·技术 / TECHNOLOGY·

基于五元组加载荷特征的在线流量分类方法

黄盛林[1, 2, 4]，王恩海[1]，何燕玲[3]，王伟[4]

1. 中国科学院计算机网络信息中心，北京 100190
2. 中国科学院大学，北京 100049
3. 西南科技大学，四川 绵阳 621010
4. 北龙中网 (北京) 科技有限责任公司，北京 100190

摘 要： 准确的流量分类是解决网络拥塞、网络安全监管、流量计费等研究的基础。为了解决在线混合流量 (加密与非加密) 一次性快速分类问题，本文结合传统特征提取和载荷特征提取的优点，提出五元组加载荷 ASCII 出现频次的特征提取，并以此提出一种在线流量快速分类方法。实验表明，在相同算法下，使用本文的特征提取比使用载荷特征提取，整体分类准确率提高了近 4%；基于五元组加载荷特征提取、使用 C4.5 算法的在线流量快速分类方法是可行的。

关键词： 在线流量分类；五元组；载荷特征；C4.5 决策树

doi: 10.11871/j.issn.1674-9480.2015.05.004

Online Traffic Classification Method Based on Five-Tuple Flow with Payload Signature

Huang Shenglin[1, 2, 4], Wang Enhai[2], He Yanling[3], Wang Wei[4]

1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Southwest University of Science and Technology, Mianyang, Sichuan 621010, China
4. Knet Co., Ltd, Beijing 100190, China

Abstract: Accurate traffic classification is the keystone of the network congestion, the network security supervision and the traffic billing. To solve the problem, how the encrypted and unencrypted online traffic mixed flow can

基金项目：国家自然科学基金 (61171109，6137503)；中国科学院计算机网络信息中心“一三五”重点项目 (CNIC_PY_1402)

be quickly identified at once, we constructed the characteristic vector of 5-tuple flow and ASCII occurrences of the payload, and put forward an online traffic classification method based on it. Experiments show that in comparing to the method of traffic load feature extraction, our method increase the accuracy by nearly 4%, and the online traffic classification method is feasible with using the C4.5 decision tree.

Keywords: online traffic classification; 5-tuple flow ; payload signature; c4.5 decision tree

引言

随着互联网的迅速发展，互联网用户规模和网络流量都日益增大。随之带来的网络拥塞、网络安全及网络监管等方面的问题也日趋严重。准确的流量分类是解决这些问题的基础。此外，准确的流量识别对流量计费和应用趋势分析等研究领域也具有极其重要的意义 [1]。基于网络端口的流量分类方法，由于只从协议端口入手，因此无法应对 P2P 等网络端口不固定的新型网络应用；基于深度报文检测的分类方法，需要匹配载荷中具有具体含义的特征字段，因此存在隐私纠纷问题，且对于加密流量识别方面欠佳；基于行为特征的分类方法，需要针对特定网络应用的行为进行研究，如果网络应用自身升级改进，该方法就会逐步失效。针对前面三种方法的缺陷，基于机器学习的分类方法被寄予厚望。

基于机器学习流量分类的核心是流量特征提取和分类算法的选取。文献 [2] 使用传统特征提取，采用基于关联的快速过滤机制 FCBF (Fast correlation-based Filter) 算法和核估计 (kernel estimation) 技术对原始的朴素贝叶斯算法进行了改进，将流量识别的准确率提高到 95% 左右。文献 [3] 在文献 [2] 基础上提出基于信息熵的 C4.5 决策树的流量识别方法。实验表明，利用 C4.5 决策树方法处理流量识别问题，在分类稳定性和数据处理效率上比改进的朴素贝叶斯有优势。文献 [4-5] 在文献 [2] 基础上，尝试了半监督学习算法，也取得了不错的效果。文献 [6] 提出载荷特征提取的加密流量快速识别方法，采用决策树算法针对 80 端口的加密流量进行识别，取得了较好的效果。但是，这些研究都针对离线流量的分类，并未考虑在线流量分类的适用情况。在线流量分类，对特征提取和分类方法的效率都具有更高的要求。文献 [7] 吸取了传统特征提取的优点，利用 TCP 流开始的前 5 个数据包 (排除三次握手数据包)，计算数据包大小、负载大小和到达间隔时间等网络流量的统计特征，使用 NBTree 等进行分类实验，探讨了在线流量分类的适用性，但仅仅针对 TCP 流量。

