基于机器学习的网络流量分类研究发展报告

摘　要：机器学习方法不依赖匹配协议端口或解析协议内容，而是利用网络流的各种统计特征识别网络应用，近年来得到了广泛关注和快速发展. 本文总结了基于机器学习的网络流量分类方法自 2004 年来的研究进展，并且按有监督、无监督与半监督的区别进行分类、分析与比较. 重点讨论了基于机器学习的网络流量分类研究的挑战与方向，即解决样本标注瓶颈、样本分布不平衡与动态变化、实时与连续分类以及分类算法可扩展性等核心问题.

关 键 词：机器学习；网络流；网络流量分类；统计特征

Advances in Machine Learning Based Network Traffic Classification

Abstract: ML (machine learning) employs statistical network flow characteristics to assist in the IP traffic classification identification and classification, which is different with traditional methods that depend on well known application port numbers or deeply inspecting the contents of packet payloads. ML-based network traffic classification has been researched widely and developed rapidly. This survey reviews the significant works that cover the dominant period since 2004, and categorize, analyze and compare them according to their choice of ML strategies which include supervised, unsupervised and semi-supervised learning algorithms. We importantly discuss the orientations and challenges for the employment of ML-based traffic classifiers in operational IP networks. More specifically, the key issues such as sample labeling bottleneck, skewed data distribution, real-time and continuous classification and scalability of classification algorithms are discussed.

Key words: machine learning; network flow; network traffic classification; statistical characteristics

1　引　言

目前，互联网应用正向纵深方向发展，新的应用模式 (如 P2P)与应用需求不断涌现，网络流量不断增长并呈现多样化，给互联网运营与管理带来巨大压力与挑战. 实时网络流量分类对帮助互联网服务提供商了解网络运行状态、优化网络运营与管理具有重要的意义. 借助网络流量分类，网络管理者可以实时将网络中所有流量按不同应用类型进行划分与分析，为部署服务质量控制 (QoS)机制提供依据，并针对不同类型的应用提供不同的服务质量等级，从而避免减轻网络拥塞，确保关键业务服务质量，维持网络高效通畅运行. 同时，依靠流量分类，网络服务提供商可以预测网络业务的发展趋势，合理的规划网络基础体系结构，使用户得到更好的上网体验. 另外，在网络安全方面，流量分类是入侵检测系统 (intrusion detection system，IDS)的核心部分[1]，可发现网络中的突发流量 (如蠕虫传播、大规模分布式拒绝服务攻击等)与未知协议流量，从而及时采取防御遏制措施.

传统的流量分类主要基于端口与基于分组深度解析两种方法. 使用端口进行流量分类[2]是通过检查分组的传输层端口号，然后根据 IANA 定制的知名端口号与注册端口号列表将分组与应用对应起来. 然而，随着网络应用的不断更新发展，使用端口号进行流量分类的缺陷日益明显. 已有研究表明[3,4]，流行的 P2P 与被动 FTP 等新型网络应用普遍利用随机端口进行数据传输，进而导致基于端口的流量分类方法已不再适用.

基于端口的流量分类具有众多不可靠因素，因此在工业级产品中广泛采用基于数据包深度解析的方法[5]，主要依靠分析数据包的有效负载来判断其是否包含与已知应用匹配的特征. 此方法需以 2 个假设作为应用前提：除数据包源与接收者外的第三方能提取每个 IP 分组载荷明文；需要分类的每种已知应用的语法与特征. 在此 2 个假设成立的情况下，基于数据包深度解析的流量分类具有准确率高的特点. 然而，随着应

用负载加密与新型应用的不断涌现，无法获取数据包负载明文以及未知应用语法与特征则导致此方法的有效性逐步下降. 另外，此方法需要丰富的存储资源与计算能力，难适应高带宽网络流量实时在线分类的应用.

为了克服上述两种方法的不足，近年来许多研究者开始利用机器学习方法解决流量分类问题. 机器学习方法不依赖匹配协议端口或解析协议内容识别网络应用，而是利用流量在传输过程中表现出来的"网络流"(flow)的各种统计特征区别网络应用，方法本身不受动态端口、载荷加密甚至网络地址转换的影响. 本文将对近年来基于机器学习的网络流量分类方法进行分类综述，比较指出常见方法的优势与不足并讨论未来利用机器学习进行网络流量分类研究的挑战与方向.

2 基于机器学习的网络流量分类

2.1 基于流统计特征的机器学习流量分类

机器学习在众多领域已有应用，包括搜索引擎、医疗诊断、文本与手写识别、负载预测、市场与销售数据分析等等[6,7]. 在本文中，网络流(flow)按照五元组进行定义：源-目IP地址、源-目端口和协议. 流量分类问题可以抽象为以下步骤：

第1步. 若已知网络流类型集合 $C = \{C_1, C_2, \cdots, C_k\}$ 与网络流集合 $T = \{t_1, t_2, \cdots, t_n\}$，其中网络流 t_i 可以表示一个为由各种统计特征构成的特征向量 $A_i = (A_{i1}, A_{i2}, \cdots, A_{im})$；

第2步. 利用已知类型的网络流集合与其特征来学习训练分类模型 $f: T \rightarrow C$，并利用训练建立的分类模型对未知类型的网络流进行分类.

