基于机器学习的IP流量分类研究

刘 琼 刘 珍 黄 敏
(华南理工大学软件学院 广州 510006)

摘 要 IP流量分类是Internet研究和流量工程的重要基础,近年来网络应用类别和Internet流数量在快速增长。流量分类技术不断面临新的挑战。对基于机器学习的IP流量分类方法进行了系统性研究。给出了这类流量分类方法的数学描述;通过深入研究有监督和无监督机器学习方法在流量分类中的应用,从数据预处理、模型构建和模型评估3个方面评述这类技术的研究现状,并指出存在的问题;总结得出现阶段基于机器学习的IP流量分类技术存在数据偏斜、标识瓶颈、属性变化和实时分类等4个方面的共性问题;最后展望了流量分类技术的未来发展方向并介绍了作者正在进行的工作。

关键词 流量分类,机器学习,网络流,网络测量
中图法分类号 TP393.06 **文献标识码** A

Study on Internet Traffic Classification Using Machine Learning

LIU Qiong LIU Zhen HUANG Min
(School of Software Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract Internet traffic classification is one of the key foundations for research works and traffic engineering in Internet. The categories of Internet applications and the number of Internet flows increase fast last years. The technique challenges are coped with development of traffic classification all the time. A research was carried out systematically in Internet traffic classification using machine learning in the paper. A mathematical description was given for the technology of traffic classification first. After surveying the technology of traffic classification based on supervised learning and unsupervised learning, we commented on the data preprocessing, classification model and performance evaluation etc. three aspects, and indicated the shortages now in these studies. Then, four common issues were summarized, which are data skew, label bottleneck, attribute changing and classify timely respectively. At the end, the prospect and our going work in this area were pointed out.

Keywords Traffic classification, Machine learning, Traffic flow, Network measurements

1 引言

随着Internet规模和应用范围的日益扩大,流量监控、网络管理以及计费服务等备受重视。网络运营者进行网络规划需要分析主流网络应用的行为;Internet服务提供商对网络资源进行QoS管理需要区分网络应用;网络安全需要识别恶意流等。IP流量分类或应用识别已成为网络研究和网络应用的重要基础。

然而,互联网应用数量的迅速增加和应用类别的复杂多变使网络流量分类不断面临新的挑战。随着端口跳变技术、伪装技术的大量使用,传统的利用熟知端口号(known ports)识别网络应用[1]的方法已经不再可靠。通过解析应用层协议匹配特征(signature matching)字段的基于报文载荷(payloads)的流量分类方法[2,3,6,7],曾经被誉为最为准确的流量识别方法目前仍为大多数商用系统所采用。但是,面对应用协议的频繁更新、载荷加密技术的普及、新应用频出等状况,特征匹配的有效性在下降;而且维护特征库耗费计算资源和计算时间,适用范围非常有限。基于传输层行为模式的流量分类方法[5,44],虽然不依赖端口号或报文载荷内容,有较好的可扩展性,但是网络应用的传输层行为极易受到网络环境的影响,而且随着网络应用的自身完善,这类方法的可用性仍然受到限制。由此可知,利用网络应用的外在属性进行浏览分类,前景不容乐观,因而迫切需要引入新的理论和方法,从根本上解决流量分类技术的可持续发展问题。

基于机器学习(ML, Machine Learning)分类或识别网络应用是近年的研究热点[8]。这类技术不依赖匹配协议端口或解析协议内容识别网络应用,而是利用从被传输的流量(traffic)数据中抽取的"流"的先验特征(priori knowledge)或计算流的统计信息(statistical information)达到区分网络应用的目的,这类方法不受动态端口、载荷特征甚至网络地址转换(NAT)的影响,其效果、灵活性以及可扩展性等,较之前述各种方法都有所突破。这对解决流量分类及相关问题,特别是

到稿日期:2010-01-12 返修日期:2010-03-31 本文受华南理工大学985科研启动经费(D606001II)资助。
刘 琼(1959—),女,博士,教授,博士生导师,主要研究方向为多媒体通信技术、网络测量,E-mail:liuqiong@scut.edu.cn;刘 珍(1986—),女,博士生,主要研究方向为网络测量、流量分类;黄 敏(1978—),女,博士,副教授,主要研究方向为计算机网络技术。

对于我国大量使用内网地址进行互联网流量分类尤为重要。

本文系统研究基于 ML 进行 IP 流量分类的方法，首先论述了运营网中进行 IP 流量分类的重要意义；指出传统 IP 流量分类技术存在的不足，简述基于 ML 进行 IP 流量分类的技术优势。第 2 节介绍相关概念，对基于 ML 进行 IP 流量分类的方法进行数学描述，归纳总结研究内容和分析步骤，并研究和描述了分类器评价指标。第 3 节深入分析有监督和无监督机器学习方法在流量分类中的应用，从数据预处理、分类器和分类器评估 3 个方面评述近年来的主要进展，指出其中存在的不足和亟待解决的问题。第 4 节总结得出现阶段基于机器学习的 IP 流量分类研究中存在数据偏斜、标识瓶颈、属性变化和实时分类等 4 个方面的共性问题，指出了未来的发展方向。最后总结全文并简单介绍了正在开展的工作。

2 ML 方法与 IP 流量分类

2.1 基本概念

单向流(Uni-directional flow)：由共享 5 元组(five-tuple)的 IP 报文(packets)组成，5 元组包括源/目的 IP 地址、源/目的 IP 端口号和网络层协议代码。

