第四周

网络流量分类和特征提取研究内容小结

Wu You

2018 年 8 月 15 日

目录

1. 课题背景介绍

1.1. 当前网络环境分析

由于互联网技术发展迅速，同时也为国家发展带来了巨大的经济效益，我国的互联网规模在一段时期内还将不断扩大。而随着互联网的不断普及，用户人数不断增加，同时产生了多种类的新型网络应用，网络已经覆盖了生产生活的各个方面，社会发展对网络的依赖性也逐渐增强。而网络规模的告诉发展也带来一定问题，首先一些使用新型的网络应用如基于 P2P 或 VoIP 的软件对于网络的占用率较高，影响其他网络应用的运行; 其次随着电子商务的兴起，网络中信息价值不断提高，导致网络安全问题日益突出；最后，互联网信息扩散十分迅速，为不良信息的传播提供了条件。[12]可以看出，加强对网络流量的监督和管理是十分重要的。

当前网络环境有流量大、种类多、发展快的特点，无论是网络的监督管理，还是安全防护，在全流量的基础上进行无疑是事倍功半，迫切需要对流量进行分类以进行有针对性的工作，因此网络流量的分类技术需要不断升级以应对日益复杂的网络环境。流量分类识别是进行各项网络分析工作的第一步，同时，在该方向上的研究成果也可以部分迁移至后续的分析步骤中，故流量分类技术的研究对于当前社会具有重大意义。

1.2. 流量分类的主要工作内容

有关网络流量分类的研究可以分为流量数据采集和预处理、特征的提取、算法的选择与改进以及分类系统的部署等问题。其中，行业内进行数据采集的方法已经较为完善，并且有多种数据集可用于研究，但对于其后的特征提取问题，目前还不存在十分完美的解决方案。作为流量分析的重要环节，流量特征选取对于分类性能有着决定性的作用，其中不仅需要考虑耗时、准确率等问题，特征选取策略的及时更新也十分重要，因此需要研究流量特征的自动提取方法。图 1 中给出了流量分类的主要工作过程，并分析了特征提取技术的研究内容，理想的分类系统可以在较少

的人工指导下运行，并保持良好的分类性能。

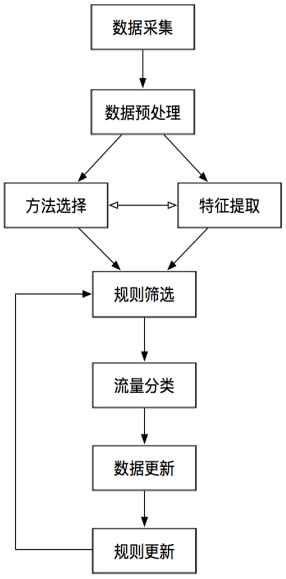

图 1 流量分类工作流程

2. 网络流量分析技术

网络流量分类是根据不同网络应用流量的特点，实现网络流量的对比、分类，在网络流量管理、信息监测或者网络安全防护工作上，流量分类都是一项重要技术手段。流量特征的提取，需要依据算法的选择，而算法的选择，取决于流量分析的具体方式，三者相互制约。根据分析层面可以将流量分析技术分为三类，即基于端口的分析、深度包检测（Deep Packet Inspection，DPI）以及深度流检测（Deep Flow Inspection，DFI）。

2.1. 基于端口的流量识别

早期互联网应用种类较少，不同协议的应用只需根据标准端口号即可进行区分，但互联网环境日渐复杂，出现了多种端口伪装技术，同时新兴的互联网协议，如 P2P 协议使用了动态端口技术导致该方法的准确率大大降低，目前只能作为辅助分析手段。[13]