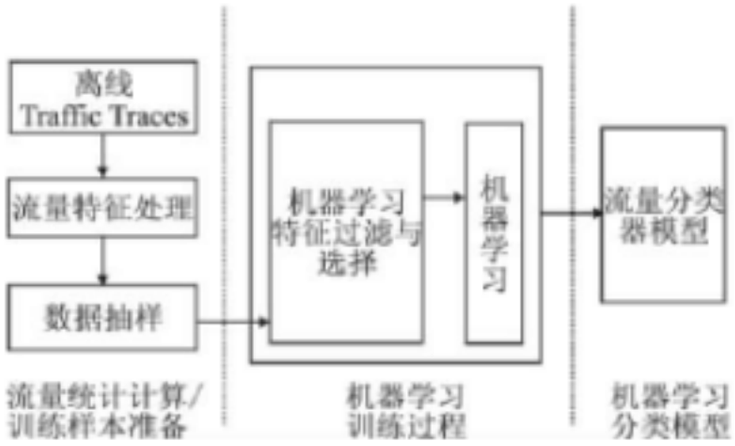

图1 有监督机器学习流量分类训练过程

Fig. 1 Training process of supervised ML based IP traffic classification

基于有监督机器学习流量分类的一般过程如图1、图2所示. 在图1中，首先利用已标注应用类型的网络流作为训练样本集，并提取网络流统计特征(如流持续时间、分组数量、最大分组长度、分组到达时间间隔等等)，同时，如果训练样本数量庞大，通常需要进行数据抽样处理以降低训练复杂度；其次，在训练过程中，通常先利用特征过滤或特征选择算法求出对分类识别最有效的特征组合，有利于减少算法学习时所需要的数据量，减少执行时间和提高分类正确性. 然而，特征选择算法往往会导致特征选择的局部最优性，进而使得分类结果不稳定. 在图2中，利用分类器对网络实时流量进行分类，根据分类器需求提取网络流相应的统计特征作为分类器输入，分类器经过预测计算并给出分类结果. 另外，利用实时采集的网络流量，可实时更新分类器以达到更好的分类性能.

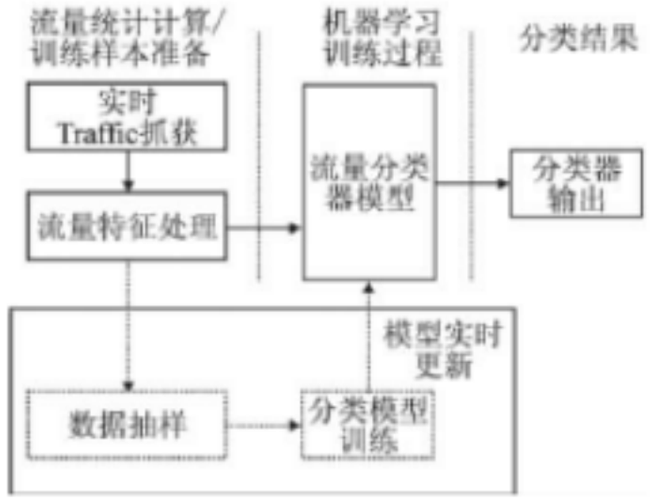

图2 机器学习流量分类器在线分类应用

Fig. 2 Data flow within an operational supervised ML traffic classifier

2.2 特征选择算法

选取能代表网络应用本质区别的流特征，对于流量应用分类非常重要. 在数据挖掘与机器学习应用中，特征选择通常用于对数据进行预处理，包括降维、去冗余、过滤无关特征、提高学习精度等等. 特征选择算法可分为过滤(filter)与封装(wrapper)两种方法. 其中，Filter特征选择算法的评价函数与分类器无关，尽管具有通用性强、算法复杂度低的特点，但对某一个具体的分类器选择的特征子集也许并不是最优的. 这一类特征选择算法较多，如基于相关性的子集搜索方法CFS[8]、基于一致性的子集搜索方法CON[9]、FCBF[10]等.

Wrapper[11]方法与其相反，采用分类器的错误概率作为评价函数，因此对特定的分类器可以找到最优的特征子集，但算法复杂度很高，此类方法的代表算法有基于遗传算法的wrapper方法[12]等. 另外，目前无监督的特征选择算法还比较少，在样本类别未知的情况下，需要选用无监督的特征选择算法，如Dash等特出的一种基于熵的Filter模型[13].

2.3 机器学习分类方法性能评估策略

针对某一机器学习分类模型，模型评估是指评价分类模型在未知样本集上处理分类问题的能力，其关键指标是对未知样本的预测准确率. 若网络流量中包含n条网络流样本，分别属于m种网络应用类型，那么对类型 i，其测试结果的邻接表如表1所示，其中：

表1 类别 i 测试结果邻接表

Table1 The contingency table for category i

Category i		Expert judgments	
		True	False
Classifier judgments	Positive	TP	FP
	Negative	FN	TN

TP(true positive)：类型 i 中的样本被分类模型正确预测的样本数，记为 TP_i；

FN(false negative)：类型 i 中的样本被分类模型预测为其它类型的样本数，记为 FN_i；

FP(false positive)：不属于类型 i 的样本被分类模型预测

为类型 i 的样本数，记为 FP_i.