因此，为了尝试解决在线混合流量一次性快速分类的问题，本文首先结合传统特征提取和载荷特征提取各自的优点，提出五元组加载荷 ASCII 频次的特征提取；然后在此基础上，提出使用 C4.5 算法的一种在线流量分类方法；最后进行仿真实验分别测试分类准确率和效率。本文其余部分安排如下：第 1 节详细介绍五元组加载荷的特征提取；第 2 节介绍在线流量分类方法；第 3 节进行两组分类仿真对比实验，并给出结果分析；第 4 节是本文的结语。

1　五元组加载荷的特征提取

1.1　不同特征提取优缺点分析

特征提取是模式识别的关键问题之一，特征提取结果的好坏直接影响着分类器的分类精度和泛化性能。理想情况下，特征提取是寻找必要的、足以识别目标的最小特征子集的过程 [8]。但针对在线流量的分类，除了分类算法本身的分类效率外，同时也需要兼顾特征获取的效率。

文献 [2-5] 的网络流量属性选择，属于传统的特征提取。其优点是能比较直观的反映网络协议层面的流量属性，从文献 [2-5] 中实验表明分类准确性也很好。但该特征提取基于语义完整的 TCP 双向流，从报头相关属性、分组长度相关属性、时间相关属性等，提取出 249 项网络流量属性，其中有 100 多项是通过傅里叶变换技术获得 [9]。从流量数据中分离出 TCP 双向流和使用傅里叶变换都需要消耗不小的计算

代价，更重要的是，如果从在线流量中获取语义完整的 TCP 双向流的话，需要等待一次完整的 TCP 通信结束。文献 [6] 针对加密流量采用载荷特征提取，该方法无需对数据包载荷进行深入分析，使用 256 维向量描述载荷中 256 个扩展 ASCII 字节发生的频次，很好的避免了载荷数据的隐私纠纷问题。另外，该方法还加入了时间信息，对特定应用程序流量构建 n×256 的特征矩阵，并计算量化后的均值和方差，也需要消耗不小的计算代价。

因此，本文首次提出五元组加载荷的特征提取方法。结合传统特征提取中易提取的五元组特征和载荷特征提取中易提取的 ASCII 出现频次特征，既保留了部分传统特征提取和载荷特征提取的优点，又克服了特征不易构造的缺点。而且五元组与载荷 ASCII 出现频次相对独立，不会出现特征重复等新的问题。综上，不同特征提取的比较如表 1 所示。

1.2 PCAP 格式数据报文的解析

PCAP (Process Characterization Analysis Package) 文件格式是 BPF (柏克莱封包过滤器，Berkeley Packet Filter) 保存原始数据包的格式，很多软件都在使用，比如 Tcpdump、Wireshark 等。

基本格式为：文件头 数据包头 数据报 数据包头 数据报 ……

详细说明如图 1 所示。24 字节的文件头后接着是 16 字节的数据包头，Packet Data 部分就是熟悉的以太网数据报文。传统特征提取，需要提取分组长度相关属性和时间相关属性，会使用到文件头和数据包头的数据。载荷特征提取和本文特征提取只需要数据报文的数据，因此这里不进一步说明文件头和数据包