2.2. 基于深度包检测的流量识别

深度包检测技术是一种基于数据包载荷的流量分析技术。由于相同应用的数据包内容有一定的相似性，可以从数据包的内容中提取一定的特征作为规则，并建立规则库，通过内容匹配来进行流量分类。这种方法的优点是识别精度较高，但是由于需要获取数据包内容，涉及了用户隐私问题，同时在高速网络中，分析的实时性无法得到保证。现今网络环境迅速变化，依赖人工建立规则库的开销巨大，且无法对规则库进行及时的更新，而大量的规则也会引起分析速度下降等问题，因此现在 DPI 的研究方向除了算法识别精度的提升外，还包括算法运行速度的提升以及规则的筛选问题。

2.3. 基于深度流检测的流量识别

深度流检测技术是一种基于网络流量整体统计特征的分析技术。该技术主要依赖机器学习算法，可以实现流量的自动、准确分类。但是由于需要提取网络流的整体特征，很难做到流量的实时分析，且训练得到的模型在处理未知类型流量时无法拥有良好的健壮性。目前针对该技术的研究主要方向有算法的选择、对比和优化，特征的有效选取以及未知流量的识别等。

当前的机器学习算法按照样本数据的形式，可分为无监督学习、有监督学习以及半监督学习。无监督学习算法的主要作用是对数据进行聚类，这类算法不需要初始数据的分类标签，可以自动将数据进行划分，但无法得到数据的类别，常见的无监督学习算法有 K 均值、DBSCAN、谱聚类等。有监督学习算法要求样本数据带有标签，这类算法性成一种从数据属性到数据类别的映射模型，并利用该模型对未来数据进行分类，常见的有监督学习算法有决策树、支持向量机（Support Vector Machine）、贝叶斯方法以及神经网络等。虽然有监督学习算法可以实现数据的有效分类，但对初始数据的标定工作难度较大，半监督学习则可以有效解决这一问题。半监督学习算法首先对数据进行初始聚类，之后再使用少量已标定数据确定类别信息。

目前基于机器学习的流量分类，已成为国内外研究的研究重点。文献 [6]中采用了改机的朴素贝叶斯算法进行流量分类，通过在概率公式中引入改进的核函数，结合一定的特征选择策略，使算法准确率达到了 95% 以上，作者在另一篇文章 [?]中介绍了所使用的 249 种流量特征，并公开了其所使用的数据集。文献 [12]中使用了一种基于 K 均值改进的半监督学习算法，作者改进了 K 均值算法的初始中心选择策略，与传统 K 均值算法和 KKZ 算法相比，性能有所提高，但仍不能保证其实时性，并且在算法性能的对比环节缺少与有监督学习算法的对照，无法体现文中算法的优势。同时，作者也没有解决参数选取、特征选取等问题。文献 [15]中，使用了传统反向传播（Back Propagation, BP）神经网络对数据进行分类，但相对网络结构较为简单，并且未对网络结构做出解释。文献 [11]中使用了基于柔性神经树（Flexible Neuron Tree）的算法，并使用 PIPE 算法优化 FNT 的树结构，同时使用粒子算法（PSO）来进行 FNT 的参数优化。

无论 DPI 技术还是 DFI 技术，都存在一些共同的问题值得研究。例如数据方面可能会存在样本不均文献 [11]中提出了样本数据不均衡的问题，并提出了一种模型来改善此问题。为了保证分类系统实时性和健壮性，算法的运行效率以及自我更新能力也需要深入研究。还有一些算法本

身带有参数, 使用这类算法时还需考虑参数选择对分类效果的影响。由于两种技术各有其弊端, 也有人提出使用混合的流量识别方法，从而更好地贴合应用场景，这也是目前流量分类研究的发展趋势。

3. 网络流量特征提取

特征的提取对于流量识别的速度、精度等有直接影响，人工特征提取的方式开支较大，且有效性和时效性无法得到保证，故自动特征提取成为现在研究的重点。基于 DPI 和 DFI 的流量分析所使用的特征有很大区别，因此相应的特征提取技术也有所不同。