基于以上概念，下面给出评价分类模型准确性的 3 个常用指标：类准确率 (recall)、类可信度 (precision)以及整体准确率 (overall recall)的描述，计算方法如公式 (1)~(3)所示：

$$recall(i) = \frac{TP_i}{TP_i + FN_i} \tag{1}$$

$$precision(i) = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

$$overall\_recall = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + FN_i} \tag{3}$$

在这 3 个评价指标中，分类模型的整体准确率应用最广，它反映了分类模型正确预测样本数占总样本数的比例. 类 i 的准确率表示类 i 所有样本中被分类模型正确预测的样本的比例，类 i 的可信度表示被预测为类 i 的样本中真实类 i 样本的比例. 类准确率、类可信度反映了分类模型对单个应用类型的预测能力. 整体准确率反映了分类模型对多种应用类型的综合预测能力.

3 基于机器学习的网络流量分类方法分析与比较

3.1 基于有监督机器学习的流量分类

有监督机器学习分类是基于已标注类型的样本集进行机器学习并建立分类规则，将未知样本分类为已知的类型. 有监督机器学习方法一般检测率高，但要求样本数据事先正确标记类别，无法对未知应用类型的流量进行分类；为生成具有良好的泛化性能的检测模型，往往需要利用大规模标注过的训练数据提高学习算法结果的准确度，但是标记必须由人手工完成.

3.1.1 最近邻与线性判别式分析

Huang 等人[14] 提出基于 K-最近邻 (K-nearest neighbor, KNN)分类器的流量分类方法. 该方法针对 6 种主要应用流量进行分类实验，达到了 90% 以上的分类准确率. 然而，K 近邻方法是基于实例的学习方法，在对测试集样本分类时，需要逐个计算测试样本与训练样本之间的相似度，因而此方法通常会导致较大的处理开销. 同时，该方法所使用的样本集与测试集过小，不足以证实该方法在实际使用时的有效性.

Roughan 等人[15] 提出使用 K-最近邻机器学习方法进行网络流量应用分类. 该方法共使用了 5 类特征：分组层次、流层次、连接层次、流与连接内部特征、同一源目主机间的多条并发流的特征，同时使用十折交叉认证评价分类方法. 然而，实验结果说明该方法随着流量应用类型数量的增加，分类错误率明显上升.

3.1.2 贝叶斯算法

Moore 与 Zuev 等人[16] 引入有监督的朴素贝叶斯 (Naïve bayes, NB)机器学习方法进行流量分类与应用识别. 但该方法要求样本各个特征满足条件独立并遵循高斯分布，然而实际应用中的原始网络流量特征很难满足上述条件，因此其分类准确率只有约 65%. 为解决此问题，Moore 等人采用特征选择方法对特征集合进行过滤，并使用核密度估计对朴素贝叶斯方法进行了改进，分类准确率得到提高达到 95% 以上. Moore 等人[17] 进一步提出将贝叶斯神经网络应用于流量分类，此方法与前 2 种贝叶斯方法相比具有更高的分类准确率，训练的分类器对同一天内采集的流量数据达到了近 99% 的分类准确率，对相隔八个月的流量数据也达到约 95% 的分类准确率.

应用朴素贝叶斯及其改进算法进行流量分类具有实现简单、分类高效的特点，但是朴素贝叶斯是一种传统的参数估计方法，依赖训练集样本先验概率分布，然而实际获取的未知流量集样本分布往往与训练集不同，因此朴素贝叶斯方法具有潜在的分类不稳定性.

3.1.3 决策树

王宇等人[18] 提出基于 C4.5 决策树分类器的有监督网络流量分类方法，讨论特征选择和 boosting 增强方法 2 种改进策略. 文中实验结果表明，C4.5 分类器的训练复杂度适中，准确率高且分类速度快. 徐鹏等人[19] 引入 C4.5 决策树方法来处理流量分类问题. 该方法利用训练数据集中的信息熵来构建分类模型，并通过对分类模型的简单查找来完成未知网络流样本的分类. 该文经过实验说明 C4.5 决策树方法具有以下优点：与 NB 方法不同，C4.5 决策树不依赖于网络流样本分布的先验概率，因此在网络流样本分布变化时依然具有较好的分类准确率；具有较快的流量分类速度，在对待测网络流样本进行分类时，仅需进行特征值比较，计算量小，在处理大规模流量分类问题时具有明显的性能优势.

Li 等人[20] 分别利用 C4.5 和 AdaBoost + C4.5 算法构建了决策树，该方法利用关联的过滤方法筛选出 12 个最优特征，并只处理每个 TCP 流的前 5-6 个分组以提高分类的实时性. 经过不同采集时间的测试集验证，决策树方法对 P2P 流的预测准确率达到 99% 以上. Raahemi 等人[21] 提出基于 CVFDF 决策树算法的 P2P 流量识别方法，该算法能够得知每个决策节点测试所需要的最少样例数，并能使决策树动态生长，从而满足识别模型动态更新的需要.

