具有显著提高准确率和鲁棒性的基于极限学习机的流量分类*

施 燕 陈荣荣 刘亚帆 顿 涵
(上海大学 特种光纤与光接入网省部共建重点实验室 上海 200072)

摘 要：流量分类是网络管理员进行网络流量监控从而实现有效管理的重要手段。因此，准确地对流量进行分类具有重要意义。流量分类的两个重要评判标准是分类器的准确率和效率。本文提出了一种准确率高、鲁棒性强的流量分类方案，该方案第一次将最近几年提出的一种新的机器学习算法-极限学习机引入网络流量分类领域进行研究并进行针对性优化。同时也提出了一种自动生成流量分类器训练集的方案，使该系统对新的网络应用具有更强的自适应性和扩展性。本文使用 VoIP 和 WWW 流量作为流量分类的两个类别。实验结果表明该方案相比其他文献提出的 C4.5，RandomForest，NaiveBayes 和 KNN 具有更高的准确率、稳定性和鲁棒性。其中当测试数据集在训练数据集后当天收集时，本文分类器具有 93%的高准确率，其他算法具有类似的准确率；当测试数据集在训练数据集后 1 月和 2 月收集时，本文分类器仍保持 85%的高准确率，而其它算法的准确率只有大概 60%左右，具有较大偏差无法应用到实际的流量分类系统中。实验结果表明，提出的流量分类方案具有准确率高，鲁棒性和扩展性强，可应用到流量分类实践中。

关键词：流量分类；机器学习；极限学习机；VoIP；WWW
中图分类号：TN915 **文献标识码**：A **国家标准学科分类代码**：510.4030

Extreme learning machine based traffic classification with significant improvement on accuracy and robustness

Shi Yan Chen Rongrong Liu Yafan Dun Han
(Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200072, China)

Abstract: Network traffic classification plays an important role for network administrators to supervise the traffic flows in order to manage the network efficiently. Therefore, accurate classification of traffic flow is of great significance. The quality of traffic classification lies in the classifier's accuracy and efficiency. This paper firstly implements an accurate and robust traffic classification solution using a recent new machine learning algorithm "Extreme Learning Machine". This paper also proposes a way to automatically generating training dataset for traffic classifier, making the system adaptable and scalable for new network applications. In this paper, VoIP and WWW traffic are used as two categories for the traffic classification. Experiment results indicate that this solution is highly accurate, more stable and robust for the classification of our traffic flow samples compared with other methods such as C4.5, Random Forest, Na? ve Bayes and KNN sited in other literatures. The classifier proposed in this paper has 93% accuracy when test dataset collected after training dataset, and other algorithms have similar accuracy. When test dataset collected one and two months later than training dataset, it still keep 85% high accuracy while other algorithm only reach about 60% at most, whose accuracy deviation are too large to apply to the practical traffic classification system. Thus, it is quite accurate and robust, with great scalable for engineering practice.

Keywords: traffic classification; machine Learning; extreme learning machine; VoIP; WWW

1 引 言

随着网络的高速发展，越来越多的新网络应用不断涌现。流量分类是流量统计、流量监控、入侵检测及提高服务质量等网络行为的基础。因此，准确的进行流量分类具有重要意义。

收稿日期：2015-12
* 基金项目：国家自然科学基金重点（61420106011）、上海市重点学科(15511105400)资助项目

流量分类方法主要有：基于端口分类、基于 DPI 分类及基于 DFI 分类。传统的基于端口分类的技术通过对比 IANA 分配的端口号与数据包包头的端口号来进行流量分类，其高效快速，易于实现，但对于使用动态端口及伪造端口的应用无法识别。DPI 分类技术解决了基于端口号分类的问题，通过提取数据包的应用特征来准确识别流量，但需获取数据包的负载，效率低，且对于内容加密的应用无法识别。为了解决端口和 DPI 分类技术的局限性，DFI 技术通过分析流量的统计特征来进行流量分类，其中机器学习算法常用于 DFI 分类技术[1]，该方法高效且具备很强的鲁棒性。机器学习的一个重要的挑战是特征选择，本文选用包长和包间隔作为流量统计特征[2]。已经有很多机器学习算法用于流量分类：Moore 等人采用 NaiveBayes[3] 和贝叶斯神经网络[4] 进行流量分类；Munther 等人对 C4.5 和 RandomForest 流量分类进行对比[5]。