双向流(Bi-directional flow)：在同一源/目的 IP 之间的 TCP 连接中，包含方向相反的两个单向流，两个完整单向流集合称作全流(Full-flow)。

流量类别(class)：一种流量类别通常指由一个或多个相同网络应用引发的 IP 流量。

实例(instances or examples)：通常指单向流，可能是完整单向流，也可能只是完整单向流的一部分。在 ML 中，每个实例由一个特征(feature)值向量表述，每个特征值代表这个实例的一种属性。

机器学习(ML)[8]：按某种预定模式模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，使之不断改善自身的性能。流量分类 ML 的输入是一个实例集(特征值向量组成的矩阵，dataset)，输出是已学到知识的描述或规则。

IP 流量分类一般包括两种[10]：(1)从 IP 混合流(如 HTTP、DNS、SSH、P2P、游戏等混合流)中识别并提取具有兴趣特征的个别流，如在线游戏(online game)流或 P2P 流等；(2)区分 IP 混合流中所有的网络应用类别，这种流量分类更具有一般性。

一般来说，分类方法面向特定任务，构造普适分类器难度较大。目前用于 IP 流量分类的 ML 方法主要分为两类：监督学习(supervised learning)和非监督学习(unsupervised learning)。

2.2 ML 方法

监督学习算法用于已知类别的数据集，创建出输入/输出之间的关系规则(分类器)，再利用分类器将输入的未知特征向量映射至已知输出类别。知识关系规则可以表达为流程图(flowchart)、决策树(decision tree)等分类规则(classification rules)。监督学习流量分类器的建立过程包括两个步骤[8]：学习阶段(training phase)通过探察(examine)训练集，构建分类器；验证阶段(testing phase)采用训练好的分类器分类已标识的实例集(作为未知实例集)。

基于监督学习的流量分类方法描述如下：

设训练数据集 TS 由分属 M 类的 N 个实例组成，则

$$TS=\{\langle x_1,y_1\rangle,\langle x_2,y_1\rangle,\cdots,\langle x_N,y_M\rangle\} \quad (1)$$

每个实例以元组 $\langle x_i,y_j\rangle$ 表示($i=1,2,\cdots,N;j=1,2,\cdots,M$)，其中 x_i 表示第 i 个实例的特征值向量，y_j 表示输出类 j 的值，训练集 TS 中的预标识输出类值为 $\{y_1,y_2,\cdots,y_m\}$。学习阶段的任务是通过输入 TS 来确定关于输入特征向量的函数 $f(x)$(分类算法)的相关参数。

测试集(ts，testing dataset)包含与训练集 TS 一致的预标识输出类值 $\{y_1,y_2,\cdots,y_m\}$ 及其特征向量集，验证过程将 ts 作为未知集，利用分类器 $f(x)$ 对 ts 进行类映射，并进行性能评价；根据评价结论接受或进一步改进分类器。

目前已有不少监督学习算法用于 IP 流量分类，如决策树(Decision tree)算法、朴素贝叶斯(Naïve Bayes)算法等；各算法的差别在于分类模型的构建和算法的优化。

非监督学习方法又称为聚类(clustering)法，与监督学习的主要差别在于聚类方法无需预标识输出类别 $\{y_1,y_2,\cdots,y_m\}$，而是采用内在启发式(internalized heuristics)方法进行聚簇或分组，使具有类似属性的实例向量相聚成簇(clusters or groups)。相似性由距离度量法以及相似函数的非度量法衡量，如欧氏空间(Euclidean space)。数据集实例成簇的方式可能是全分离的，每个实例仅属于一组(类)；也可能出现重叠(overlapping)。一个实例可能按不同概率落入几个组(类)。聚簇还可以分层完成，归入顶级的实例由合并(自底向上)和分裂(自顶向下)两种方法来完成分层聚类，等等。聚类完成后，再根据需要对所形成的“簇”进行类标识。如采用 K 均值(K-means algorithm)算法，实例集成员按数值域划分成簇；采用增量聚簇(incremental cluster)算法，实例集成员形成层次分组形态；采用概率聚类(incremental cluster)算法，实例集成员按概率归类。聚簇方法的类标识代价将大大减小。

2.3 模型评价指标

设实例集合包括 N 个样本，不失一般性，建立表 1 所列的 $m(m\geqslant 2)$元分类问题的混淆指标矩阵。

表 1 m 元分类问题的混淆指标矩阵

Classes		Prediction				
		1	2	3	…	n
Actual	1	n_{11}	n_{12}	n_{13}	…	n_{1m}
	2	n_{21}	n_{22}	n_{23}	…	n_{2m}
	3	n_{31}	n_{32}	n_{33}	…	n_{3m}
	⋮	⋮	⋮	⋮	⋮	⋮
	m	n_{m1}	n_{m2}	n_{m3}	…	n_{mm}

在基于机器学习的 IP 流量分类研究中，类别预测准确率(A，Accuracy)是一项最为关键的评估指标(metric)，它反映了分类器对未知数据的判断(decisions)能力。此外，还有两项指标，分别是分类精度(P，Precision)和召回率(R，Recall)。不失一般性，对表 1 所列的 m 元分类问题进行二类简化。设兴趣流量类别为 X，非 X 类别为 $\overline{X}$，则有准确率 A：对参加分类的实例集合，被正确分类的实例数占实例(instances)总数的百分比例；召回率 R：被正确分类为 X 的实例数占 X 实例总数的百分比例；精度 P：被正确分类为 X 的实例数占被分类为 X 的实例总数的百分比例。

