基于关键字的单协议分类

郑 杰 李建平
(电子科技大学计算机科学与工程学院 成都 611731)

摘 要 网络协议是网络通信中一系列标准的集合,未知协议的识别和分析对网络监管、保障网络安全具有重大意义。协议识别技术多种多样,但大都不适用于二进制的协议识别。在此针对现有的协议识别技术的局限性,提出了一种在双方单协议通信环境下的多种类型二进制数据帧的协议识别方法。该方法首先利用 n-gram 技术对数据帧进行分割,然后利用无监督的特征选择算法提取特征串集合,从而利用聚类算法实现协议消息的识别。最后在 ICMP 上对该方法进行评估,消息识别的准确率和召回率均可达到 90%以上。

关键词 协议识别,单协议,无监督,特征选择,聚类算法
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.014

Classification of Single Protocol Based on Keywords

ZHENG Jie LI Jian-ping
(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract Network protocols are sets of standards for certain network communications. The protocol identification and analysis have great significance for network management and security. Although there are all kinds of protocol identification technology, most of them are not suitable for the binary protocol identification. To address this issue, the paper proposed a novel method of protocol identification which can classify the same protocol into several messages in the environment of single protocol communication. This method utilizes n-gram to segment the data frames and then extracts the set of keywords using unsupervised feature selection algorithm. At last, it implements the identification of different type of messages using clustering algorithm. Finally the method was evaluated on ICMP. The results show that the rate of precision and recall can both reach more than 90%.

Keywords Protocol identification, Single protocol, Unsupervised, Feature selection, Clustering algorithm

1 引言

随着网络应用的不断发展,网络通信环境也变得更加复杂。在真实网络中,无法识别的网络流量占 30%左右,其中甚至包括恶意病毒和其他不确定的数据[1]。恶意病毒通信应用通常使用专有协议,窃取涉密数据,严重危害网络安全。控制和管理这类未知流量,需要对其通信结构进行研究。而通信结构又主要反映了协议内容,因此有效地识别协议类型,对维护网络安全具有重要意义。

最早研究的协议识别技术是基于端口的协议识别技术[2],该技术能识别的协议为在 IANA(Internet Assigned Numbers Authority)[3]中注册端口号的协议。但是随着网络协议的不断发展,大量协议采用自由端口,这使得基于端口的协议识别技术已经不适用于现在的网络环境。文献[4-6]详细论述了导致基于端口算法失效的原因,正是由于这些原因,基于端口识别协议的准确性已经低于 50%。

基于测度的协议识别技术是利用协议规范的不同所造成的流测度的差异来识别各个协议。其中流属性包括流时间、最大包长、最小包长、报文间隔、流方向以及出字节数和入字节数等一系列与流相关的测度。Wright[7]等人提出了一种基于隐式马尔可夫模型进行网络流量分类的方法,对 SMTP、HTTP、FTP、Telnet 等协议进行了识别,取得了较好的效果。Charles Wright[8]等人又在此基础上进行优化,建立了一种只使用报文大小和时间间隔的混合模型,使其识别准确性提高了大约 30%。Moore[5]等人通过直接测量和利用傅立叶变换技术给出了网络流的 249 种属性,以后的研究大都采用这 249 种属性的子集进行特征优化,以期得到协议识别的最简洁的属性集合。Bernaille 等[9]采用无监督聚类的学习方式,仅采用每条流的前 5 个报文的大小构造五维空间,以 K 均值算法进行聚类。对不可知流,采用一个简单的启发式方式——将其归类为最普遍的流量,该方法准确性一般,但是计算代价非常小。

基于测度识别协议无需分析报文体的内容,而是采用机器学习的方式,对于已知协议的报文进行训练,使其把握该类

到稿日期:2015-01-05 返修日期:2015-04-09 本文受中国工程物理研究院科技发展基金(2012A0403021),NSAF 联合基金(U1230106),国家信息安全发展计划(2013F098)资助。

郑 杰(1976-),男,博士生,主要研究方向为信号处理、信息分析,E-mail:1120103@qq.com;李建平(1962-),男,博士,教授,主要研究方向为小波指纹加密系统、分布式网络监控系统。