如今，WWW 和 VoIP 流量占据大量网络流量。web 浏览器访问网站时会产生大量 WWW 流量，VoIP 应用程序如 Skype 也会在视频会议和视频通信中产生很多流量。对于这些流量的分类非常重要，如 IT 需要识别视频会议流量并对其分配更多的网络带宽以确保流畅的视频会议。

最近，黄广斌[6] 提出一种称为极限学习机（ELM）的快速机器学习算法，其基于具有任意隐层节点的单隐层前馈神经网络（SLFN）。ELM 随机选择输入权重和隐层偏置，通过 Moore-Penrose 广义逆来得到相应的输出权重。相比传统的基于梯度下降的算法，ELM 不但学习速度快并且具备良好的泛化性能，同时也解决了梯度下降法的一些问题如训练速度慢、容易陷入局部极小点、学习率的选择敏感等。

由于 ELM 具有这些优点，它已被应用到很多分类领域，如模式分类[7]、前车检测[8]、人类行为分类[9] 和图像分类[10] 等。但是 ELM 暂时还未应用到流量分类中。本文首次针对 VoIP 和 WWW 流量实现 ELM 的流量分类器并优化验证，相比其他机器学习算法，获得了很高的准确率和很强的鲁棒性。

2　实　　现

ELM 流量分类系统架构如图 1 所示。ELM 分类器是重要的组件，其实现了 ELM 算法。大部分机器学习分类系统在流量分类之前必须先训练得到一个分类器，并且很多流量分类系统依赖一个手动分类好的训练数据集，这对于新的网络应用具有局限性，无法自适应。为了应对越来越多的新应用出现，提出了一种自动收集训练数据系统，可用来训练流量分类器。

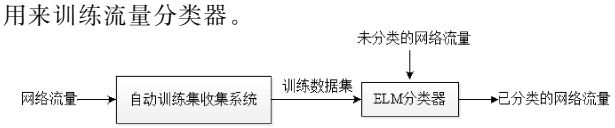

图 1　ELM 流量分类系统架构

2.1　ELM 算法

ELM 结构如图 2 所示，从结构来看，ELM 是 1 个 SLFN。相比传统的神经网络算法（如后向传播算法，基于梯度下降迭代），ELM 不需要任何迭代来调整输入权重和隐层偏置。在 ELM 算法中，一旦输入权重和隐层偏置被随机确定，隐层输出矩阵就唯一确定。训练单隐层神经网络可以转化为求解一个线性系统。ELM 算法具体步骤[11] 如下：

Step1：随机选择输入层权值 a_i 和隐层偏置，b_i，$i=1,2\cdots,N$；

Step2：计算隐层输出矩阵 $\boldsymbol{H}$；

Step3：计算输出权重矩阵 $\boldsymbol{\beta}=\boldsymbol{H}^{\dagger}\boldsymbol{Y}$，其中 $\boldsymbol{H}^{\dagger}=(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}$ 是 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆。