以上研究表明，C4.5 决策树方法在处理流量分类问题时具有较好的分类稳定性与准确率，但决策树方法根据训练数据集的局部信息对样本特征进行过滤，会导致特征选择的局部最优性，进而导致分类结果的不稳定，同时在高维样本学习时存在复杂度过高的问题.

3.1.4 支持向量机

徐鹏等人[22] 提出一种基于支持向量机 (support vector machine, SVM)的流量分类方法. 该方法利用非线性变换和结构风险最小化 (structural risk minimization, SRM)原则将流量分类问题转化为二次寻优问题，具有良好的分类准确率和稳定性. 该文通过在实际网络流集合上与朴素贝叶斯算法的对比实验说明 SVM 方法具有几种优势：网络流特征不必满足条件独立假设，无须进行特征过滤；能够在先验知识相对不足的情况下，仍保持较高的分类准确率；不依赖于样本空间的分布，具有较好的分类稳定性. 然而，该方法使用了 Moore 等人提出的 247 个流特征，流特征数量过多将导致实际应用时过高的特征计算负载.

Rui W 等人[23]使用 V-SVM 作为二值支持向量机并用于 P2P 流量识别. 基于网络连接数相关的统计特征,该方法将网络流划分为 P2P 流与 non-P2P 流两类,然而,网络连接数相关的统计特征依赖于应用的连接模式,分类结果的稳定性容易受到网络环境的影响.

应用 SVM 方法进行流量分类具有较好的分类稳定性与准确率,并且 SVM 方法在小样本训练空间时依然可保持较好的分类性能. 然而,由于网络流量分类应用中通常是大样本训练集,而 SVM 方法在训练集规模较大时具有速度慢的缺陷,因此难以满足网络流量分类的实际需求,如何改进 SVM 方法将其应用于大规模流量样本训练分类是个有意义的研究问题.

3.1.5 神经网络

Shen 等人[24]提出基于 BP 神经网络的 P2P 流识别方法,该方法采用基本的三层 BP 神经网络层次结构,将几种常见流特征作为输入,输出为 P2P 与 non-P2p 两种,并基于经验决定隐层节点数目取 4. 通过实验表明,该方法识别正确率为 96.3%,但 BP 神经网络学习学习速率固定,网络的收敛速度慢,需要较长的训练时间. Sun 等人[25]提出利用概率神经网络 (probabilistic neural network, PNN)解决流量分类问题,文献中将 PNN 方法与支持向量机、RBF 神经网络进行比较,PNN 方法取得最高的分类准确率. 相对于 BP 神经网络,PNN 具有训练速度快,收敛性好等特点,但此方法仅仅限于 web 与 P2P 两类流量分类,流类型数量有限,其用于实际流量分类的能力有待进一步考验. Raahemi 等人[26]将 fuzzy ART-MAP 神经网络应用于 P2P 流的识别领域,与 BP 神经网络相比,该方法具有增量学习能力,能够在变化中学习新信息,且不会造成系统不稳定,破坏已学习知识信息. 实验表明,该方法识别率高于 80%,而非增量学习方法的识别率只有 78%. 目前,将神经网络用于流量分类的研究还局限在小规模流量数据与较少流量类型的应用,尚无研究对其分类准确率与稳定性进行深入的分析.

3.1.6 实时流量分类

Nguyen 等人[27]将网络流按协议的不同阶段 (建立连接、数据传输、结束连接)划分为不同子流,统计分析每条子流的特征向量并构造流量分类模型. 此方法将流的特性按阶段进行划分统计,在对网络流进行分类处理时,可不需要首个数据分组或等待网络流结束,提高了流量分类的实时性与实用性. 然而,作者只是验证了此方法对某种在线游戏应用流量识别的有效性,其应用于多种类型网络流量分类的有效性有待进一步验证. Nguyen 等人[28]进一步研究如何自动提取合适的子流用于训练分类模型,文中方法采用 EM 聚类方法对子流进行聚类,并提取最能代表应用流特征的子流构造训练集,此方法能进一步降低分类模型训练的计算复杂度.

Huang 等人[29]利用应用协议在早期协商阶段的行为特征对其进行识别,该研究发现,不同的应用协议在具有独特的协商阶段,文中针对协议协商阶段提取对应统计特征用于机器学习分类,通过多种机器学习方法 (Naïve Bayes, SMO, Bayesian network, partial decision tree, C4.5)的对比实验证明,该方法能有效对应用协议早前阶段进行识别,其中 partial decision tree 方法达到 97.24% 的分类准确率.

