文章编号:1007-5321(2011)01-0103-04

在线聚类的网络流量识别

张 剑[1,2], 钱宗珏[1], 寿国础[1], 胡怡红[1]
(1. 北京邮电大学 信息与通信工程学院, 北京 100876; 2. 青岛理工大学 电子与通信工程学院, 山东 青岛 266033)

摘要: 针对网络流量在线识别的难题, 提出一种聚类算法和在线流量识别方案. 以网络数据流的若干初始数据包作为子流, 提取子流的统计特征, 应用基于滤波器算法的属性相关性算法提取子流最佳特征子集, 并提出基于密度的在线带噪声空间聚类算法对子流特征向量进行聚类, 采用优势概率业务实现聚类和应用类型的映射. 实验结果表明, 该方案具备识别新应用类型和加密数据流的功能, 且能实现在线的网络流量分类.

关 键 词: 流量识别; 在线聚类算法; 特征选择
中图分类号: TP181 **文献标志码**: A

Network Traffic Identification Based on Online Clustering

ZHANG Jian[1,2], QIAN Zong-jue[1], SHOU Guo-chu[1], HU Yi-hong[1]
(1. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Communication and Electronic Engineering, Qingdao University of Technology, Shandong Qingdao 266033, China)

Abstract: To solve the problem of network traffic identification online, a clustering algorithm and a traffic identification scheme is proposed. The scheme uses a few number of the initial data packets in the flows as a sub-flow, extracts the statistical features from sub-flows, and extracts the best feature subset of sub-flows by applying correlation-based filter approach. The network traffic flows are clustered by on-line density based spatial clustering of applications with noise algorithm, and mapped to application types by the dominant application in clusters. Experiments show that the scheme can identify new application types and encrypted flows, and can be implemented in online network traffic classification.

Key words: traffic identification; online clustering algorithm; feature selection

随着网络业务类型的多样化, 网络流量的测量与分析成为关注的热点[1]. 及时准确地识别网络流量对于流量工程及网络安全管理等有重要意义. 网络流量识别技术主要有端口方法、深度包检测方法、流统计分析方法和机器学习等[2]. 端口的方法实现简单, 但越来越多的应用采用动态端口技术, 限制了该方法的应用. 深度包检测方法通过分析数据包中的载荷来判断应用类型, 具有准确性高的优点, 其缺陷是无法检测加密数据流, 另外存储开销和计算量大, 不适合高速在线流量分类. 流量统计特征及机器学习对流量进行分类[3-5]的方法中, 分类算法运行速度快、准确度高, 但需要训练集建立流量模型, 训练集的选择对分类的准确度有较大影响; 分类的业务类型是已知的, 对未知类型无法识别.

本文提出一种基于密度的在线噪声空间聚类(OL-DBSCAN)算法及接入网流量分类方案. 该方案采用流的若干初始数据包作为子流[4], 基于子流的 OL-DBSCAN 算法满足接入网要求对数据流的早

收稿日期: 2010-03-09
基金项目: 国家高技术研究发展计划项目(2008AA01Z218)
作者简介: 张 剑(1974—), 男, 讲师, E-mail: zhangj9860@sina.com; 钱宗珏(1930—), 男, 教授, 博士生导师.

期识别要求. 该方案包含了自适应在线聚类算法、深度包检测(DPI)检测机制,具备识别加密数据流、新应用类型的能力, 能适应数据流特征随时间变化.

1 接入网流量在线分类方案

本文提出了基于聚类算法的流量在线分类和识别方案, 如图1所示,网络数据流速度在线实时处理数据流模块主要包括子流的生成、特征提取、聚类、应用层协议类型映射等;在线低速处理模块包括子流应用识别、应用类型映射表的生成、流DPI及异常流量检测结果的反馈等. 方案的核心部分是将网络流量以子流的形式采用OL-DBSCAN算法进行聚类, 并应用DPI技术完成对抽样子流应用层协议的识别.