定义 4 个中间指标变量：假负(FN，false negative)、假正(FP，false positive)、真正(TP，true positive)和真负(TN，true negative)。表 2 表达了 4 个中间指标变量与二类 X、$\overline{X}$ 之间

的关系，则有 FN：被错误分类为 $\overline{X}$ 的样本数占 X 实例总数的百分比例；FP：被错误分类为 X 的样本数占 $\overline{X}$ 类样本总数的百分比例；TP：被正确分类为 X 的样本数占 X 类样本总数的百分比例（$TP=100\%-FN$），即召回率 R；TN：被正确分类为 $\overline{X}$ 的样本数占 $\overline{X}$ 类样本总数的百分比例（$TN=100\%-FP$）。

表 2　二类问题的混淆指标矩阵

Abstract classes		Prediction	
		X	$\overline{X}$
Actual	X	TP	FN
	$\overline{X}$	FP	TN

显然，优质分类器希望 FP 和 FN 较小；按百分比字节数量或流的数量表示，准确率 A、召回率 R 和精度 P 可分别表示为

$$A=\frac{(TP+TN)}{FN+FP+TP+TN}\times 100\% \tag{2}$$

$$R=TP \tag{3}$$

$$P=\frac{TP}{TP+FP}\times 100\% \tag{4}$$

2.4　研究内容

基于 ML 的 IP 流量分类方法，其关键是构建分类器。构建分类器包括数据预处理（Data Preprocessing）和分类器生成（Classifier Generating）两部分，如图 1 所示。

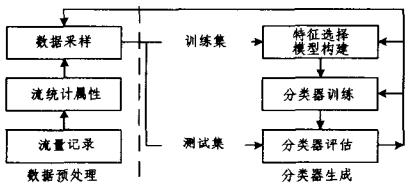

图 1　构建分类器

数据预处理主要包括通过网络测量设备捕获/收集（或人为生成）的混杂的 IP 流记录（traffic traces）；进行流分离；对每流记录进行统计处理（flow statistics processing），求取每流的各种属性值（如平均包长、包到达时间间隔的标准差、每流总长、包到达时间间隔等），构建每流属性向量；数据采样（Data Sampling）构造数据集（包括训练集和测试集）。

分类器生成包括训练集输入预选的分类算法，通过学习，求取并优化算法模型的参数；生成分类器；通过测试集进行分类验证和评估，判断分类器性能；认可或进一步优化、改善分类器。

3　研究现状

3.1　数据预处理

使用属性向量表述流量，是机器学习流量分类方法的基础。已有的 Internet 流属性种类繁多，而且大部分属性对网络环境变化比较敏感。如何选择既能有效区分流量，又具有一定稳定性的流属性，是数据预处理研究的核心问题。数据预处理过程包括成流（flow generation）、特征提取（feature extraction）、特征选择（feature selection）和采样构造数据集等多个环节。

随着网络规模的扩大和网络速度的提高，网络应用类别剧增，利用单报文信息进行流量分类不仅存储开销过大，而且准确性或可靠性都不尽如人意。目前的网络流量分类通常以“流（即实例）”作为基本处理单元，即截取大量数据报文后进一步“成流”。单向流收集简单，且有较好的推广性，适用于核心网络或边缘网络的流量分类，但是不能利用双向通信之间的关联性。双向流截取同一网络连接的双向报文序列，既能反映两个单向流的特征，也能反映成对单向流之间的关联性，为流量分类提供了更深层次的依据。但是路由的非对称性[11]使得双向的流并不总是经过同一链路，因此双向流数据多用于对边缘网络出口流量的分类研究。

特征提取指通过对个体流群体某种特征的特殊计算产生有代表性的网络应用行为的数值属性。Roughan 等人[11]率先使用流持续时间（flow duration）、平均报文长度（mean packet lengths）和报文到达间隔时间增量（inter-arrival variability metric）等属性将网络流映射至不同的 QoS 级别。由于只使用了少数几种属性，网络应用仅被粗略地划分为交互式应用、块传输应用、流媒体应用和事务性应用等 4 类。Moore 等人[12]通过直接计算或傅立叶变换等方法给出了双向流的 248 种统计属性，试图为基于机器学习的流量分类提供完整的属性集合。

然而，Williams 和 Zander 等人[13]指出由于网络速度较高，经有限时段截取的流量记录中通常包含数十万乃至数百万条流。如果对每条流的属性都进行诸如傅立叶变换之类的运算，计算开销难以接受。因此，他们选取了 22 种简单实用的双向流属性描述真实网络环境的流。为了进一步提高机器学习方法的在线分类能力，Bernaille 等人[14]提出测取每个 TCP 连接中的前 5 个数据报文的长度来构造流分类属性向量。该方法具有处理简单、实时性强的特点，但它存在网络拥塞的环境中，极易由于数据报文的乱序到达[5]而失效。为了有效处理网络流的行为特征在数据交换过程中动态变化的问题，Nguyen 和 Armitage[15,16]提出了多子流模型（sub-flow model）。该方法先将网络流按时间顺序划分为若干子流，通过组合多个子流的属性集合来构建新的流属性集。该方法能够提高分类器的准确性，但子流属性也存在依赖网络运行状况的问题，因此该方法的鲁棒性和推广性还有待进一步考察。