3.1.7 几种不同有监督学习方法的比较

Wang 等人[30]研究了机器学习方法在流量分类领域的应用,该研究采集 Sun Yat-Sen 大学校园网流量并选取其中 7 种网络应用作为实验数据集,包括 FTP、MSN、PPLive、QQ、Web、Thunder、QQGame. 文献共采用了 Naïve Bayes (NB)、Decision Tree、Bayesian Neutral Network (BNN)、Naïve Bayes Tree (NBT)与 AdaBoost 五种机器学习方法进行实验,结果表明 5 种方法的分类准确率比较接近,都在 90% 左右,即机器学习方法能有效地应用于流量分类.

Williams 等人[31]研究比较了几种常见机器学习流量分类方法,包括离散化贝叶斯 (Naïve Bayes with discretisation, NBD)、核密度估计贝叶斯 (Naïve Bayes with kernel density estimation, NBK)、C4.5 决策树、贝叶斯网络 (Bayesian network)、贝叶斯树 (Naïve Bayes tree, NBT)5 种算法. 文中利用公开共享的 NLANR 流量文件以及预选的 22 个流特征,对 5 种算法的训练时间、分类时间、分类准确率进行了比较. 作者首先对比各种方法基于原始特征集与经过特征选择后的分类性能,结果表明几种方法整体分类准确率接近,并且经过特征选择处理后,分类性能保持稳定. 该研究还发现 C4.5 决策树拥有最快的分类速度,然后依次是 NBD、贝叶斯网络、NBT、NBK. 在分类模型建立阶段,贝叶斯树需要最长的训练时间,耗费时间按降序排列依次是 NBT、C4.5、贝叶斯网络、NBD、NBK. 同时,该文实验结果表明特征选择有助于减少各算法模型的训练时间与提高分类速度.

3.2 基于无监督机器学习的流量分类

无监督机器学习方法根据流量统计特征的相似性进行聚合分簇,然后建立各个簇与类的映射关系. 无监督机器学习具有能够自动发现新应用的特点,但其检测精度与分类速度明显低于有监督的分类方法.

3.2.1 EM 算法

McGregor 等人[32]率先将期望最大化 (expectation maximaization, EM)算法应用于网络流量聚类,该方法针的样本集包含 HTTP、FTP、SMTP、IMAP、NTP 与 DNS 六种流量,利用 EM 算法可将具有相同特性 (大文件传送、多交互等)的流聚到同一簇,然而,该方法需要进一步研究如何建立聚类簇与流应用类型之间的映射关系.

Erman 等人[33]引入 EM 聚类方法来处理流量分类问题,通过与 Bayes 的分类方法进行比较,获得了更为准确的分类结果. 此类方法无须已标记类型的训练样本,因此具有发现新型网络应用的能力,但此类方法通常需要手工标记各个聚类的应用类型,而且大规模样本聚类时间通常较长.

Zander 等人[34]提出基于 autoclass 方法对网络流量进行无监督学习分类,该方法使用 EM (Expectation Maximaization)方法从训练集中得到最佳的聚类簇,并以此训练构建分类器. 同时,作者使用从不同的网络位置收集的流量来验证该方法的有效性,获得的分类平均准确率为 86.5%. 然而,由于聚类簇数量较多并影响分类性能,因此该方法需要进一步研究如何缩减聚

类簇数量并基于聚类簇建立有效的流量分类规则.

3.2.2 K-Means 算法

Bernaille 等人[35] 提出了采用 TCP 连接的前五个数据分组的大小来代表不同的网络流,此方法可以尽可能早地识别出流的应用类型,有效的提高了分类实时性. 该方法包含离线学习与在线分类两个阶段,离线学习阶段采用 K-Means 方法对训练集流量进行聚类划分,并给出每个聚类簇的描述以及包含应用类型的组成;在分类阶段,计算每个未知流与各个聚类簇中心的欧式距离来确定其所属的簇以及应用类型. 然而,该方法采用的网络流特征依赖于分组的到达顺序,而在实际网络环境中,路由动态性、网络拥塞往往会影响到分组到达顺序,因此,Bernaille 等人的方法无法保证分类的稳定性与实用性.

王宇等人[36] 利用 K-Means 聚类方法检测未知流并描述建立未知流的应用特征. 该方法首先利用 X-Means 算法将具有相似统计特征的流聚为簇,相对于 K-Means 算法,X-Means 算法可自主决定聚类簇数量. 随后,将每个聚类簇应用类型标注为簇内占大多数的流的应用类型,基于聚类簇建立分类模型,并只利用流的前 32 字节的统计特征作为训练输入,采用 Naïve Nayes 与 C4.5 算法进行实验,分别达到 93.96% 与 92.72% 的分类准确率. 作者需要在挖掘聚类簇中未知流应用特征方面做进一步研究.

Hirvonen 等人[37] 提出一种两阶段的网络流量分类方法,提取网络流初始建立阶段与平稳阶段的特征来描述网络流,并采用无监督的 K-means 聚类方法分两个阶段训练分类器. 然而,该方法提取的第一阶段的特征依赖于连接的前 4 个数据分组的到达顺序,这将影响此方法的分类准确率.