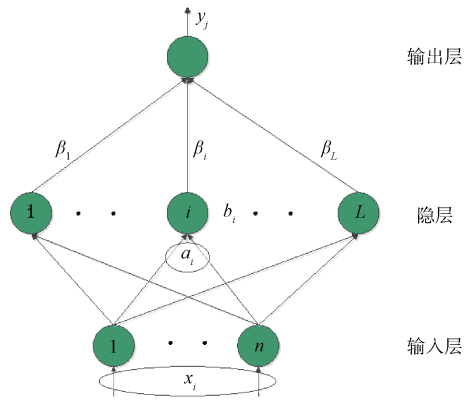

图 2　ELM 结构

ELM 算法的重要因素和挑战是：激活函数和隐层节点数。在图 2 中，输入层到隐层输出的函数叫做激活函数，激活函数有两个子函数：组合函数和转移函数，如图 3 所示。组合函数相当于多输入与多输出的映射，可以将输入线性或非线性的映射表示为 $S=g(a,b,x)$，其中 S 为内激活函数。转移函数将内激活函数变换到隐层输出。本文试验了不同的激活函数及隐层节点数来优化 ELM 网络流量分类器。

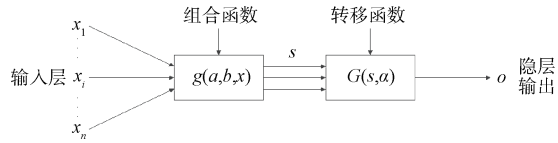

图 3　ELM 激活函数

2.2　自动数据收集系统和特征选择

训练数据集对于基于机器学习流量分类系统的准确率非常重要。大的数据集将提高准确率，鲁棒性和稳定性，但是，这需要花费很多时间来手动对网络流进行分类。另外，如果新的应用流量需要分类，它需要先收集训练数据并训练分类器，这对于需要快速对新应用进行流量分类的实际系统不实用。因此，本文提出并设计

了一个自动数据收集系统并将其集成到 ELM 流量分类系统中。

如图 4 所示,socket-hook 和 libpcap 用于自动收集流量训练数据。首先,系统使用 socket-hook 钩子技术来拦截所有的套接字函数,从而获取应用信息以及五元组信息。五元组信息包括源地址 IP,目的地址 IP,源端口,目的端口和协议,用于唯一的标识流量。同时,libpcap 会抓取数据包并分析其五元组信息及流量统计特征信息,然后,进行五元组信息匹配。如果五元组匹配,将提取流量统计特征以及应用信息,这些信息将用于训练 ELM 分类器的训练数据集。

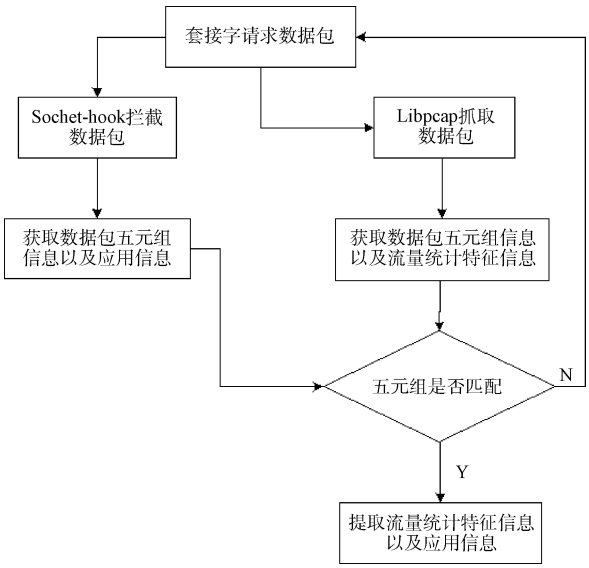

图 4　自动数据收集系统

本文选择数据包的包长和包间隔及其变种作为流量统计特征,即:最小包长、最大包长、平均包长、包长方差、最小包间隔、最大包间隔、平均包间隔以及包间隔方差。

3　实验和结果

3.1　不同激活函数的 ELM 分类器

常见的转移函数有 sina,tanh,sigmoid 和 gaussian 函数,组合函数通常是线性函数和径向基函数。线性函数的表达式为 $g(a,b,x)=ax+b$,径向基函数的表达式为 $g(a,b,x)=b\|x-a\|$,其中"a"是输入权重,"b" 是隐层偏置。本文使用不同的激活函数的 ELM 分类器来对 VoIP 和 WWW 流量进行分类,从而获取最佳的 ELM 分类器激活函数。由于输入权重和隐层偏置是随机选取的,所以每次运行输入权重和隐层偏置都会不同。因此,为了避免随机性导致偏差,本文对每个激活函数运行 20 次然后计算其流量分类的平均准确率和平均分类时间来评价 ELM 分类器。在该实验中,隐层节点数为 40,测试数据流量为2 200,测试结果如表 1 所示。

