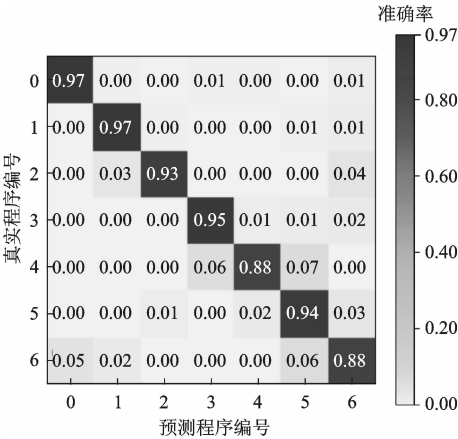
(a)OSOINN-RBF



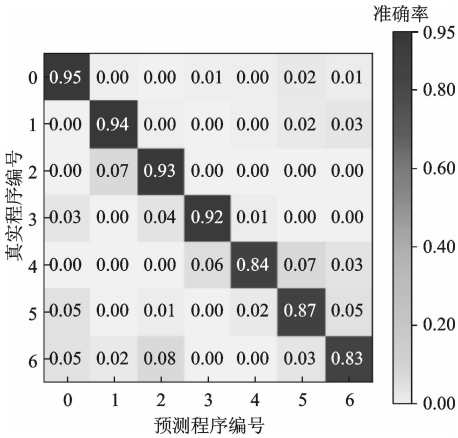
(b)SOINN

图 6 初始数据集分类结果

从图 6 和图 7 可以看出,无论初始数据集还是在线检测的数据集,OSOINN-RBF 算法的分类准确率均优于 SOINN 算法;在线检测中,OSOINN-RBF 算法不仅提高了 WWW、MAIL、BULK、DATA-



(a)OSOINN-RBF



(b)SOINN

图 7 在线检测数据集分类结果

BASE 这几种应用类别的分类准确率,对于未知类别 P2P 和网络攻击 ATTACK 也具有 94%和 88%的准确率,高于 SOINN 的 87%和 83%,而 SOINN 算法在面对未知流量时,所有应用类别的准确率均有所下降。实验结果证明了当出现新的流量类别时,OSOINN-RBF 算法不仅具有较高的识别率,整体分类性能也更为稳定。

4 结 论

本文设计了一种半监督流量分类算法——OSOINN-RBF,为网络流量的在线分类提供了一种新的解决方法,在多个数据集上进行了实验,对比了几种在线学习算法和离线算法,结果表明:

(1)本文采用的半监督学习算法在少量标记样本的情况下,具有较快的在线学习速度和较高的分类准确率,考虑真实网络环境中标记样本较少,本文算法具有较高应用价值。

(2)本文的算法具有较快的学习速度,在线检测中时间开销较小,满足流量分类的实时性要求。

(3)当网络流量环境变化,出现新的未知流量时,本文算法相比 SOINN 在线学习算法具有更好的动态更新能力,对未知类别具有较好的识别率,分类性能也更为稳定。

参考文献:

- [1] 叶春明,王珍,陈思,等. 基于节点行为特征分析的网络流量分类方法 [J]. 电子与信息学报, 2014, 36(9): 2158-2165.
YE Chunming, WANG Zhen, CHEN Si, et al. Internet traffic classification based on hosts behavior analysis [J]. Journal of Electronics and Information Technology, 2014, 36(9): 2158-2165.
- [2] LI Jun, ZHANG Shunyi, LU Yanqing, et al. Internet traffic classification using machine learning [C]// 2007 2nd International Conference on Communications and Networking in China. Piscataway, NJ, USA: IEEE, 2007: 239-243.
- [3] 顾纯祥,吴伟森,石雅男,等. 基于自编码器的未知协议分类方法 [J]. 通信学报, 2020, 41(6): 88-97.
GU Chunxiang, WU Weisen, SHI Yanan, et al. Method of unknown protocol classification based on autoencoder [J]. Journal on Communications, 2020, 41(6): 88-97.
- [4] SYARIF I, PRUGEL-BENNETT A, WILLS G. Unsupervised clustering approach for network anomaly detection [C]// 4th International Conference on Networked Digital Technologies. Berlin, Germany: Springer Verlag, 2012: 135-145.
- [5] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [EB/OL]. [2020-03-20]. <https://arxiv.org/abs/1609.02907>.
- [6] PACHECO F, EXPOSITO E, GINESTE M, et al. Towards the deployment of machine learning solutions in network traffic classification: a systematic survey [J]. IEEE Communications Surveys & Tutorials, 2018, 21(2): 1988-2014.
- [7] ILIYASU A S, DENG Huifang. Semi-supervised encrypted traffic classification with deep convolutional generative adversarial networks [J]. IEEE Access, 2019, 8: 118-126.
- [8] 王勇,周慧怡,俸皓,等. 基于深度卷积神经网络的网络流量分类方法 [J]. 通信学报, 2018, 39(1): 14-23.
- [9] WANG Yong, ZHOU Huiyi, FENG Hao, et al. Network traffic classification method basing on CNN [J]. Journal on Communications, 2018, 39(1): 14-23.
- [9] LI Y, WANG Y, LIU Q, et al. Incremental semi-supervised learning on streaming data [J]. Pattern Recognition, 2019, 88: 383-396.
- [10] GU C, ZHANG S, XUE X, et al. Online wireless mesh network traffic classification using machine learning [J]. Journal of Computational Information Systems, 2011, 7(5): 1524-1532.
- [11] FURAO S, OGURA T, HASEGAWA O. An enhanced self-organizing incremental neural network for online unsupervised learning [J]. Neural Networks, 2007, 20(8): 893-903.
- [12] SHEN F, HASEGAWA O. Self-organizing incremental neural network and its application [C]// 20th International Conference on Artificial Neural Networks. Berlin, Germany: Springer Verlag, 2010: 535-540.
- [13] REZAEI S, LIU X. Deep learning for encrypted traffic classification: an overview [J]. IEEE Communications Magazine, 2019, 57(5): 76-81.
- [14] ASAD M, ASIM M, JAVED T, et al. DeepDetect: detection of distributed denial of service attacks using deep learning [J]. The Computer Journal, 2020, 63(7): 983-994.
- [15] FURAO S, HASEGAWA O. An incremental network for on-line unsupervised classification and topology learning [J]. Neural Networks, 2006, 19(1): 90-106.
- [16] FURAO S, OGURA T, HASEGAWA O. An enhanced self-organizing incremental neural network for online unsupervised learning [J]. Neural Networks, 2007, 20(8): 893-903.
- [17] MCMAHAN H B. Follow-the-regularized-leader and mirror descent: equivalence theorems and L1 regularization [C]// 14th International Conference on Artificial Intelligence and Statistics. Brookline, MA, USA: MIT Press, 2011: 525-533.
- [18] 邱天宇,申富饶,赵金熙. 自组织增量学习神经网络综述 [J]. 软件学报, 2016, 27(9): 2230-2247.
QIU Tianyu, SHEN Furao, ZHAO Jinxi. Review of self-organizing incremental neural network [J]. Journal of Software, 2016, 27(9): 2230-2247.
- [19] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques [C]// SIGMETRICS 2005: International Conference on Measurement and Modeling of Computer Systems. New York, USA: ACM, 2005: 50-60.