在机器学习流量分类中，使用无关（irrelevant）或冗余（redundant）特征对分类器的准确性有负面影响，但是过少的特征维数不足以对流量进行完美分类。选择有区分意义的特征子集是设计过程中最关键的一步。目前在网络流量分类中使用较多的特征选择方法是过滤方法（filter method）和包装方法（wrapper method）[17]。过滤法基于一般数据特征进行运作，在进行机器学习之前评估和选择最佳子集，它对机器学习算法是无偏的，而且这种方法计算简单、可移植性强，但不能保证给出最优的特征子集[18,19]。包装法在子集（subsets）选择过程中封装了机器学习算法，最终又将子集用于机器学习过程，虽然它对机器学习算法是有偏的，但是它能给出最优特征子集。包装法的缺点是计算开销过大，不适用于高维特征选择[20]。

训练集和测试数据集（datasets）：面对巨大规模的原始实例集，需要进行数据集采样，择出有效代表网络流量的训练集和测试集；控制 ML 算法的搜索空间；保证分类性能等。目前，仅有少数文献研究抽取不同数量实例组成数据集，进而对分类器性能做出实验比较。但这些研究论文仅笼统地将 ML 算法面对有限数量的预标识数据集，而且采用 N 倍交叉验证

(N-fold cross-validation)法或"留一"法组成训练/测试集,进行分类器的学习/验证[15]。如何判断采集方法或数据集空间对流量分类的影响,尚未查到专题研究论文。除此之外,先验知识(例如样本空间与算法之间的关系)在流量分类数据预处理中的作用,也未能查到系统性研究论文。

3.2 分类器

分类器指所选择机器学习算法经数据集训练学习得出的分类规则。目前,应用于流量分类的机器学习方法主要有监督(supervised)学习方法和非监督(unsupervised)学习方法。

监督学习方法通过数据集训练分类算法,且需要对每个输入样本提供类别标识。Roughan 等人[11]首次使用 K 近邻(K nearest neighbor, K-NN)、线性判别式分析(Linear Discriminant Analysis, LDA)和二次判别式分析(Quadratic Discriminant Analysis, QDA)等 3 种监督类机器学习算法对网络流量进行分类。然而 K 近邻方法需要逐个计算测试样本与训练样本之间的相似度,处理开销较大,而且分类性能极易受到噪声数据的干扰。为此,Zuev 和 Moore[21]提出了基于概率模型的朴素贝叶斯(naïve bayes)方法。该方法要求参与分类的各属性条件独立且服从高斯分布。在流量分类问题中,流属性集合很难满足这些条件,因此该方法的整体准确率只能达到 65%左右。为了避开条件独立和高斯分布假设,Moore 等人[17]采用 FCBF(Fast Correlation-Based Filter)算法对属性集合进行过滤,并使用核评估(kernel estimation)技术改进朴素贝叶斯方法。他们的工作使分类器的整体准确率提高至 95%以上。Auld 等人[22]提出的基于贝叶斯方法的神经网络模型,将分类准确率提高到 99%,而且能够以 95%的准确率分类 8 个月以后的网络流量。Williams 和 Zander 等人[13]对 NBD(naïve bayes using discretisation)、NBK(naïve bayes using kernel density estimation)[23]、C4.5(C4.5 Decision Tree)、BayesNet(bayes network)[24]、NBTree(naïve bayes tree)[25]等 5 种主要的监督方法进行了比较研究。结果表明,在这 5 种方法中,除了 NBK 的整体准确率在 80%左右,其他 4 种方法都超过了 90%。但是他们的比较仅得出了个别性能指标的区别,并没有分析产生这种区别的原因,甚至没有指出这些算法在论文条件下是否存在可比性,其参考意义有待考证。

综上,监督方法在流量分类中的应用有比较完整的理论依据,但是在类别预标识、数据集的代表性、分类器的推广能力和参数稳定性等方面存在瓶颈。而且监督学习方法计算资源消耗较大,对于实际应用,特别是实时流分类,存在难以逾越的问题。此外,由于监督方法的二类本质,有学者认为它更适合用于分类兴趣流[26]。

无监督学习算法无需类别预标识,对输入数据集自动形成"聚类(簇,cluster)",再进行由簇到流量类别的映射。2004 年 McGregor 等人[27]首次提出使用期望最大化算法(expectation maximization, EM)将流量聚簇映射到不同的网络应用类别,但没有分析基于聚簇算法的分类器性能。Zander 等人[28,29]基于 EM 算法的一种实现(AutoClass)对聚类方法的分类性能进行了分析。他们的实验结果表明,使用 AutoClass 方法,流量分类器整体准确率达到 87%左右。Erman 等人[30]对 K 均值(K-means)、DBSCAN(density-based spatial clustering of applications with noise)[31]和 AutoClass 等 3 种聚类算法进行了对比分析,结果表明,AutoClass 算法的模型训练时间较长,但分类的整体准确率最高;K 均值算法的模型训练时间较短,但整体准确率最低;DBSCAN 算法的计算性能和分类效果居中,这种算法不仅能够有效排除噪声干扰,而且能够将大量的网络流汇聚在少数几个大簇中,具有良好的推广能力。Erman 等人[32]将 AutoClass 算法与监督方法中的朴素贝叶斯方法进行了详细的比较,结果表明 AutoClass 的整体准确率略高于朴素贝叶斯算法,达到 91.19%。虽然 AutoClass 算法的训练时间远远大于朴素贝叶斯方法,但它能使 80%的网络流聚集到 50%的簇中。由此可见,利用 AutoClass 算法构建分类器,无需将所有簇都准确映射到相应的流量类别,有效减少了聚簇映射的时间开销。在 Bernaille 等人[33]提出的实时流量分类器中,采用了 K 均值、高斯混合模型(Gaussian mixture model)和谱聚类(spectral clustering)等 3 种非监督方法,结果表明这些聚类算法不仅能够以超过 90%的准确率有效分类知名应用流量,而且能够以 60%的概率发现新型网络应用。