表 1　不同激活函数的 ELM 分类器

激活函数		准确率	时间
组合函数	转移函数		
线性函数	sine	0.213 95	0.006 01
	tanh	0.445 61	0.001 99
	sigmoid	0.445 73	0.005 99
	gaussian	0.341 84	0.006 02
径向基函数	sine	0.682 13	0.003 04
	tanh	0.661 98	0.003 99
	sigmoid	0.661 01	0.006 01
	gaussian	0.800 67	0.005 99

从表 1 可以看出,总体来说径向基组合函数具有更高的准确率,更短的分类时间以及更强的稳定性。并且,径向基组合函数和 gaussian 转移函数具有最高的准确率和相对高效的时间。因此,径向基函数组合函数和 gaussian 转移函数的激活函数是 ELM 流量分类器的最佳选择。

3.2　不同隐层节点数的 ELM 分类器

从上面的实验来看,总体准确率并不是很高。这里研究 ELM 流量分类器另外一个重要因素—隐层节点数。本文对不同隐层节点数的 ELM 流量分类器进行实验。结果如图 5 所示。

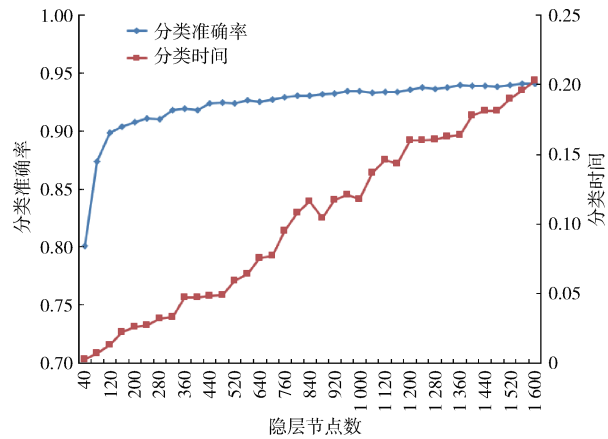

图 5　不同隐层节点数的 ELM 分类器

从图 5 可以看出,总体上,随着隐层节点数的增加,流量分类准确率也提高,并且当隐层节点数为 960 时,准确率达到相对稳定,此时准确率为 0.934 74。但是,隐层节点数增加,流量分类时间也明显增加,这是因为隐层节点数增加会导致矩阵变得更大从而需要更多的计算。随着隐层节点数的增加,时间有可能在某些局部范围降低,这是因为 ELM 分类器的输入权重和隐层偏置是随机产生的。

综合考虑流量分类的准确率和时间,隐层节点数为 960 是最佳参数,此时有很高的准确率和相对快的分类速度。

3.3　ELM 流量分类器进行流量分类

为了更好的验证 ELM 流量分类器的实用性，本文进一步实验并且和其他机器学习算法进行比较。

首先，使用自动数据收集系统收集了 20 000 个 WWW 和 VoIP 流作为训练数据集，训练不同机器学习算法实现的流量分类器。收集完训练数据后，再次收集 10 000 个流作为测试数据集并随机从测试数据集中选择一些流来测试，从而比较不同算法的分类器，结果如图 6 所示。