(下转第 78 页)

- 2018: 4490-4499.
- [12] ANANDAKUMAR M, RAMIYA, RAMA R, et al. A super voxel-based spectro-spatial approach for 3D urban point cloud labelling [J]. *International Journal of Remote Sensing*, 2016, 37(17): 4172-4200.
- [13] 邵磊, 董广军, 于英, 等. 结合多尺度体素和高阶条件随机场的点云分类 [J]. *计算机辅助设计与图形学学报*, 2019, 31(3): 385-392.
- SHAO Lei, DONG Guangjun, YU Ying, et al. A point cloud classification method based on multi-scale voxel and higher order random fields [J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(3): 385-392.
- [14] CHARLES R Q, SU HAO, KAICHUN M, et al. PointNet: deep learning on point sets for 3D classification and segmentation [C]// 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 77-85.
- [15] CHARLES R Q, YI Li, SU Hao, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [C]// 31st Annual Conference on Neural Information Processing Systems. New York, USA: Neural Information Processing Systems Foundation, 2017: 5100-5109.
- [16] 白静, 徐浩钧. MSP-Net: 多尺度点云分类网络 [J]. *计算机辅助设计与图形学学报*, 2019, 31(11): 1917-1924.
- BAI Jing, XU Haojun. MSP-Net: multi-scale point cloud classification network [J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(11): 1917-1924.
- [17] SIMONOVSKY M, KOMODAKIS N. Dynamic edge-conditioned filters in convolutional neural networks on graphs [C]// 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2017: 29-38.
- [18] WANG Yue, SUN Yongbin, LIU Ziwei, et al. Dynamic graph CNN for learning on point clouds [J]. *ACM Transactions on Graphics*, 2019, 38(5): 112-124.
- [19] FAUVEL M, CHANUSSOT J, BENEDIKTSSON J A, et al. A spatial-spectral kernel-based approach for the classification of remote-sensing images [J]. *Pattern Recognition*, 2012, 45(1): 381-392.
- [20] YOUSEFHUSSEIN M, KELBE D, IENTILUCCI EJ, et al. A multi-scale fully convolutional network for semantic labeling of 3D point clouds [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017, 143(5): 191-204.
- [21] ZHAO Ruibing, PANG Mingyong, WANG Jidong. Classifying airborne lidar point clouds via deep features learned by a multi-scale convolutional neural network [J]. *International Journal of Geographical Information Science*, 2018, 32(5): 960-979.
- [22] WANG Zhen, ZHANG Liliqiang, ZHANG Liang, et al. A deep neural network with spatial pooling (DNNSP) for 3D point cloud classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(8): 4594-4604.

(编辑 刘杨)

(上接第 69 页)

- [20] DRAPER-GIL G, LASHKARI A H, MAMUN M S I, et al. Characterization of encrypted and VPN traffic using time-related [C]// Proceedings of the 2nd International Conference on Information Systems Security and Privacy. Rome, Italy: SciTePress, 2016: 407-414.
- [21] SUN G, CHEN T, SU Y, et al. Internet traffic classification based on incremental support vector machines [J]. *Mobile Networks and Applications*, 2018, 23(4): 789-796.
- [22] MUNTHER A, RAZIF R, ABUALHAJ M, et al. A preliminary performance evaluation of K-means, KNN and EM unsupervised machine learning methods for network flow classification [J]. *International Journal of Electrical and Computer Engineering*, 2016, 6(2): 778-784.
- [23] ZHAO Zhirou, MEHROTRA K G, MOHAN C K. Online anomaly detection using random forest [C]// 31st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Berlin, Germany: Springer Verlag, 2018: 135-147.
- [24] WANG W, ZHU M, WANG J, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks [C]// 15th IEEE International Conference on Intelligence and Security Informatics. Piscataway, NJ, USA: IEEE, 2017: 43-48.

(编辑 刘杨)