收集并标识大规模样本集是非常耗费的工作,聚簇方法的优势不言而喻。但"簇"与"类"的映射不存在一一对应的关系,通常参考簇与类之间的概率倾向进行类别映射,但这种倾向并非一定存在;不同聚类算法,甚至同一聚类算法在选择不同的参数时都可能导致不同的结果,而且通常要求事先指定聚簇数目。目前,聚类方法在流量分类中的优势仅表现在发现网络新的应用方面。

无论是监督方法还是非监督方法,不同研究者使用或评价了不同的算法模型。如何裁决分类器的本质差异尚待进一步研究。

3.3 分类器评估

在流量分类中,对于整体准确率 A、分类精度 P、召回率 R 等指标,准确率应用最广。由式(2)、式(3)、式(4)可以看出,分类器评价指标之间存在关联性,需要根据应用折中取舍,近期基于 ML 的 IP 流量分类研究尚未顾及 R、P 指标的折中。这类问题的处理可以借鉴"受试者操作特征(ROC, receiver operating characteristic)"或"奈曼-皮尔逊准则(NPC, Neyman-Pearson criterion)"。ROC 曲线表达了 TP(作为函数)与 FP(作为自变量)之间的数值关系,便于在阈值范围分析模型性能;NPC 准则多用于在 FP 阈值范围内寻求最大 TP[26]。

流量分类评价指标计量多分为按"流"或按"字节"度量。长期以来,多数流量分类选择按"流"估算分类器性能指标,目前也有许多学者按"字节"估算分类器性能指标。Erman 等认为 Internet 绝大多数"流"为小流(mice flows),这些流占网络总字节流量的比例甚小;而 Internet 的极少数"流"为大流(elephant flows),占据了字节总流量的绝大部分。如果出于提供 IP 流 QoS 服务而进行流量分类,适宜按流估算分类器性能指标;如果出于运营商进行网络监管或载荷模式分析,则适宜按字节估算。

无监督 ML 分类器的性能评估是一个更为复杂的问题,且需要在聚类标识、形成分类规则之前进行。不同的聚类算法可能产生不同的成簇结果,甚至对于相同的算法,选取不同的参数或不同的聚类秩序都可能导致不同的成簇。让用户了解数据中究竟隐藏了多少"簇"或最佳的"簇"数值范围;判断最终的成"簇"是有意义的"类组"还仅只是纯算法结果。比较不同聚簇算法以及分类器的可用性和效率、簇内成员的相似

度、簇间成员的分离度、簇标识成本/资源等，都是亟待研究的重要课题。

Halkidi 等人从外部准则(external criteria)、内部准则(internal criteria)和相对准则(relative criteria)等方面研究了成簇的有效性。外部准则利用指定结构比较和验证聚类结果；内部准则通过检验数据集的遗传结构来评价聚类结果；相对准则注重在某些参数和假设条件下寻优聚类算法[34-38]。

分类器的效率是指分类器对未知样本进行分类的速度，以分类器对测试数据集上进行分类的时间来衡量。随着互联网速度的日益提高，分类器的效率对于网络流量的在线分类显得尤为重要。Bernaille 等人[14]提出使用在线分类的计算复杂度来评价流量分类器。但他们只给出了基于端口、基于特征字段方法和基于聚类方法等的计算性能的定性比较，没有给出具体的量化分析。Williams 等人[13]在分析 5 种机器学习算法时采用正态化分类速度(normalized classification speed)描述分类器的分类效率。从他们的实验结果来看，利用 C4.5 决策树分类器分类速度最快，NBD，Bayes Net，NB-Tree 和 NBK 的分类速度依次递减。

分类器的鲁棒性指分类器处理未知样本的时间稳定性和推广能力。时间稳定性按分类器准确率随测试数据集采集时间不同而产生的差异来衡量。Auld 等人[22]首次利用与训练数据集相差 8 个月的测试数据集评估了流量分类器，他们的实验结果表明基于贝叶斯神经网络的流量分类器具有良好的时间稳定性。分类器的推广能力以分类器准确性随测试数据集采集地点的变化所产生的差异进行评估。Zander 等人[39]利用不同采集地点的数据集对源自 C4.5 决策树算法的分类器进行了交叉验证(cross validation)，评估这种分类器的可推广性。

4 未来的挑战与方向

基于机器学习的流量分类技术的研究目前尚处于起步阶段。随着网络应用的快速发展，流量属性在不断变化，流量分类的新问题不断涌现。当前运用机器学习方法处理流量分类问题主要面临数据偏斜、标识瓶颈、属性变化和实时分类等 4 类亟待解决的共性问题。

4.1 数据偏斜问题

数据偏斜问题指一个数据集中不同类别的样本数量可能存在量级上的差别，分类器容易被样本数量占优的大类所淹没，而导致小类样本的分类准确率偏低。然而，在 Internet 流量中不同类型的网络流数目通常存在较大差异，因此在基于机器学习的流量分类研究中数据偏斜问题不可避免。以 Moore 等人[17]的实验结果为例，如表 3 所列，WWW 和 MAIL 等大类的分类精度普遍较高，而 P2P 和 ATTACK 等小类的分类精度则低于 30%。