表1 不同特征提取的比较

Table 1 Comparison of different feature extraction

类别	描述	特点
传统特征提取	构造五元组、分组长度相关属性、时间相关属性的特征向量	仅依赖报头数据；基于语义完整的 TCP 双向流且部分属性需要傅里叶变换获取；分类效果好，构造耗时。
载荷特征提取	构造载荷中 256 个扩展 ASCII 字节出现频次的特征向量	仅依赖载荷数据；易于构造，但仅针对加密流量识别，混合流量识别有待考察验证。
五元组加载荷特征提取	构造五元组加上载荷 SCII 频次的特征向量	依赖报头和载荷数据；结合传统特征提取和载荷特征提取的优点，且易于构造。

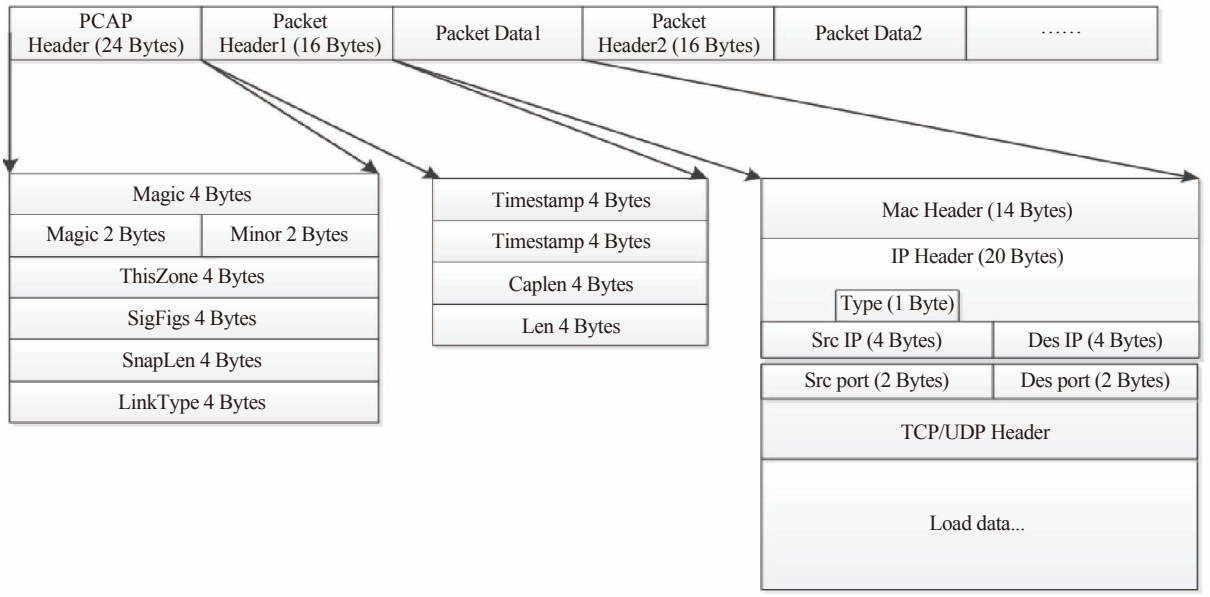


图1 PCAP 文件格式详细说明

Fig. 1 File format description of PCAP

头各字段的含义，详细请参阅文献 [10]。

五元组信息，从数据报文里面的 IP 报头能解析出协议类型、源 IP 地址、目标 IP 地址，从 TCP 或者 UDP 报头解析出源端口、目标端口。载荷特征，从载荷数据 (Load data) 中统计 ASCII 出现的频次。

1.3 五元组加载荷特征向量的构造

线下模型的训练与测试，使用事先采集到的 PCAP 格式样本数据。首先进行数据预处理，将样本数据划分为训练数据和测试数据。事先从训练数据中按应用类型分离出的 PCAP 格式数据，按照应用类型命名，如：http.pcap，并保存至训练数据文件夹下，测试数据同理。然后编程处理生成特征向量的文件。

处理训练数据设计算法步骤如下：

Step 1：五元组 (5 项) 加上单纯统计载荷中 ASCII 出现的次数 (256 项)，共 261 维；扫描训练数据文件夹中的文件名称，确定样本种类 class (1 项)，生成 262 维 ARFF 文件头，写入新建的 training.arff 文件。