3.1. DPI 分析特征提取

基于 DPI 的流量分析主要使用了流量内容的模式特征，除字符串以外，还可以包括比特串、正则表达式以及序列模式等，这类特征的提取方法是利用数据挖掘算法分析数据包的内容字段与其流量类别之间的关联度，并提取出关联度较高的部分。下面将介绍几种用于模式特征提取的方法。

文献 [12]中使用了一种基于序列模式挖掘的特征提取方法，该算法是经典 PrefixSpan 算法的改进, 直接在数据包的十六进制数序列中进行挖掘, 通过模式增长的方式来生成连续序列模式。其中还使用了偏移属性约束, 在一定程度上减少了算法的计算量。但该算法没有考虑规则库的筛选和更新问题, 文中提出可使用*增量序列模式挖掘*的方法来实现规则库的更新。文献 [17]中在 AutoSig 算法的基础上进行了一定修改，提出了一种自动协议指纹挖掘算法以用于 P2P 流量的特征提取。实验结果表明该算法可以挖掘出人工未能总结出的协议指纹。文献 [15]中使用了固定比特流算法进行流量特征提取，并进行了规则集的筛选工作，结果表明该算法运行时间相对于类 Apriori 算法和类 SPADE 算法有明显缩短。文中还详细介绍了数据的预处理和规则的后处理策略，有一定参考价值。最后，作者还提出可以使用 PCA 的方法来缩短特征提取时间, 并精简规则库。

3.2. DFI 分析特征提取

基于 DFI 的流量分析主要使用了流量整体的统计特征，在特征的选取方面，Moore 提出了针对 TCP 网络流量的 249 种统计特征，利用这些特征，只需了解 TCP 数据包的头部信息即可进行分类 [?]，这对于算法的实时性有重大意义。可以看出 DFI 技术所面对的主要是特征的选择，而不是提取问题，合理的特征选择可以提高算法的精度、速度，从而实现流量的实时检测。下面将介绍几种统计特征选取的方法。

文献 [14]介绍了一种基于关联规则的特征选择算法，该算法基于 Apriori 算法原理，首先寻找各个特征取值与类别之间的关联规则，并按照长度最短、置信度最大、支持度最大的顺序进行

规则挑选，从而提取出可以对数据进行有效分类的属性，作为分类算法的特征。该算法在样本量较大时运行速度较慢。文献 [15]中，首先确定了 37 个特征，将训练模型所使用的特征数限定为 3 个，并利用神经网络算法对所有组合的分类效果进行比对，找到了其中最有效的 10 个特征组合和 10 个无效特征组合。结果表明，有效组合与无效组合所包含特征的重合率很低，说明了单个特征的有效性不会因为组合的变化而有太大波动。但文中并未阐述为何将特征数限定为 3 个，也没有提出一种有效的特征选择策略，仅仅遍历的所有可能的组合得出结果，与实际应用仍有较大差距；同时，作者只利用了 BP 神经网络进行测试，结果可能存在偏差。文献 [11]中基于柔性神经树（Flexible Neuron Tree，FNT）进行了流量的特征选择与识别。

4. 网络流量分类中的主要问题

网络流量分类是一项明确的工作，但由于网络环境的复杂性，再加上在这一工作中使用了多学科的应用技术，使得这一工作中仍可细分为很多部分，这里列举一些更加细化的研究问题，以及相关的研究工作总结