表 3 样本数量与分类精度的对比

	WWW	MAIL	BULK	SERV	DB	P2P	ATTACK	MMEDIA
Flows	328091	28567	11539	2099	2648	2094	1793	1152
Precise(%)	98.29	90.56	66.66	36.73	80.88	28.92	10.38	76.91

为此，Erman[26]和 Zander[28]等人采用了欠抽样技术(undersampling)来严格控制数据集中大类的样本数量，以达到抑制数据偏斜的效果。然而这种简单修正方式改变了训练数据集的数据分布，对基于概率的机器学习算法(如朴素贝叶斯方法等)造成了负面影响。而且大类样本中仅有部分被用于模型构建，可能丢失大类样本的分布特征，从而导致分类准确率显著下降。因此，“基于机器学习的流量分类”如何有效处理数据偏斜问题还有待进一步研究。

4.2 标识瓶颈问题

标识指在构造训练数据集时预先确定样本类别。随着网络规模的日益增大，网络应用类别日益增多。为了保证分类器的准确率，需要尽可能提供大规模的标识样本，这造成了实际操作的难度，也制约着分类器的性能。当前，样本标识困难包括以下 3 个原因。

公开的网络流量记录不包含应用层负载，只能使用基于默认端口的方法进行标识。然而 Karagiannis 等人[40]指出大量的 P2P 应用使用随机端口传输数据，基于默认端口的流量识别方法难以准确标识 P2P 等新型网络应用。

包含应用层负载的网络流量记录虽然可以利用基于特征字段的流量分类方法进行标识，但这种方法计算和存储开销过大。而且，采集和解析这类流量记录受用户隐私问题的约束[11]；随着负载加密技术的普及，基于特征字段的样本标识技术的可操作性也受到质疑[41]。

基于监督 ML 的模型评估采用独立的、大规模的训练集和测试集，且预标识对应的输出类别。利用人工网络流量记录构造的数据集[42]，类别标识简单，通过网络测量获取的数据集标识困难，且数据集的规模和范围受到限制，很难代表真实的网络环境；还易引发分类器的过拟合(overfitting)或欠拟合(underfitting)问题。

Erman 等人[43]提出一种基于半监督(semi-supervised)学习的网络流量分类方法。该方法可以利用少量的已知样本和大量的未知样本作为输入来构建非监督分类器，再利用标识样本实现簇到类的映射。这种混和方法(Hybrid Approach)结合了监督和非监督学习技术，从理念上有效解决了标识瓶颈问题。但是该方法的实用性和有效性尚有待进一步考证。

4.3 属性变化问题

如何处理网络流属性的动态变化，一直是流量分类研究中的关键问题。从端口跳变(port hopping)技术的使用到负载加密技术的兴起，网络流属性的变化都引起了流量分类技术的变革。造成网络流属性发生变化的原因很多，总体上可以划分为主观因素和客观因素。主观因素来自人为修改应用代码或改变设置参数，如端口跳变技术和负载加密技术。而客观因素源自网络环境变化，以目前广泛使用的报文平均到达间隔(mean inter-arrival time)为例，当分类器所处的网络环境发生拥塞现象时，报文的平均到达间隔将发生显著变化，而这种变化可能引起分类器部分失效[44,45]。

为了解决这些问题，研究人员对网络应用行为[46,47]进行了深入的研究，越来越多的新属性被应用于流量分类的相关领域[48,49]。正如 Bernaille 等人[50]所说，流量分类研究的主要挑战是不断探讨规避(evasion)策略。网络流量分类的任何流属性都可能因人为修改而发生变化。探讨对环境和噪声不敏感而且具有自相似性(self-similarity)的网络流属性，将是一项长期且富于挑战性的工作。

4.4 实时分类问题

实时分类问题要求分类器能够快速、及时、连续且准确地对网络流进行分类。通常的基于完整流的分类器[8,13,33]不适合实时流量分类，而利用“流”中的部分字节或分组作为特征构建的分类器则有望提高分类速度。

然而，已有方法尚不能解决网络恶意攻击和流属性变化

问题，并且这类方法都建立在难于实现的假设之上，如准确地截获起始网络流；不存在流定位差错或分组丢失等问题。虽然 Nguyen 的多子流模型[15]基于滑动窗口（sliding window）方法获取数据流，无需忧虑流的初始位置，但子流属性依然赖以网络运行状况，其鲁棒性还有待考察。

网络流量分类技术落实到实时分类才能体现其应用价值，目前这方面的研究正在起步，从分类属性的选择到模型算法确定、分类器的生成及其性能指标的确定等均尚未得出结论。

结束语 本文简要描述了利用机器学习方法解决流量分类问题的意义和方法；在深入研究多篇学术论文的基础上评述了数据预处理、分类器和分类器评估，并指出了存在的问题，重点给出现阶段基于机器学习的流量分类研究中存在的数据偏斜、标识瓶颈、属性变化和实时分类 4 个共性问题。指出流量分类技术要走向实用，能否进行实时分类是“验金石”。

有监督机器学习的流量分类方法，其预标识问题与借助机器学习方法进行流量分类以摆脱传统流量分类方法弊端的初衷相矛盾，而无监督机器学习方法仅在新应用发现方面具有理论优势。我们认为机器学习方法比较灵活，但是流量分类方法的历史沿革不宜全盘否定。半监督机器学习允许数据集同时包括有标记和无标记样本，从无标记样本中挖掘更多的有用信息正是该方法的优势所在。目前，我们提出了一种自学习半监督机器学习模型，对比实验取得了比较好的效果。