3.2.3 DBSCAN 算法

Yang 等人[38] 提出基于 DBSCAN 聚类算法的无监督流量分类方法,DBSCAN 算法具有 3 个优点:只需要少量领域知识即能确定输入参数;能形成任意形状的聚类簇;适用于大规模数据集. 实验结果表明,DBSCAN 算法最高可达到约 87% 的分类准确率,但其分类准确率受到聚类结果的影响,其中聚类结果由输入参数决定. 然而文献中输入参数通过实验进行选择,这将影响模型的通用性.

3.2.4 不同聚类算法的比较

Erman 等人[39] 较系统地比较分析了 3 种聚类算法 (K-Means, DBSCAN 与 AutoClass)的流量分类性能,利用 2 个数据集进行实验:Auckland 大学公开共享流量文件与 Calgary 大学校园网采集的流量 (前者包含 9 种应用,后者包含 4 种应用). 该文献利用整体分类准确率与聚类簇的数量来评价各算法的有效性,分类准确率为被正确分类的样本数量占所有待分类样本数量的比率,聚类簇的数量是影响算法分类性能的重要因素. 实验结果表明,AutoClass 算法具有最高的分类准确率,但其聚类簇的数量却最多,这在分类阶段将带来较高的计算代价;K-Means 的分类准确率随着聚类簇的数量增加稳定上升,在 K 值为到达某个阈值后,分类准确率上升不明显;DBSCAN 算法的分类准确率最低,但其聚类簇的数量最少. 可以看出,聚类方法能有效的处理流量分类问题,选择合适的聚类算法以及确定适中的聚类簇数量对模型准确率与计算复杂度有重要影响.

3.3 基于半监督机器学习的流量分类

半监督机器学习主要关注当训练数据的部分信息缺失 (如数据的类别标签缺失、部分特征维缺失等)的情况下,如何获得具有良好性能和泛化能力的分类器,利用半监督机器学习解决流量分类问题是近年来的研究新热点.

柳斌等人[40] 将半监督机器学习方法用于解决流量分类问题,利用少量的标记数据辅助 K-Means 聚类过程,确定簇与流量类型的映射关系. 该文献中提出一种基于熵函数的组合式特征选择方法,首先计算所有特征的熵,前 d 个最优特征构成候选的特征子集,其次采用顺序后退搜索方法,已分类器本身的分类准确率为评估标准,藉此去除冗余特征. 经过与有监督学习 Naïve Bayes (NB)算法、无监督 K-Means 算法的对比实验,该半监督机器学习方法分类准确率高于 K-Means 算法,低于 NB 算法. Erman 等人[41] 与 Shrivastav 等人[42] 同样提出基于半监督机器学习的流量分类方法,在聚类阶段中均采用了 K-Means 方法,研究实验结果说明,半监督机器学习方法能有效的解决流量分类问题,尽管其分类准确率一般低于有监督学习方法,但却无需人工预先对所有训练样本的应用类型进行标注.

4 主要挑战与方向

基于机器学习的流量分类技术经过近年来的不断发展,特别是直接从机器学习等领域借鉴最新的研究成果,已能初步解决大部分数据量相对较小、标注比较完整、离线等特点的流量分类问题和应用. 但是,基于机器学习的流量分类技术的实际应用仍受到很多问题的困扰,其主要挑战包括:

4.1 样本标注瓶颈

有监督机器学习算法分类准确率高,且需要大量的标注样本,然而实际训练集样本数量一般较小,已标注的样本所能提供的信息有限;另外,已标注样本相对于大量实际网络流样本数量较少,其样本空间的数据分布与实际网络流样本空间有差异. 提供尽可能多的标注样本需要艰苦而缓慢的手工劳动,大幅增加了建立分类器的代价,这就产生了标注瓶颈的问题. 因此,如何利用少量的已标注样本和大量的未标注样本训练一个好分类器将成为未来网络流量分类研究方向之一. 基于这一思想,使用半监督机器学习的网络流量分类方法逐渐成为研究热点.

4.2 样本分布不均衡

通过对实际网络流量的类别组成分析,发现网络流样本集的类别分布往往是不均衡的,即类别间的样本数量存在数量级的差距,这是导致分类效果不理想的一个重要因素. 比如,现有实际网络流量中,web、mail、p2p 三类网络应用流已占绝大部分. 在样本分布不均衡的情况下,分类器容易被大类淹没而忽略小类. 另一方面,在实际的流量分类问题中,建立分类器时所获得训练样本数量相对于海量的未知数据则显得非常有限,而且实际网络流样本空间分布存在动态变化的情况,从而难以模拟未知样本的空间分布,无法保证分类结果的稳定性. 徐鹏等人[22] 对 SVM、NB 方法在样本分布受控情况

下的健壮性的对比研究，结果表明，SVM 方法对样本分布的健壮性要好于 NB 方法，这也印证了 SVM 方法的泛化性以及 NB 方法对样本空间分布先验概率的依赖性. 采用合适的分类算法与改进优化策略来解决样本分布不平衡问题是流量分类的难点问题. 同时，目前所有方法在稀有类别上的准确率均较低，相关的研究仍需进一步深入.