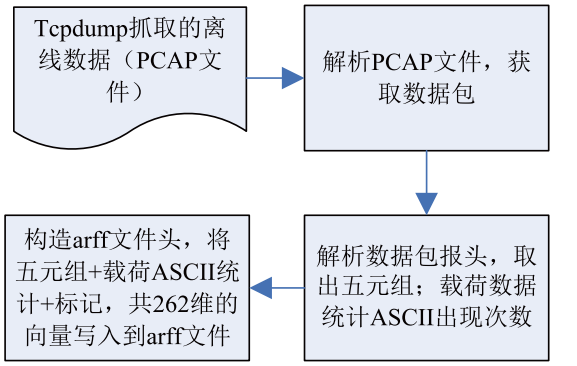
Step 2：解析训练数据文件夹下 PCAP 文件，生成特征向量写入 training.arff 文件。

伪代码描述如下：

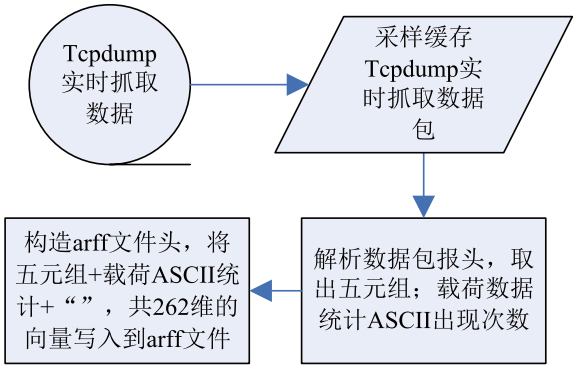
```
参数定义：
    训练数据文件列表 File_List；当前处理文件 File
    流量应用类型 Type；当前读取的数据流字节数
    ByteNum
    特征向量 Vector；training.arff 输出文件 OutFile
For (int i; i<File_List.length; i++){
    File = File_List[i];
    Type = name of File; /* 获取文件名称，即应用类
        型*/
    ByteNum = File.read (Pcap_Header); /* 读取 PCAP
        文件头 Pcap_Header*/
    while(ByteNum > 0){
        ByteNum = File.read(Packet_Header); /* 读
            取数据包头 Packet_Header*/
        IF (ByteNum < 0)
            Break;
        ByteNum = File.read(Packet_Data); /* 读取
            数据报 Packet_Data*/
        Vector = get_vector(Packet_Data); /* 解析出
            五元组和载荷 ASCII 频次 */
        OutFile.write (Vector+”,”+Type+”\n”);
            /*将特征向量写入ARFF文件*/
    }
    File.close()
}
OutFile.close()
```

测试数据处理方式，同理。

待分类的在线流量数据处理方式稍有不同，具体如图 2 所示。在线未知数据打上默认标记后，输入到分类器进行分类。



(a) 离线数据特征向量构造



(b) 在线数据特征向量构造

图2 特征向量的构造

Fig. 2 Eigenvector structure

2 在线流量分类方法

2.1 基于 C4.5 算法的流量分类器

用于流量分类的机器学习方法主要包括无监督方法和全监督方法，此外还有将这两种方法相结合而产生的半监督方法。半监督使用少量的标记数据，无监督基本不使用标记数据，很难保证分类准确性，且所训练出的系统都很难具有较强的泛化能力[1]。因此，使用大量标记数据的全监督分类算法更多被采用到流量分类上面来。

C4.5 决策树算法是一个重要的全监督分类算法，在基于机器学习的网络流量分类的研究领域中占据重要的地位。文献 [3, 11] 进行了 C4.5 算法在网络流量分类中的研究并进行了相关实验，在 C4.5、朴素贝叶斯、贝叶斯网络、SVM 等算法的比较研究中，C4.5 具有最好的分类准确率及分类效率。C4.5 主要有以下几个要点：