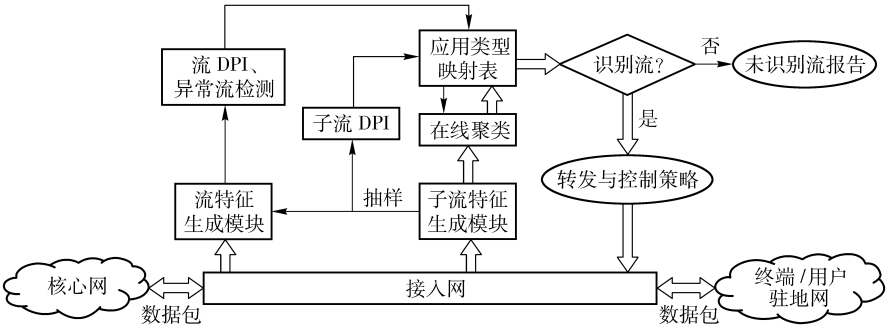

图1 基于聚类的接入网络流量识别和流量分类处理示意图

1.1 子流特征生成模块

在流量识别和分类的研究中广泛采用5元组定义流[6], 即{source IP, destination IP, source PORT, destination PORT, protocol type}. 文献[7]采用对流中所有数据包进行统计特征分析, 该方法只能用在离线的数据流分析,不能满足实时在线的流分析.

子流特征生成模块以流的前 n (5~6)个数据包形成子流[8], 并计算其到达时间、间隔、包大小等参数的均值、方差等统计特征, 作为聚类的参数. 方案中采用NetMate实现子流生成及特征提取功能.

为减少计算复杂度和资源开销, 剔除冗余信息特征,提高应用类型识别准确度, 运用基于filter相关性特征选择(CFS)方法从流量属性特征集中选取最佳的特征子集.

1.2 子流应用识别模块

抽样可降低对检测设备大内存和中央处理器(CPU)能力的需求. 目前广泛采用的2类抽样机制是数据包抽样和流抽样. 数据包抽样相对简单,但在推断如原始流分布等流特性方面还不够准确. 流抽样克服了数据包抽样的局限性,其典型抽样方法为随机流抽样和智能流抽样. 本文方案采用随机流抽样.

通过抽样获得子流子集 $S=\{s_i,\ 0<i\ll N\}$, 应用层协议类型及异常流量类型集为 $Y=\{y_j,\ 0<j\leqslant K\}$, 应用DPI技术确定了一部分子流的协议类型, 即生成$\{s_i,\ y_j\}$映射关系表.

1.3 应用映射模块

该模块实现簇到应用类型的映射功能. 采用基于簇内应用层协议类型最大概率完成映射. 设聚类形成的簇集为 $C=\{c_i,0<i\leqslant M\}$, 其中 M 为聚类数,计算簇内各类协议的概率为

$$P(Y=y_i\mid c_j)=\frac{N_{\text{um}}(s_i\in c_j,y_i)}{N_{\text{um}}(s_i\in c_j)} \tag{1}$$

其中 N_{um} 为满足条件的子流数. 簇 c_j 的应用类型映射为

$$L(c_j)=\max P(Y=y_i\mid c_j) \tag{2}$$

如果簇中没有应用类型映射的对象, 则标记该簇的应用层协议类型为未知 T_{unknown},即

$$L(c_j)=T_{\text{unknown}} \tag{3}$$

2 聚类算法

2.1 算法选择问题描述

目前一些聚类技术的分类方法采用基于距离的原理,缺点在于所形成的簇是球形的分割面. 基于密度的聚类算法可形成任意边界形状的簇. DBSCAN算法是基于密度算法的代表.

聚类算法根据流量特征直接进行聚合, DBSCAN算法需要2个输入参数 v_{eps}、$v_{\text{min-pts}}$. v_{eps} 确定任一子流特征向量邻近距离;$v_{\text{min-pts}}$ 为任一子流特征向量在邻近距离内的其他子流特征向量的最小数目. 对于一个给定的子流特征向量 $\boldsymbol{q}$,在 $\boldsymbol{q}$ 的 v_{eps} 范围内的其他子流特征向量数达到 $v_{\text{min-pts}}$,则 $\boldsymbol{q}$ 被定义

为核对象；而在 $\boldsymbol{q}$ 的 v_{eps} 范围内的其他子流特征向量定义为 $\boldsymbol{q}$ 的直接密度可达对象.

2.2　OL-DBSCAN 算法

基本 DBSCAN 算法步骤如下：

1）所有子流特征向量设置为未归类.

2）选择任一向量为聚类的子流特征向量 $\boldsymbol{p}$.

3）如果 DBSCAN 判断 $\boldsymbol{p}$ 是核对象，则将 $\boldsymbol{q}$ 和所有 $\boldsymbol{q}$ 的直接密度可达对象定义为一个新簇.

4）如果 $\boldsymbol{p}$ 不是核对象，则将 $\boldsymbol{p}$ 归为噪声对象.

5）DBSCAN 算法遍历每个未聚类对象.