4.1. 针对特定协议或应用类型的分类问题

网络流量分类工作中，根据特定的场景需求，需要有针对性地进行某种流量的有效识别。例如对 P2P 流量、Tor 流量、VPN 流量或者视频、音频流量的识别等。其中针对 P2P、Tor、VPN 等的识别工作是为了加强网络监管，防止技术滥用，而对流媒体的识别则主要用于合理分配带宽资源，从而提高 QoS。文献 [1]的作者进行了 VPN 流量数据的采集，并对 VPN 和非 VPN 流量进行了分类，同时也对流量的应用类型进行了分类，之后，作者又在 2017 年收集了 Tor 流量数据，并完成了类似研究工作。文献 [2]中用 ANN 和 SVM 对 Tor 数据进行了分类，并对比了两种算法的效果，最后通过分析结果来说明 Tor 对于互联网用户的保护作用。[4]中的作者于 2018 年进行了 VPN 数据的采集，并用利用多层感知器（神经网络）算法对其进行了分类。由于网络中部分攻击手段具有固定的模式，因此流量分类也可直接用于部分威胁的检测，Wang 等人在 [8]中利用网络流量进行了恶意流量的检测，并应用了近几年备受关注的卷积神经网络（Convolutional Neural Network），通过截取数据包的前 784 个字节，并将其转化为灰度图片，作者发现相同应用的数据包的灰度图之间具有较强的相似性，根据这一特点，作者直接使用 CNN 来进行流量的分类。但由于没有使用时间特征，该方法仍可改进，如使用 RNN 进行训练。Wang 和 Yang 在 [9]中研究了加密 P2P 流量的识别，使用了隐式马尔科夫模型，但只分析了较为简单的流量行为。Mujtaba 则对通过隧道传输的流量进行了识别 [7]，作者指出，对于隧道流量，PSD 是一个有价值的特征。

[3]文献 [16]的作者则研究了 SSL 流量的分类，同时使用 SMOTE 算法解决数据不均问题。[10]中，作者对 P2P 流量进行了分类，其中也使用了重抽样的方法来解决数据不均衡的问题。高

长喜等人在文献 [18]中结合了 DPI 和 DFI 的方法，进行了加密流量的应用类型识别。

4.2. 数据不平衡问题的处理

数据不平衡是机器学习领域的十大问题之一，在流量分类中这个问题尤其明显，特别是当研究目标为某类特殊流量时，因为网络中最常用几类的应用和协议如 Web 浏览等所产生的流量占据了一半以上，而流量分类问题常关注的流量类型则占比较小，不均衡问题严重影响了分类效果。为解决此问题，目前主要有以下解决方案：重抽样、Cost-sensitive 方法以及特征选择方法，三种方法各有其优劣。

重抽样方法可以有效降低由数据不平衡带来的分类器偏差，但从理论上来说，这种方法改变了数据的抽样分布，对于基于概率假设的分类器有较大影响，但对于树分类器、神经网络等影响并不大。重抽样方法分为过采样、欠采样、以及组合方法，过采样方法从少数类中抽取样本，并加入训练集中，这种方法会增加训练时间，同时可能造成过拟合；欠采样方法将多数类中的部分数据从训练集中剔除，使数据均衡化，这一方法降低了训练时间，但存在丢失重要信息的风险；组合方法结合二者的特点，但具体应用仍有待研究（没怎么看到过）。

Cost-sensitive 方法不会改变数据分布，且消耗的计算资源也较少，但是目前没有成熟的理论来确定如何定义合理的 Cost，特征选择方法可以在一定程度上提升分类器性能但由于特征选择是数据分析中重要的一步，其主要目的是提升算法的整体性能，难以与数据不均问题的优化目标相结合，且该方法需消耗大量计算资源。针对重抽样方法存在的问题，可以通过结合梯度增强决策树（Gradient Boosting Decision Tree，GBDT）算法来解决。GBDT 结合了多个简单决策树，通过一定的决策集成策略，使得多个弱决策树组合成一个强分类器。这类方法的有以下几个优势：首先，组合分类器的算法可以有效防止过拟合的发生，因为在训练过程中，每个基分类器都只利用训练集的部分数据，且最终决策需要综合全部基分类器的结果，少数过拟合的现象并不影响整体的性能；其次，GBDT 算法基于决策树，树算法无需对样本分布进行假设，所以可以有效结合采样方法；第三，这类方法对噪声不敏感，分类精度高；最后，Ensemble 的性能受到基分类器的影响，且每个基分类器的分类效果越好，则整体效果也越好，所以在特征选择时，可以用决策树来代替 Ensemble，而 GBDT 的基分类器是决策树算法，决策树是一种简单快速的算法，这将显著提高特征选择的效率。树算法的另一个优势在于，特征的有效性可以被明确地定义，因为在这类算法中，有效的特征就是信息增益最大的特征，因此可以首先利用全部特征对 GBDT 进行一次训练，再从训练结果中对特征的重要性进行排序，选取一定数量的特征后再应用特征选择算法