(1) 根据信息增益率 (information gain ratio) 来选择属性。克服了用信息增益选择属性时偏向选择值多的属性的不足；

(2) 在决策树构造过程中进行剪枝；

(3) 能够完成对连续属性的离散化处理；

(4) 能够对不完整的数据进行处理。

本文构造的五元组加载荷 ASCII 出现频次的特征向量，其中载荷 ASCII 出现频次并无明确的分布，存在近乎连续的属性。而且对于 261 维网络属性，完全有必要在决策树构造过程中进行剪枝。因此，C4.5 决策树比较适合作为本文分类器的分类算法。更为重要的是，决策树构造完成之后，待分类的在线流量只需要匹配树中节点，就可以确定其类别，效率很高。

本文构造的 C4.5 算法描述如下：

算法：C4.5 决策树流量分类算法 *C4.5_Traffic_Classification()*。

输入：

◆ 数据划分 *D* 是五元组加 256 个 *ASCII* 频次和对应类标号集合；

◆ *attribute_list* 是候选属性的集合；

输出：一棵决策树。

方法：

(1) 创建一个节点 *N*；

(2) **IF** *D* 中的元组都是同一类 *C* **then**

返回 *N* 作为叶节点，以类 *C* 标记；

(3) **IF** *attribute_list* 为空 **then**

返回 *N* 作为叶节点，标记 *N* 为 *D* 中的多数类；

(5) **For each** *attribute_list* 中的属性

计算信息增益率 *information gain ratio*；

(6) *N* 的测试属性 =*attribute_list* 具有最高信息增益率的属性；

(7) **IF** 测试属性为连续型 **then**

找到该属性的分割阈值；

(8) **For each** 由节点*N*分裂的新叶子节点{

IF 该叶子节点对应的样本子集*D**为空 **then**

分裂叶子节点生成新叶节点，将其标记为 *D* 中出现最多的类;

Else

该叶子节点上执行*C4.5_Traffic_Classification* (*D**, *D**. *attribute_list*) 递归；

}

(9)计算每个节点的分类错误，进行剪枝。

(10)返回 N;

2.2 在线流量分类框架

本文的在线流量分类框架如图 3 所示。该框架主要包含 3 个模块：数据采集处理模块、特征向量快速构造模块以及分类器快速分类模块。数据采集处理模块是前期的部署工程性工作，这里不阐述。特征向量快速构造模块及分类器快速分类模块如 1.3 节、2.1 节所述。