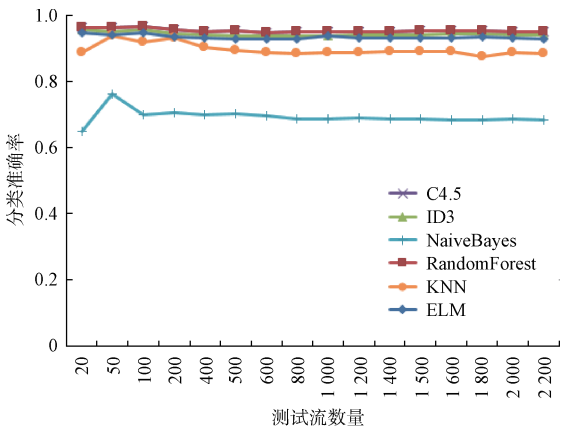

图 6　ELM 和其他算法流量分类比较

从图 6 可以看出，C4.5，ID3，RandomForest 算法有较高的准确率，接近 95%，ELM 准确率是 93%。同时，对于不同数量的流，C4.5，ID3，RandomForest 和 ELM 都能稳定的保持很高的准确率。但是，对刚刚收集完训练数据集后的流进行分类不符合实际系统和工程实践，实际系统中，流量分类发生在训练好的流量分类器之后，并且希望尽可能保持较长的生命周期。本文进一步通过实验来评估这种场景。本文在 1 个月后收集了 10 000 个 WWW 和 VoIP 数据流，随机从中选择一些流来测试上面已训练的分类器。结果如图 7 所示。

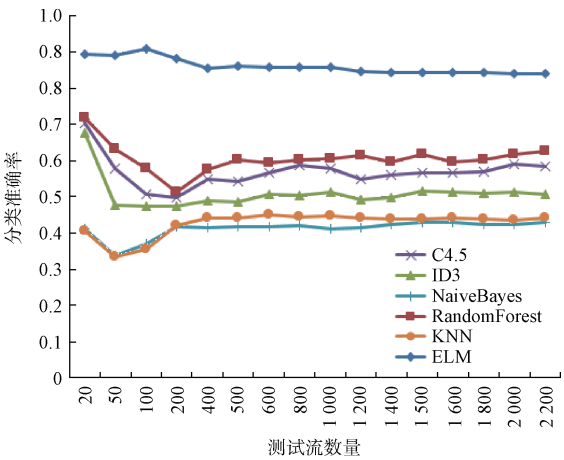

图 7　ELM 和其他算法流量分类比较

实验结果表明，在该场景下，ELM 仍具备很高的准确率，大约 85% 左右；但是，其它算法如 RandomForest，C4.5，ID3 准确率大幅度下降到 60%左右甚至更低。因此，当对一段时间后的流进行分类时，ELM 更稳定并且具有高的准确率，但是其他机器学习算法的准确率降低很多。

图 6 和图 7 对比可知，流量分类器在测试数据流数量小时由于数据流较大的随机性分类比较不稳定，当测试样本数达到 1 400 时，由于此时随机性非常小分类比较稳定。显然，1 个月后收集的测试数据集与训练数据集的相关性比在同一天收集的测试数据集与训练数据集的相关性更小。因此，当测试数据集与训练数据集相关性小时，ELM 分类器也能保持很高的准确率，而其他算法准确率则有很大的降低。为了进一步验证该结果，本文在 2 个月后收集了另外的 10 000 个 WWW 和 VoIP 数据流，并从中随机选择 1 400 个测试数据流进行流量分类。从不同测试数据集中随机选择 1 400 个测试数据流进行流量分类的比较如图 8 所示。