4.3 实时与连续分类

现有大多数研究工作都是利用完整流的统计特征，因此分类器只有在接收到完整的 IP 流后才能进行预测分类，无法满足实时要求. Huang[29] 与 Bernaille[35] 等人研究利用流的头几个分组的特征建立分类器，但却无法应对流首个分组丢失的情况. Nguyen 等人[27-28] 利用滑动窗口将流按阶段划分、统计特征并建立分类模型，摆脱了对网络流进行处理需要首个数据分组或等待网络流结束的限制，提高了流量分类的实时性与实用性. 未来的研究应更多的关注机器学习流量分类方法的在线实时应用的需求.

4.4 分类算法可扩展性

面对互联网海量的各类网络应用流，大规模的流量分类已经成为紧迫的需求. 同时，为了获得更好的分类性能，还通常利用提取高维统计特征、构造大规模样本集来构建分类器. 大规模流量分类将面对庞大的类别数量和训练样本数，从而带来问题：分类器的训练构建时间以及存储随类别、样本数量以及特征维度的增长关系；分类算法是否在较大规模样本下保持有效. 比如，SVM 方法尽管具有较好的分类稳定性与泛化性能[22]，但其训练过程所需的时间和空间复杂度分别为 O(m^3)和 O(m^2)，m 为支持向量数量，与样本数量成正比，因此不适用于对大规模数据集进行训练. 因此，需要对现有机器学习方法进行改进以适应大规模网络流量分类.

5 结束语

本文总结了基于机器学习的网络流量分类方法近年来的研究进展，并且按有监督、无监督与半监督的区别进行分类、分析与比较，重点讨论了近期所面临的一些实际应用需求、数据特点的问题及最新成果，并对将来的一些研究工作进行了展望.

网络流量分类技术有着广泛的应用需求，利用机器学习解决流量分类问题已具有一定的实用性. 随着相关应用的发展及需求的不断提升，仍有很多值得研究的问题，比如：解决大规模流量分类问题的途径和方法；可靠、有效及快速的在线分类；缓解样本标注瓶颈以及样本数据分布带来的影响等. 随着机器学习和数据挖掘领域理论和技术研究的深入，针对不同实际应用和数据的特征改进与优化学习算法，这将成为网络流量分类相关研究和应用的重点.

References:

[1] Snort [EB/OL]. http://www.snort.org,2008.

[2] Internet assigned numbers authority [EB/OL]. http://www.iana.org,2008.

[3] Karagiannis T, Broido A, Brownlee N, et al. Is P2P dying or just hiding [C]. Global Telecommunications Conference, 2004, 3: 1532-1538.

[4] Madhukar A, Williamson C. A longitudinal study of P2P traffic classification [C]. In: Proc. of the 14th IEEE Int'l Symp. on Modeling, Analysis, and Simulation. Monterey, 2006.

[5] Moore A W, Papagiannaki K. Toward the accurate identification of network applications [A]. In: Dovrolis C, ed. Proc. of the PAM 2005. LNCS 3431 [C], Heidelberg: Springer-Verlag, 2005: 41-54.

[6] Fisher H D, Pazzani J M, Langley P. Concept formation: knowledge and experience in unsupervised learning [M]. Morgan Kaufmann, 1991.

[7] Witten I, Frank E. Data mining: practical machine learning tools and techniques with Java implementations (second edition) [M]. Morgan Kaufmann Publishers, 2005.

[8] Hall M. Correlation-based feature selection from machine learning [D]. New Zealand: Department of Computer Science, Waikato University, 1998.

[9] Dash M, Liu H. Consistency-based search in feature selection [J]. Artificial Intelligence, 2003, 151 (1-2): 155-176.

[10] Yu Lei, Liu Huan. Feature selection for high-dimensional data: a fast correlation-based filter solution [C]. Proceedings of the 12th International Conference on Machine Learning, 2003: 856-863.

[11] Kohavi R, John G H. Wrappers for feature subset selection [J]. Artificial Intelligent, 1997, 97 (1-2): 273-324.

[12] Park J, Tyan H R, K C-C J. GA-based Internet traffic classification technique for QoS provisioning [C]. In Proc. 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Pasadena, California, December, 2006.

[13] Dash M, Choi K, Scheuermann P, et al. Feature selection for clustering-a filter solution [C]. In: Proc. of the Second International Conference on Data Mining, 2002: 115-122.

[14] Huang Shi-jun, Chen Kai, Liu Chao, et al. A statistical-feature-based approach to Internet traffic classification using machine learning [C]. International Conference on Ultra Modern Telecommunications & Workshops, 2009: 1-6.

[15] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification [C]. In Proc. ACM/SIGCOMM Internet Measurement Conference (IMC) 2004, Taormina, Sicily, Italy, October, 2004.

[16] Moore A, Zuev D. Internet traffic classification using Bayesian analysis techniques [C]. In ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005, Banff, Alberta, Canada, June, 2005.

[17] Auld T, Moore A W, Gull S F. Bayesian neural networks for Internet traffic classification [J]. IEEE Trans. Neural Networks, January, 2007, 18 (1): 223-239.