由于 DBSCAN 算法遍历整个数据集，不能用于在线数据流的实时分类，所以在该算法的基础上提出 OL-DBSCAN 算法如下：

1）初始化：以 DBSCAN 算法生成聚类，并积累数据流集 Φ.

2）当 Φ 满足聚类 $c_i(0<i\leqslant M)$ 的分类质量时，改进算法对实时数据流特征向量 $\boldsymbol{p}$ 直接进行聚类判决.

3）如果 $\boldsymbol{p}\in c_i$，则将 $\boldsymbol{p}$ 标签置为 c_i，同时用 $\boldsymbol{p}$ 替换聚类 c_i 中数据流时间最早的向量.

4）如果 $\boldsymbol{p}\notin c_i$，则将 $\boldsymbol{p}$ 标签置为 T_{unknown}，保存 $\boldsymbol{p}$ 至 Φ 中，由应用映射模块将 T_{unknown} 类型的数据流向量提交至异常流量检测模块判别类型.

5）当聚类质量下降时，调整算法参数并依据积累数据流集对其进行遍历聚类，形成新簇.

为解决数据流随时间变化所产生的概念漂移问题，OL-DBSCAN 算法评估聚类质量，根据实验结果分析，定义如下 3 个标准：

1）所有聚类中状态为 T_{unknown} 的子流数占所有子流数的比例不大于 10%；

2）聚类数与应用类型之比 $Q\geqslant 3$；

3）应用层协议映射表的任一簇中，概率最大的应用层协议的概率不小于 60%.

若不能满足其中任一标准，则聚类质量不符合聚类标准，将依据 Φ 重新创建各簇，并基于新的簇重新生成簇-应用类型映射表.

3　实验

3.1　数据集

为测试算法的有效性，在北京邮电大学通信与测量实验室收集包含完整数据包的 Comtest 数据集，用 DPI 技术对数据集进行应用识别，作为评估聚类算法的依据. 数据集包含的应用类型如表 1 所示.

表 1　Comtest 数据集中基本应用层协议类型

网络层协议	传输层协议	应用层协议类型	流/%
IPv4	generic	routing	9.8
	UDP	data	10.7
		DNS	0.2
		OICQ	20.4
		BOOSTRAP	1.4
	TCP	HTTP	28.2
		FTP	24.9
IPv6	TCP	bit torrent	1.0
		ICMPv6	0.3
未识别流			3.1

应用 NetMate 实现子流生成及特征提取功能，共生成 38 项特征. 利用 Weka3.7 评估子流数据集的特征子集，从中选取最优特征子集进行聚类. FilterSubset 算法选取 4 个特征，如表 2 所示.

表 2　FilterSubset 算法选取的特征子集

特征	含义
min_fpktl	最小前向包长度
max_fpktl	最大前向包长度
min_bpktl	最小后向包长度
std_fiat	前向到达间隔标准差

3.2　结果及分析

对子流不进行抽样而直接进行应用识别，以测试 OL-DBSCAN 算法对流量识别的能力.

测试指标选取精确度 $p_{i\text{-precision}}$ 和全局准确度 $p_{\text{overall-accuracy}}$. 令 $d_{i\text{-tp}}$ 为正确判断应用类型 i 的流数，$d_{i\text{-fp}}$ 为错误判断其他应用类型为应用类型 i 的流数，则精确度和全局准确度分别为

$$p_{i\text{-precision}}=\frac{d_{i\text{-tp}}}{d_{i\text{-tp}}+d_{i\text{-fp}}} \tag{4}$$

$$p_{\text{overall-accuracy}}=\frac{\sum_{i=1}^{n}d_{i\text{-tp}}}{\sum_{i=1}^{n}(d_{i\text{-tp}}+d_{i\text{-fp}})} \tag{5}$$

基于 Filter 特征子集对 Comtest 数据集进行聚类测试，全局准确度如图 2 所示. 结果显示，v_{eps} 对全局准确度影响较大，而 $v_{\text{min-pts}}$ 影响较小. 实验将主要基于 v_{eps} 来调整 M. 对于主要的应用类型，若 $v_{\text{eps}}=0.03$，$M=6$ 时，web 类型精确度为 89%，文件传输协议(FTP)类型为 93%，即时通信(OICQ)类型为 92%.