参 考 文 献

[1] 刘琼，徐鹏，杨海涛，等. Peer-to-Peer 文件共享系统的测量研究[J]. 软件学报，2006，17(10)：2131-2140

[2] Moore A W, Papagiannaki K. Toward the Accurate Identification of Network Applications [C]// Proceedings of PAM 2005. Boston, 2005

[3] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification [C]// Proceedings of the MASCOTS'06. Monterey, USA, August 2006

[4] Sen S, Spatscheck O, Wang D. Accurate, scalable in network identification of P2P traffic using application signatures [C]// Proceedings of WWW 2004. New York, NY, USA, May 2004

[5] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark [C]// ACM SIGCOMM. Philadelphia, PA, USA, 2005

[6] Paxson V. Measurements and Analysis of End-to-End Internet Dynamics [D]. Berkeley, California: University of Caifornia, 1997

[7] Vigna G, Robertson W, Balzarotti D. Testing network based intrusion detection signatures using mutant exploits [C]// Proceedings ACM Conference on Computer and Communications Security. Washington, DC, 2004

[8] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining [M]. Boston: Addison Wesley, 2006

[9] Frank J. Machine learning and intrusion detection: Current and future directions [C]// Proceedings of Computer Security Conference. Washington, D. C, October 1994

[10] Williams N, Zander S, Armitage G. Evaluating machine learning methods for online game traffic identification [EB/OL]. 060410C. Centre for Advanced Internet Architectures (CAIA). http://caia.swin.edu.au/reports/060410C/CAIA-TR-060410C.pdf, Apr. 2006

[11] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification [C]// Proceedings of ACM SIGCOMM Internet Measurement Conference. Taormina, Sicily, Italy, 2004

[12] Moore A W, Zuev D, Crogan M. Discriminators for use in flow-based classification. Technical Report [R]. Queen Mary University of London, Cambridge. http://www.dcs.qmul.ac.uk/~awm/publications/moore2005discriminators.pdf, 2005

[13] Williams N, Zender S, Armitrage G. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(5): 5-15

[14] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic Classification on the Fly [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26

[15] Nguyen T T T, Armitage G. Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world IP networks [C]// Proceedings of IEEE Conference on Local Computer Networks. Florida, USA, 2006

[16] Nguyen T T T, Armitage G. Synthetic Sub-flow Pairs for Timely and Stable IP Traffic Identification [C]// Proceedings of Australian Telecommunication Networks and Application Conference. Melbourne, Australia, 2006

[17] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques [C]// Proceedings of ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York, USA, 2005

[18] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) [M]. Elsevier Inc, 2005

[19] Yu L, Liu H. Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution [C]// Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003). 2003

[20] Zander S, Nguyen T, Armitage G. Automated Traffic Classification and Application Identification Using Machine Learning [C]// Proceedings of Local Computer Networks (LCN 2005). Sydney, Australia, 2005

[21] Zuev D, Moore A W. Traffic Classification Using a Statistical Approach [C]// Proceedings of PAM 2005. Boston, USA, 2005

[22] Auld T, Moore A W, Gull S F. Bayesian Neural Networks for Internet Traffic Classification [J]. IEEE Transaction on Neural Networks, 2007, 18(1): 223-239

[23] John G H, Langley P. Estimating Continuous Distributions in Bayesian Classifiers [C]// Proceedings of Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 1995

[24] Bouckaert R. Bayesian Network Classifiers in Weka. Technical Report [EB/OL]. Department of Computer Science, Waikato University, 2005. http://www.cs.waikato.ac.nz/~remco/weka.bn.pdf, 2005

[25] Kohavi R. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid [C]// Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD). Portland, OR, 1996

[26] Nguyen T T T, Armitage G. A Survey of Techniques for Internet Traffic Classification Using Machine Learning [J]. 2008 of IEEE Communications Surveys and Tutorials, 2008, 10(4): 56-76

[27] McGregor A, Hall M, Lorier P, et al. Flow Clustering Using Machine Learning Techniques [C]// Proceedings of PAM'04. Antibes Juan-les-Pins, France, 2004

[28] Zander S, Nguyen T, Armitage G. Self-Learning IP Traffic Classification Based on Statistical Flow Characteristics [C]// Proceedings of PAM'05. Boston, USA, 2005

（下转第 66 页）

sensor networks: A survey[J]. Pervasive and Mobile Computing, 2008, 4(3): 303-334

[2] 俞靓,王志波,骆吉安,等. 面向移动目标追踪的无线传感器网络QoS指标体系设计[J]. 计算机学报, 2009, 32(3): 441-462

[3] 王雪,王晟,马俊杰. 无线传感网络移动节点位置并行微粒群优化策略[J]. 计算机学报, 2007, 30(4): 563-568

[4] 张轮,陆琰,董德存,等. 一种无线传感器网络覆盖的粒子群优化方法[J]. 同济大学学报, 2009, 37(2): 262-266

[5] 郭怡婷,王俊年. 无线传感器网络中基于微粒群算法的优化覆盖机制[J]. 计算机与现代化, 2009, 6: 1-4

[6] 林祝亮,冯远静. 基于粒子群算法的无线传感网络覆盖优化策略[J]. 计算机仿真, 2009, 26(4): 190-193

[7] Yu C S, Shin K G, Lee B. Power-stepped protocol: Enhancing spatial utilization in a clustered mobile ad hoc network[J]. IEEE Journal on Selected Areas in Communications, 2004, 22(7): 1322-1334