图中两处阴影：“在线数据特征向量快速构造”和“样本数据训练好的快速分类器”，是本文在线流量分类方法的关键。

3 仿真实验与结果说明

3.1 实验环境

3.1.1 实验数据

实验样本数据是西南科技大学校内网交换机

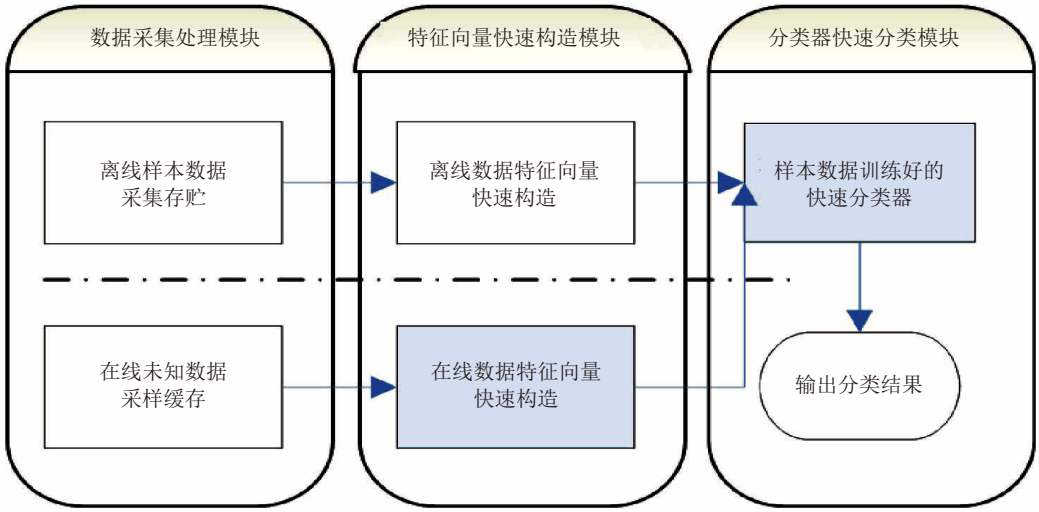


图3　在线流量分类框架

Fig. 3　Online classification framework

口，2013 年 10 月 21 日 16:27~16:28 两分钟采集到的 400M 流量数据 (共 2 个 pcap 文件，每分钟 200M)。从流量数据中识别分离采样出如表 2 所示的样本数据。样本数据中包括 QQ、HTTPS 加密流量，也包括 DNS、SMTP、HTTP 非加密流量。

将总数据集中前一分钟分离出的样本数据作为训练数据，详细如表 3 所示。后一分钟分离出的样本数据作为测试数据，如表 4 所示。

表2　总实验数据集信息

Table 2　The total experimental data set information

Type of Flow	Num of Flow	Percent(%)
HTTP	20 424	45.60
DNS	19 181	42.83
QQ	2402	5.36
HTTPS	2272	5.07
SMTP	507	1.14
Total	44 786	100

表3　训练数据集信息

Table 3　Training data set information

Type of Flow	Num of Flow	Percent(%)
HTTP	11 104	45.13
DNS	9936	40.38
QQ	1874	7.62
HTTPS	1408	5.72
SMTP	284	1.15
Total	24 606	100

表4　测试数据集信息

Table 4　Test data set information

Type of Flow	Num of Flow	Percent(%)
HTTP	9320	46.18
DNS	9245	45.81
QQ	528	2.62
HTTPS	864	4.28
SMTP	223	1.11
Total	20 180	100

3.1.2　仿真环境

本文的 PC 环境是 Inter i5，4GB 内存，Win7 系统。在实验 2 中，我们分别使用物理机 Win7 作为数据接收端和 VMware 虚拟机 Ubuntu 作为数据发送端、并使用 XCAP1.0.3 用于测试数据集的发送，模拟真实的网络环境。借助 JPCAP，通过编写 Java 语言程序，进行在线流量的分类仿真。

3.2　评价指标

本实验采用整体准确率，单类别 F-measure 值作为准确性评价指标。整体准确率 (P_{all}) 为数据集中，正确分类的样本数 (N*) 与总样本数 (N) 的比值：

$$P_{all}=N^{*}/N \tag{1}$$

F-measure 值是将“准确率P”和“召回率R”两个指标组合形成的。“召回率”为正确分类的样本数与未分类前该类别的样本总数。单类别 F-measure(i) 值：

$$F_{measure}(i)=2P_iR_i/(P_i+R_i) \quad (2)$$

其中 i 表示流量应用类型。

整体准确率值越高，且单类别 F-measure (i) 值也高，说明分类准确且稳定。

本实验采用在线相同测试数据集分类消耗时间作为效率性评价指标。消耗时间越短，代表该算法构造的分类器处理速度越快，越适用于在线流量分类。

3.3 实验与结果分析

3.3.1 实验1：不同特征提取的对比实验

为了验证本文特征提取的优势，将本文特征提取与载荷特征提取，在同样的数据集上且采用同样的 C4.5 决策树分类算法，进行对比实验。

整体准确率对比结果如表 5 所示。使用载荷特征提取，采用 C4.5 算法对混合流量分类，整体准确率为 95.98%；但使用五元组加载荷特征提取，准确率高达 99.82%，提升了近 4% 整体准确率。