聚类算法中对参数的选择是个难题，因为实际

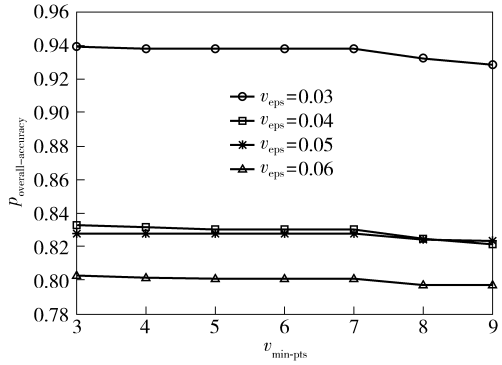

图2 FilterSubset 算法的特征子集的聚类全局准确度

应用中只有少量已知业务类型的流，聚类结果不易评价. 实验研究发现，聚类质量由算法参数决定，同时和 Q 有很强的相关性. 不同 Q 值条件下，每个 Q 值经 10 次聚类实验，将 10 次结果的全局准确度取最大值、均值和最小值，如图 3 所示. 当 $Q \geqslant 3$ 时，聚类可获得约 94% 的全局准确度. 因此将 $Q \geqslant 3$ 作为 OL-DBSCAN 算法中聚类质量评估的标准之一.

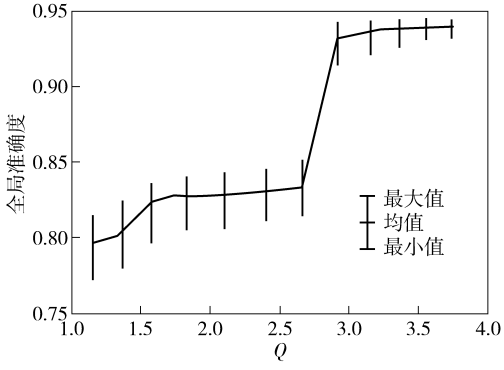

图3 Q 与全局准确度的关系

该方案的优势之一是能识别加密流及新出现的应用. 对于采用 DPI 技术未识别的流，经聚类算法处理后，识别结果如表 3 所示. 结果显示，DPI 技术未识别流中约 40% 为超文本传输(HTTP)加密流，28% 为对等(P2P)应用类型流，剩余未识别流被聚类为 3 个簇，提示可能存在新的应用类型.

表3 聚类算法对 DPI 未识别流的识别结果

占 DPI 未识别流中的比例/%	识别类型
39.7	HTTP 加密流
27.9	P2P 流
11.7	未知应用类型Ⅰ
10.3	未知应用类型Ⅱ
10.3	未知应用类型Ⅲ

4 结束语

本文提出基于聚类算法的接入网流量分类方案，采用流的若干初始数据包作为子流，提取子流的统计特征，并采用 OL-DBSCAN 算法，对流量进行在线聚类. 另外通过 3 个标准评价聚类的质量，解决数据流随时间变化所产生的概念漂移问题. 该方案可实现在线的网络流量分类，具有识别加密流和新应用类型的能力. 方案中采用最佳特征子集提取，排除相关性强的流量特征，在取得较高准确率的同时使计算量大幅度降低.

参考文献：

[1] Callado A, Kamienski C, Szabo G. A survey on Internet traffic identification [J]. IEEE Communications Surveys and Tutorials, 2009, 11(3): 37-52.

[2] Sen S, Spatscheck O, Wang Dongmei. Accurate, scalable in network identification of P2P traffic using application signatures [C] // WWW2004. NY: IEEE Press, 2004: 512-521.

[3] 马永立, 钱宗珏, 寿国础, 等. 机器学习用于网络流量识别[J]. 北京邮电大学学报, 2009, 32(1): 65-68.
Ma Yongli, Qian Zongjue, Shou Guochu, et al. Network flow identification based on machine learning [J]. Journal of Beijing University of Posts and Telecommunications, 2009, 32(1): 65-68.

[4] Karagiannis T, Papagiannaki D, Faloutsos M. Blinc: multilevel traffic classification in the dark [J]. Computer Communication Review, 2005, 35(4): 229-240.

[5] 李卫, 边江, 王盈. 动态网络流分类研究[J]. 电子科技大学学报, 2007, 36(6): 1508-1511.
Li Wei, Bian Jiang, Wang Ying. Research on dynamic network flow classification [J]. Journal of University of Electronic Science and Technology of China, 2007, 36(6): 1508-1511.

[6] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms[C] // SIGCOMM'06 MineNet Workshop. Pisa: ACM, 2006: 11-15.

[7] Moore A, Papagiannaki K. Toward the accurate identification of network applications[C] // PAM 2005. Boston: Springer-Verlag, 2005: 41-54.

[8] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 231-236.