4.2.1. 重采样结合 GBDT

GBDT 的一个有效实现是 LightGBM（Light Gradient Boosting Machine），LightGBM 于 2017 年提出，主要运用了两种优化方法，这使得 GBDT 算法的运行效率大大提高，成为目前最有竞争力的算法。第一个方法是基于梯度的单边采样（Gradient-Based One Side Sampling），作者借鉴了 Adaboost 算法中的思想而提出了该方法。Adaboost 是最早提出的 Boosting 算法，该算法

在训练新的基分类器时，会对之前之前未能正确分类的样本赋予更大的权值，强化其对后续分类器的影响，从而有效提高分类的准确率，但这样做的缺点是使得算法对噪声较为敏感。在 GBDT 中，由于树算法的特点，我们无法利用样本的权值来进行训练，但可以相应地使用梯度来代替，也就是 GOSS。GOSS 的解释如下：机器学习的本质是进行目标函数的最优化，而在优化过程中，若目标函数在某一处的梯度较小，我们可以认为该样本中的信息已经被"学习到了"，所以设定某一梯度阈值，将每次训练的样本分为两部分，对于梯度较小的那部分进行欠采样，从而弱化了这类数据的影响，强化了未被学习的样本。但我们还应该注意到，GOSS 是为了加快运行而提出的，它所带来的也只是效率的提升，因为它没有在训练过程中加强或削弱某一类的影响，而只是根据样本的梯度进行欠采样，而不是根据类别，故无法解决数据不平衡的问题。我们可以看到，GOSS 中包含了欠采样的机制，因此将 LightGBM 与欠采样相结合几乎不会带来任何提升。但正是由于 GOSS 的存在，以及 GBDT 算法本身的特点，即使训练集中存在冗余数据，也不容易造成分类器的过拟合，因此有效弥补了过采样方法的缺陷。而从 LightGBM 自身角度来看，由于每次训练都会随机在训练集中抽取样本，当少数类样本占比较小时，尤其在少数类别为多个时，可能会出现某些类别数据缺失或者训练不足的情况，这一影响在每个基分类器上都会发生，最终导致整体对少数类的训练不足。而通过使用过采样，我们提高了少数类样本的比例，使得每个基分类器都能得到足够的数据进行训练。

4.3. 特征选择算法

特征工程在所有数据分析的工作中都是十分重要的一环，在网络流量的分类中，可提取的特征有限，2005 年 Moore 提出的特征几何几乎涵盖了所有有效特征。[5] 因此，当前工作重点在于特征特征选择，网络流量分类中的特征选择包含两方面，首先从实际角度出发，一些网络特征是不容易获取的，例如完整流的统计特征需要等待整个流结束后才可以获取，为此部分研究提出利用发生于流量初期的数据包来提取特征，从而实现流量的及时分类，这对于一些持续时间较长的流量分类更有意义，如基于 UDP 协议的流量，由于 UDP 协议没有完整的确认机制，导致主机需要等待一段较长的时间才能够确认流的结束，故使用早期流量分析将有效解决这类流量的实时分类问题。其次，从算法的角度，过多的特征会降低运行速度，但随意删减特征会造成信息的丢失，所以需要有效的特征选择策略，根据需求来选取最有价值的特征，在保证分类性能的前提下，尽可能降低计算资源的消耗。