[18] Wang Yu, Yu Shun-zheng. Internet traffic classification based on decision tree [J]. Journal of Chinese Computer Systems, 2009, 30 (11): 2150-2156.

[19] Xu Peng, Lin Sen. Internet traffic classification using C4.5 decision tree [J]. Journal of Software, 2009, 20 (10): 2692-2704.

[20] Li Wei, Moore A W. A machine learning approach for efficient traffic classification [C]. Proc of the 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2007:310-317.
[21] Raahemi B, Cai Zhong-wei, Liu Jing. Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree [C]. Proc of the 20th IEEE International Conference on Tools with Artificial Intelligence, 2008:525-532.
[22] Xu Peng, Liu Qiong, Lin Sen. Internet traffic classification using support vector machine [J]. Journal of Computer Research and Development, 2009, 46 (3):407-414.
[23] Rui W, Yang L, Yuexiang Y, et al. Solving the app-level classification problem of P2P traffic via optimized support vector machines [J]. ISDA'06, Oct. 2006:534-539.
[24] Runyuan Sun, Bo Yang, Lizhi Peng, et al. Traffic classification using probabilistic neural networks [C]. Proc of the Sixth International Conference on Natural Computation, 2010:1914-1919.
[25] Shen Fu-ke, Pan Chan-ge, Ren Xiao-li. Research of P2P traffic identification based on BP neural network [C]. Proc of the 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2007:26-28.
[26] Raahemi B, Kouznetsov A, Hayajneh A, et al. Classification of peer-to-peer traffic using incremental neural networks (fuzzy ARTMAP) [C]. Proc of Conference on Electrical and Computer Engineering, 2008.
[27] Nguyen T, Armitage G. Training on multiple sub-flows to optimize the use of machine learning classifiers in real-world IP networks [C]. In Proc. IEEE 31st Conference on Local Computer Networks, Tampa, Florida, USA, November, 2006.
[28] Nguyen T, Armitage G. Clustering to assist supervised machine learning for real-time IP traffic classification [C]. In Proc. IEEE Communications Society Subject Matter Experts, 2008:5857-5862.
[29] Nen-Fu Huang, Gin-Yuan Jai, Han-Chieh Chao. Early identifying application traffic with application characteristics [C]. In Proc. IEEE Communications Society Subject Matter Experts, 2008: 5788-5792.
[30] Wang Jian-min, Qian Cheng-lu, Che Chun-hui, et al. Study on process of network traffic classification using machine learning [C]. The Fifth Annual ChinaGrid Conference, 2010:262-266.
[31] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [C]. Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review, 2006, 36 (5):5-16.
[32] McGregor M Hall, Lorier P, Brunskill J. Flow clustering using machine learning techniques [C]. In Proc. of Passive and Active Measurement Workshop (PAM2004), Antibes Juan-les-Pins, France, April, 2004.
[33] Erman J, Mahanti A, Arlitt M. Internet traffic identification using machine learning techniques [C]. In Proc. of 49th IEEE Global Telecommunications Conference (GLOBECOM 2006), San Francisco, USA, December 2006.
[34] Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning [C]. In IEEE 30th Conference on Local Computer Networks (LCN 2005), Sydney, Australia, November, 2005.
[35] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly [C]. ACM Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review, 2006, 36 (2):23-26.
[36] Wang Yu, Xiang Yang, Yu Shun-zheng. Automatic application signature construction from unknown traffic [C]. 24th IEEE International Conference on Advanced Information Networking and Applications, 2010:1115-1120.
[37] Hirvonen M, Laulajainen J P. Two-phased network traffic classification method for quality of service management [C]. Proc. of the 13th IEEE International Symposium on Consumer Electronics (ISCE2009), 2009.
[38] Yang Cai-hong, Wang Fei, Huang Ben-xiong. Internet traffic classification using DBSCAN [C]. In Proc. of WASE International Conference on Information Engineering, 2009:163-166.
[39] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms [A]. In MineNet '06: Proc. 2006 SIGCOMM Workshop on Mining Network Data [C]. New York, NY, USA: ACM Press, 2006:281-286.
[40] Liu Bin, Li Zhi-tang, Tu Hao. Network application classification method based on semi-supervised learning [J]. Microelectronics & Computer, 2008, 25 (10):113-116.
[41] Erman J, Mahanti A, Arlitt M, et al. Semi-supervised network traffic classification [C]. ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) Performance Evaluation Review, 2007, 35 (1):369-370.
[42] Amita Shrivastav, Aruna Tiwari. Network traffic classification using semi-supervised approach [C]. In Proc. of Second International Conference on Machine Learning and Computing, 2010: 345-349.

附中文参考文献:

[18] 王 宇, 余顺争. 网络流量的决策树分类[J]. 小型微型计算机系统, 2009, 30 (11):2150-2156.
[19] 徐 鹏, 林 森. 基于C4.5决策树的流量分类方法[J]. 软件学报, 2009, 20 (10):2692-2704.
[22] 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.
[40] 柳 斌, 李之棠, 涂 浩. 一基于半监督学习的应用层流量分类方法[J]. 微电子学与计算机, 2008, 25 (10):113-116.