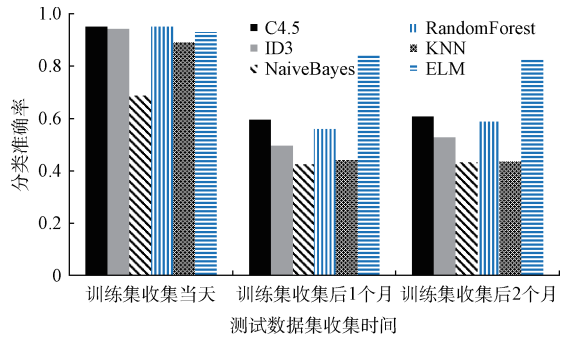

图 8　不同收集时间的测试数据集流量分类准确率对比

从图 8 对比结果来看，ELM 分类器对于不同的测试数据集稳定性更好，鲁棒性更强，准确率变化小。因此，ELM 分类器能很好地应用到实际的流量分类系统中。

4　总结与展望

本文首次提出了一种准确率高鲁棒性强的基于 ELM 的流量分类方案，同时也提出了一种自动获取训练数据集来使分类器能够快速应用到新的应用的数据流。为了验证评估 ELM 分类器的性能，做了大量实验来优化 ELM 分类器并且同其它常用于流量分类的机器学习算法进行对比。其中组合函数、转移函数及隐层节点数是影响 ELM 分类器的重要因素。实验结果表明，gaussian 转移函数和径向基组合函数的激活函数以及隐层节点数 960 时，ELM 分类器具有最佳的准确率和时间。与其他常用于流量分类的机器学习算法如 C4.5，ID3，RandomForest 对比，当测试数据集在训练数据集收集之后立即收集时，ELM 和其它算法基本上都具有 93%的高准确率；同时当测试数据集在训练数据集收集的 1 个月及 2 个月后收集时仍保持 85%的高准确

率,然而其他算法准确率却降低了至少 30%,即 60%左右。因此 ELM 鲁棒性更强,扩展性更强,准确率高,满足工程实际。

尽管本文 ELM 分类器用于流量分类已经获得非常好的结果,但仍然有很多需要探索和提升。ELM 算法有一些变种像核版 ELM 可能具有更好的性能。同时,可以验证更多的流量类型以及更长时间后的流量样本来评估本文的 ELM 流量分类器。这些将会是以后的研究方向。

参考文献

[1] 周文刚, 陈雷霆, LUBOMIR BIC,等. 基于半监督的网络流量分类识别算法[J]. 电子测量与仪器学报, 2014, 28(4):381-386.

[2] 黄志根, 陈健, 王珊. 一种基于包长和时间间隔的网络流量分类方法[J]. 电子测量技术, 2011,34(11): 109-112.

[3] MOORE A, ZUEV D. Internet traffic classification using bayesian analysis techniques [J]. ACM Sigmetrics Performance Evaluation Review, 2005, 33(1): 50-60.

[4] AULD T, MOORE A W, GULL S F. Bayesian neural networks for internet traffic classification. IEEE Transactions on Neural Networks, 2007: 223-239.

[5] MUNTHER A, ALALOUSI A, NIZAM S, et al. Network traffic classification-A comparative study of two common decision tree methods: C4.5 and random forest[C]. International Conference of Electronic Design (ICED), 2014: 210-214.

[6] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: A new learning scheme of feedforward neural networks[C]. Proceedings of International Joint Conference on Neural Networks (IJCNN2004), 2004: 985-990.

[7] SHEN Y W, KEEM S Y. On equivalence of FIS and ELM for interpretable rule-based knowledge representation[C]. Proceedings of Neural Networks and Learning Systems, 2015: 1417-1430.

[8] 张楚金, 王耀南, 卢笑,等. 基于假设验证和改进 HOG 特征的前车检测算法[J]. 电子测量与仪器学报, 2015,29(2):165-171.

[9] BUDIMAN A, FANANY M I, BASARUDDIN C. Constructive, robust and adaptive OS-ELM in human action recognition [C]. Industrial Automation, Information and Communications Technology (IAICT), 2014: 39-45.

[10] LU B, DUAN X, WANG C. A novel approach for image classification based on extreme learning machine[C]. Information Science and Technology (ICIST), 2014: 381-384.

[11] 陈绍炜, 柳光峰, 冶帅,等. 基于蝙蝠算法优化 ELM 的模拟电路故障诊断研究[J]. 电子测量技术, 2015, 38(2): 138-141.

作者简介

施燕,在读硕士,研究方向为基于机器学习的流量分类。

E-mail:shiyan2016@126.com