4.3.1. 特征选择算法

4.4. 其他方面

由于网络流量的高速性和多变性，分类模型的高效性、可扩展性等也是较为关键的问题。在一些常见的应用场景，对算法的实时性要求并不严格，更多的是强调算法的分类精度，而在网络中，若想要进行有效的监管和防护，就需要分类器及时对流量进行识别，这对算法的运行效率提

出了一定要求。同时，网络流量不同于通常的识别对象的另一点在于，网络流量的特征可能会不断变化，产生概念漂移问题，或者在某一时刻出现新的流量类型等。解决流量多变性问题，可以从两方面入手，第一，定期进行模型的更新或重新训练算法，这就需要提高算法效率，减少运行时间，从而降低重新训练的代价，第二，通过提高算法的健壮性，使模型有效期延长，降低重新训练的频率。

5. 总结

参考文献

[1] Gerard Draper-Gil, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. Characterization of encrypted and vpn traffic using time-related. In Proceedings of the 2nd international conference on information systems security and privacy (ICISSP), pages 407–414, 2016. 5

[2] Elike Hodo, Xavier Bellekens, Ephraim Iorkyase, Andrew Hamilton, Christos Tachtatzis, and Robert Atkinson. Machine learning approach for detection of nontor traffic. arXiv preprint arXiv:1708.08725, 2017. 5

[3] Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun, and Ali A Ghorbani. Characterization of tor traffic using time based features. In ICISSP, pages 253–262, 2017. 5

[4] Shane Miller, Kevin Curran, and Tom Lunney. Multilayer perceptron neural network for detection of encrypted vpn network traffic. 5

[5] Andrew Moore, Denis Zuev, and Michael Crogan. Discriminators for use in flow-based classification. Technical report, 2013. 7

[6] Andrew W Moore and Denis Zuev. Internet Traffic Classification Using Bayesian Analysis Techniques. SIGMETRICS Perform. Eval. Rev., 33(1):50–60, 2005. 3

[7] Ghulam Mujtaba and David J Parish. A statistical framework for identification of tunnelled applications using machine learning. Int. Arab J. Inf. Technol., 12(6A):785–790, 2015. 5

[8] Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye, and Yiqiang Sheng. Malware traffic classification using convolutional neural network for representation learning. In Information Networking (ICOIN), 2017 International Conference on, pages 712–717. IEEE, 2017. 5

[9] Xiaolei Wang, Jie He, and Yuexiang Yang. Identifying p2p network activities on encrypted traffic. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on, pages 893–899. IEEE, 2014. 5

[10] Weicai Zhong, Bijan Raahemi, and Jing Liu. Learning on class imbalanced data to classify peer-to-peer applications in ip traffic using resampling techniques. In Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pages 3548–3554. IEEE, 2009. 5

[11] 彭立志. 基于机器学习的流量识别关键技术研究. PhD thesis, 哈尔滨工业大学, 2015. 3, 5

[12] 林冠洲. 网络流量识别关键技术研究. PhD thesis, 北京邮电大学, 2011. 1, 3, 4

[13] 柏骏, 夏靖波, 吴吉祥, 任高明, and 赵小欢. 实时网络流量分类研究综述. 计算机科学, 40(9):8–15, 2013. 2

[14] 武建华, 宋擒豹, 沈均毅, and 谢建文. 基于关联规则的特征选择算法. 模式识别与人工智能, 22(2):256–262, 2009. 4

[15] 牟澄. 互联网流量特征智能提取关键技术研究. PhD thesis, 北京邮电大学, 2014. 3, 4, 5

[16] 陈雪娇, 王攀, and 刘世栋. 网络应用流类别不平衡环境下的 ssl 加密应用流识别关键技术. 电信科学, 31(12):83–89, 2015. 5

[17] 马婧. 网络流量特征提取与流量识别研究. PhD thesis, 北京邮电大学, 2012. 4

[18] 高长喜, 吴亚飚, and 王枞. 基于抽样分组长度分布的加密流量应用识别. 通信学报, 36(9):65–75, 2015. 6