单类别 F-measure(i) 对比结果如图 4 所示。使用载荷特征提取，采用 C4.5 算法对混合流量分类时，对 HTTPS 识别准确率很低，仅为 54.2%；对 SMTP、QQ 识别准确率分别为 75.7%，86.7%，都不算高。而使用五元组加载荷特征提取，同样采用 C4.5决策树，在所有类别中都维持了较高的 F-measure 值。

在载荷特征提取基础上，五元组加载荷特征提取使用数据报头中多获取的 5 个属性，重构决策树，使得原本较难识别的 HTTPS、QQ、SMTP 流量得到了更好的识别，从而提高了分类准确率。

3.3.2 实验2：在线流量分类对比实验

使用五元组加载荷特征提取，分别对朴素贝叶斯 (NB)、K-近邻 (KNN)、C4.5 三种算法进行在线流量分类实验对比。事先使用训练数据集训练好分类器。然后通过在线流量模拟仿真环境，从虚拟机中发送测试数据包，物理机接收数据包进行分类，并记录消耗时间、分类整体准确率和单类别 F-measure(i) 值。每种算法重复上述 10 次实验，各类指标取均值。

三种算法的整体准确率如表 6 所示。C4.5 的整体准确率最高，KNN 次之，NB 最低。单类别 F-measure(i) 值如图 5 所示，NB 算法除了在 DNS 这一类流量上的 F-measure 值较高外，其他类别的 F-measure 值都很低。KNN 算法在 HTTPS 和 SMTP 类别上的 F-measure 值也偏低。C4.5 算法最为稳定，在所有类别中都维持了较高的 F-measure 值。

朴素贝叶斯 (NB) 要求各属性相互独立且服从高斯分布，本文的流量属性显然不符合，于是准确率低；由于每类样本的数目不均，使得 K 个近邻中，大多是样本多的类别，可能是引起样本数量较少的

表5 不同特征提取的整体准确率

Table 5 Overall accuracy rate of different feature extraction

特征提取类别	载荷特征提取	五元组加载荷特征提取
整体准确率(%)	95.98	99.82

表6 三种算法的整体准确率

Table 6 Overall accuracy of three algorithms

指标	NB	KNN	C4.5
整体准确率(%)	56.13	96.69	99.82

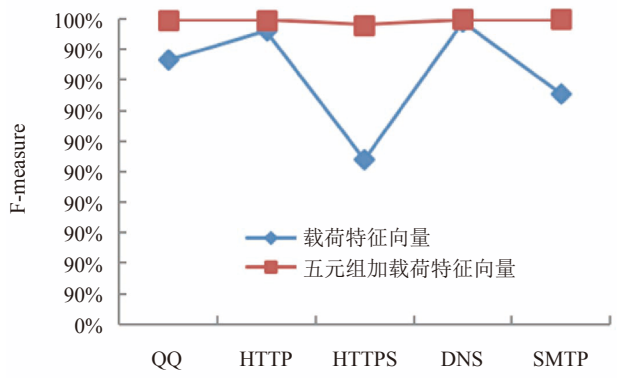


图4 单类别 F-measure(i) 对比结果

Fig. 4 Single class F-measure (i) compare results

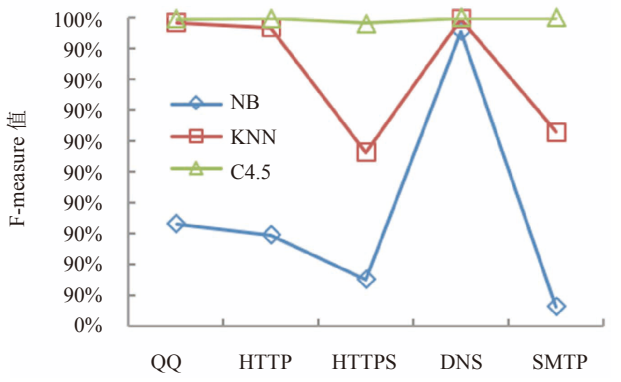


图5 三种算法的单类别 F-measure(i) 值

Fig. 5 Single value of category F-measure (i) in three algorithms

HTTPS、SMTP 分类准确率低的原因。

三种算法的消耗时间如表 7 所示。KNN 的消耗时间最长，NB 次之，C4.5 最短，仅是 NB 的 12%、KNN 的 2‰，相当于每秒钟分类 1.15 万个包 (20 180包/1.75s)。