[8] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440-442

[9] Gupta P, Kumar P R. The capacity of wireless networks[J]. IEEE Transactions on Information Theory, 2000, 46(2): 388-404

[10] Dhillon S S, Chakrabarty K, Iyengar S S. Sensor Placement for Grid Coverage Under Imprecise Detections[C]// Proc. International Conference on Information Fusion. Annapolis: IEEE Press, 2002: 1581-1587

[11] Dhillon S S, Chakrabarty K. Sensor Placement for Effective Coverage and Surveillance in Distributed Sensor Networks[C]// Proc. IEEE Wireless Communications and Networking Conference. New Orleans: IEEE Press, 2003: 1609-1614

[12] Kennedy J, Eberhart R C. Particle Swarm Optimization [C]// Proc. IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Service Center, 1995: 1942-1948

[13] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[C]// Proceedings of the 1997 Conference on Systems, Man and Cybernetics. Piscataway, NJ: IEEE Service Center, 1997: 4104-4109

[14] Kennedy J, Eberhart R C, Shi Y H. Swarm Intelligence[M]. San Francisco, USA: Morgan Kaufmann Press, 2001

[15] 张国富,蒋建国,夏娜,等. 基于离散粒子群算法求解复杂联盟生成问题[J]. 电子学报, 2006, 35(2): 323-327

[16] 周明,孙树栋. 遗传算法原理及其应用[M]. 北京: 国防工业出版社, 2004

(上接第 40 页)

[29] Cheeseman P, Strutz J. Bayesian Classification (AutoClass): Theory and Results [M]. Advances in Knowledge Discovery and Data Mining. Cambridge, MA: AAAI/MIT Press, 1996: 153-180

[30] Erman J, Arlitt M, Mahanti A. Traffic Classification Using Clustering Algorithms [C]// Proceedings of SIGCOMM Workshop on Mining Network Data. Pisa, Italy, 2006

[31] Ester M, Kriegel H, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C]// Proceedings of Knowledge Discovery and Data Mining (KDD 96). Portland, OR, 1996

[32] Erman J, Mahanti A, Arlitt M. Internet Traffic Identification Using Machine Learning Techniques [C]// Proceedings of IEEE Global Telecommunications Conference. San Francisco, USA, 2006

[33] Bernaille L, Teixeira R, Salamatian K. Early Application Identification [C]// Proceedings of Conference on Future Networking Technologies 2006 (CoNEXT 06). Lisbon, Portugal, 2006

[34] Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Second Edition)[M]. Morgan Kaufmann Publishers, 2005

[35] Rand W. Objective criteria for the evaluation of clustering methods [J]. Journal of the American Statistical Association, 1971, 66 (336): 846-850

[36] Halkidi M, Batistakis Y, Vazirgiannis M. Cluster validity methods: part I [J]. SIGMOD Rec, 2002, 31(2): 40-45

[37] Xu R, Wunsch D. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678

[38] Halkidi M, Batistakis Y, Vazirgiannis M. Clustering validity checking methods: part II [J]. SIGMOD Rec, 2002, 31(3): 19-27

[39] Zander S, Williams N, Armitage G. Internet Archeology: Estimating Individual Application Trends in Incomplete Historic Traffic Traces [C]// Proceedings of Passive and Active Measurement Workshop (PAM 2006). Adelaide, Australia, 2006

[40] Karagiannis T, Broido A, Brownlee N, et al. Is P2P dying or just hiding [C]// Proceedings of the IEEE Globecom. Dallas, TX, 2004

[41] Erman J, Mahanti A, Arlitt M, et al. Identifying and Discriminating Between Web and Peer-to-Peer Traffic in the Network Core[C]// Proceedings of the 16th International World Wide Web Conference (WWW 2007). New York, NY, USA, 2007

[42] Wang R, Liu Y, Yang Y X, et al. Solving the App-level Classification Problem of P2P Traffic via Optimized Support Vector Machines [C]// Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06). Jinan, China, 2006

[43] Erman J, Mahanti A, Arlitt M, et al. Semi-supervised Network Traffic Classification [C]// Proceedings of ACM SIGMETRICS'07. San Diego, CA, 2007

[44] 徐鹏,刘琼,林森. 改进的对等网络流量传输层识别方法 [J]. 计算机研究与发展, 2008, 45(5): 794-802

[45] 徐鹏,刘琼,林森. 基于支持向量机的 Internet 流量分类研究 [J]. 计算机研究与发展, 2009, 46(3): 407-414

[46] Xu K, Zhang Z, Bhattacharya S. Profiling Internet Backbone Traffic: Behavior Models and Applications [C]// Proceedings of the ACM SIGCOMM 2005. Philadelphia, PA, 2005

[47] Bernaille L, Soule A, Jeannin M I, et al. Blind application recognition through behavioral classification[R]. RP-LIP6-2005-02-02. Laboratory of Computer Sciences, Paris 6, Pierre & Marie Curie University, 2005

[48] Collins M P, Reiter M K. Finding Peer-to-Peer File-sharing Using Coarse Network Behaviors[C]// Proceedings of the 11th European Symposium on Research in Computer Security (ESORICS 2006). Hamburg, Germany, 2006

[49] Bartlett G, Heidemann J, Papadopoulos C. Inherent Behaviors for On-line Detection of Peer-to-Peer file sharing [R]. ISI-TR-2006-647. University of Southern California, 2006

[50] Bernaille L, Teixeira R. Early Recognition of Encrypted Applications [C]// Proceedings of PAM'07. Louvain, Belgium, 2007