应用协议的流测度特征以识别新的流量[10,11]。这种协议识别技术消耗的计算资源比较小，对于流量的变化能很好地适应，数据流是否加密对于识别的影响也比较小。但是这两种方法都需要事先对协议进行训练，然后进行识别，并不适用于未知协议识别，并且当前研究的算法都只能将流分成几大类，不能将流数据分成具体的协议类别。

基于负载的协议识别技术，对协议交互过程中产生的报文内容进行分析，根据协议固有的模式特征进行协议识别。这种协议识别技术主要采用正则表达式和固定字符串2种方式来表示协议特征。其中采用正则表达式表示特征的方式一般采用自动机知识，将正则表达式构造NFA，然后将NFA编译成DFA，再将DFA传送给匹配引擎进行协议识别[12]。Deyoung[13]在假设协议格式已知的前提下，引入文法推断中的k-RI和k-TSSI算法提取状态机信息，这种方法只适用于已知协议格式的环境中。Nohl等人[14]提出一种以偏序比对算法为基础，构建有穷自动机来识别报文的方法。王勇[15]等人提出一种基于协议关键字的自动机的构建方法，通过对关键字的提取和协议会话的分析分别建立协议的语言自动机和状态自动机，从而实现了在不需要任何先验知识的前提下对协议规范的提取。基于正则表达式的协议识别方法的识别准确率高、方式灵活，但是大量的匹配消耗了过多的计算资源，还需要在如何进行高效的识别上做进一步的研究。

采用固定字符串表示特征的研究时间较久。Subhabrata Sen等人[16]通过分析5种P2P协议的相关文档和实际报文流量，确定协议的应用层特征串，该方法可以较准确地识别出应用层协议的特征串，但是只适用于有公开协议规则的协议中。王一鹏[17]等人提出了一种基于语义的协议识别方法，这种方法通过3-*gram*对原始数据集进行切分，通过语义进行分词，然后利用LDA算法完成格式无关单元的过滤和协议特征关键字的生成，最后完成聚类和协议格式的提取。这种方法计算过程过于复杂，花费的时间代价过大，并且从语义层面实现关键字的提取，更适合于文本协议或者是语义已知的二进制协议。王勇[15]等人提出一种基于关键字的提取方法，这种方法既适用于文本协议同样也适用于二进制协议。对二进制单协议的数据输入，将其按字节进行切分和拼接得到特定协议的关键字，但是这种方法在实际的操作过程中，对拼接成的长串进行切分时很难把握切分的原则。

从理论上说，基于负载的协议识别方法总可以通过分析协议规范和实际交互的报文得到协议的特征，并且该方法的准确性是目前所有算法中最高的，文中[15,18,19]所研究的算法的误报率均小于10%，但该类算法的时空复杂度也是目前所有算法中最高的，并且随着待识别协议数量的增多而增长。使用基于负载的算法需要不断地跟踪待识别协议的发展，更新的工作量非常大。因此，该类算法通常在待识别数量较少时使用，且需要相当的工作量。

通过对现有的协议识别技术优缺点的分析，本文提出了一种基于关键字的未知协议识别方法。假设通信双方使用单协议，该协议中含有M种消息，本文提出的方法通过关键字提取和聚类，可以将待识别的协议分成不同类型的消息。由于并不清楚协议的类别信息，只能通过特征之间的相互关系进行特征选择，因此将熵值中的互信息用于特征选择，在使用互信息时，通过最大相关和最小冗余的原则可以选择出指定个数的特征。

2 准备工作

本文提出了一种基于关键字的单协议分类模型，该模型在双方使用单协议通信的情况下，可以将未知的二进制协议分成不同类型的消息。该模型包括*n-gram*的生成、特征候选集的拼接、特征选择和聚类4个模块。

(1)*n-gram*的产生：本文将*gram*定义为字节，n为字节的长度，该模块能够将原始的数据帧切分成长度为n的字节，为特征候选集的拼接做准备。该部分的2个关键问题是n值的确定和*n-grams*的筛选，与此对应的2个解决方案分别是齐夫分布和*Jaccard*参数。

首先在n取不同值时($n=1,2,3$)分别对原始数据集进行切分，对每个*n-gram*按其出现的次数进行排序，用x轴表示*n-gram*的排序，y轴表示对应*n-gram*出现的次数，2个坐标轴都用对数来表示，产生的齐夫分布曲线如图1所示，根据曲线图来决定n值的大小。