表7　三种算法的消耗时间

Table 7　Time consuming of three algorithms

指标	NB	KNN	C4.5
消耗时间(s)	14.62	912.71	1.75

使用 KNN 时，每个数据包分类需要找出与其最近的K个近邻，因此特别耗时；使用 NB 时，每个数据包分类需要计算各类别的条件概率后取最大值；而使用 C4.5 时，只需对决策树进行查找，故而 C4.5 耗时短。

4　结语

基于机器学习方法进行流量分类是近年来的研究热点。本文研究了基于机器学习的在线混合流量分类问题，提出五元组加载荷特征提取，并在此基础上提出在线混合流量的分类方法。然后，搭建实验环境，进行了两组分类对比实验。通过对比实验我们发现：(1) 采用 C4.5 算法，使用五元组加载荷特征提取比载荷特征提取，提高了近 4% 的整体准确率；(2) 在线流量分类中，C4.5 算法比 NB、KNN 算法的准确率高、消耗时间短，本文的在线流量方法切实可行。

由于本文样本数据中的流量类别比较少，训练出来的快速流量分类器能识别的流量也有限。获取更多的样本数据并与其他的在线流量分类方法对比，研究更适合在线流量分类的特征提取和分类算法，是本文下一步的研究工作。

参考文献

[1] 熊刚, 孟姣等. 网络流量分类研究进展与展望 [J]. 集成技术, 2012, 1(1): 32-42.

[2] Moore AW, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]. Proc. of the 2005 ACM SIGMETRICS Int’l Conf. on Measurement and Modeling of Computer Systems. Banff, 2005. 50-60.

[3] 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.

[4] 周文刚等. 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.

[5] 张震, 汪斌强等. 基于近邻传播学习的半监督流量分类方法 [J]. 自动化学报, 2013, 39(7): 1100-1109.

[6] 陈伟等. 基于载荷特征的加密流量快速识别方法 [J]. 计算机工程, 2012, 38(12): 22-25.

[7] 赵树鹏, 陈贞翔, 彭立志. 基于流中前 5 个包的在线流量分类特征 [J]. 济南大学学报, 2012, 26(2).

[8] 姚旭, 王晓丹, 张玉玺. 特征选择方法综述 [J]. 控制与决策, 2012, 27(2): 161-166.

[9] Moore AW, Zuev D, Crogan M. Discriminators for use in flow-based classification[J]. Technical Report, RR-05-13, Queen Mary University of London, 2005.

[10] 李山松. 利用 Wireshark 软件进行信令分析 [J]. 计算机与网络, 2013, 20: 64-66.

[11] 徐鹏等. 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.

收稿日期：2015 年 8 月 16 日

黄盛林：中国科学院计算机网络信息中心，硕士研究生，主要研究方向为数据挖掘、企业信用数据分析。E-mail: *huangshenglingf@163.com*

王恩海：中国科学院计算机网络信息中心，高级工程师，主要研究方向为数据分析、互联网管理。E-mail: *wangeh@cnic.cn*

何燕玲：西南科技大学，北京理工大学在读博士研究生，主要研究方向为网络流量测量与分析。E-mail: *heyanling@126.com*

王　伟：北龙中网 (北京) 科技有限责任公司，研究员，主要研究方向为 DNS 结构研究，数据挖掘、互联网金融信用评价研究。E-mail: *wangwei@knet.cn*