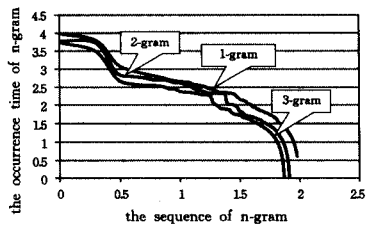

图1 不同n值所对应的齐夫分布

然后，为了防止拼接过程中产生过多的冗余串，需要对产生的*n-grams*进行筛选。本文利用*Jaccard*参数作为选择阈值的指标，其计算公式如式(1)所示：

$$J(A,B)=\frac{\sum_{i=1}^{n} T1_i * T2_i}{\sum_{i=1}^{n} T1_i^2+\sum_{i=1}^{n} T2_i^2-\sum_{i=1}^{n} T1_i * T2_i} \tag{1}$$

其中，A,B分别表示两个用字节表示的特征向量，$T1_i$和$T2_i$分别表示A和B中的第i个特征。通过记录阈值与*Jaccard*值的变化曲线，选择第一次达到最高点的*Jaccard*值所对应的阈值作为所求。本文中产生的阈值与*Jaccard*值的对应关系如图2所示。

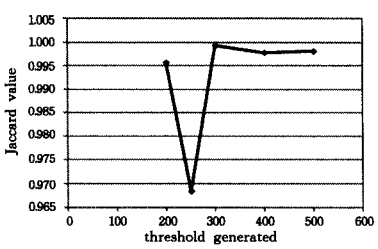

图2 *Jaccard*值与阈值的对应曲线

(2)特征候选集的拼接：特征候选集是利用筛选出来的*n-grams*拼接而成的频繁长串，拼接的依据是*n-grams*在每一帧数据中是否连续出现，如果有2个或2个以上的*n-grams*在一帧数据中是连续出现的，就将它们拼接在一起形成频繁长串。这一模块的目的是产生特征候选集，以此作为下一步特征选择的输入。

(3)特征选择：假设(2)中产生的特征候选集为 $U=\{F_1,F_2,\cdots,F_n\}$，该模块的目的在于从这 n 个特征候选集中找出能够标识每一类消息的特征，以此作为聚类的依据。本文利用熵值中的互信息作为特征选择算法，特征选择的依据是最大相关-最小冗余，即所选择的特征应该与所有的未选特征最大程度地相关，而与已选特征最小程度地冗余。

假设有 2 个随机变量 X 和 Y，X 可能的取值为 $\{a_1,a_2,\cdots,a_n\}$，Y 的取值为 $\{b_1,b_2,\cdots,b_m\}$，随机变量 X 的熵值的计算公式如式(2)所示：

$$H(X)=\sum_{i=1}^{n}p(a_i)\log\frac{1}{p(a_i)} \tag{2}$$

其中，$p(a_i)$ 表示随机变量 X 中 a_i 出现的概率。互信息表示 2 个变量间共同拥有信息的含量，具体来说，互信息的计算公式如式(3)所示：

$$I(X,Y)=H(X)-H(X|Y) \tag{3}$$

其中，$H(X|Y)$ 表示条件熵，条件熵是指一个变量依赖于另一个变量的强弱程度。具体来说，条件熵的计算公式如式(4)所示：

$$H(X|Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(b_j)p(a_i|b_j)\log 2p(a_i|b_j) \tag{4}$$

其中，$p(a_i)$ 表示随机变量 X 中 a_i 出现的概率，$p(b_j)$ 表示随机变量 Y 中 b_j 出现的概率。

本文中，特征候选集的拼接模块所得到的特征候选集用集合 $U=\{F_1,F_2,\cdots,F_n\}$ 表示，集合 U 中的每一个特征 F_i 的取值只有 f_i(表示特征 F_i 出现)和 $\overline{f_i}$(表示特征 F_i 不出现)两种。在进行特征选择之前，先介绍几个定义。

定义 1(相关度)　一个特征 F_i 的相关度就是其与整个特征集合的平均互信息，用式(5)表示：

$$\begin{aligned}\mathrm{Re}\ l(F_i)&=\frac{1}{n}\sum_{t=1}^{n}I(F_i,F_t)\\&=\frac{1}{n}\left(H(F_i)+\sum_{1\leqslant t\leqslant n,t\neq i}I(F_i,F_t)\right)\end{aligned} \tag{5}$$

定义 2(条件相关度)　一个特征 G_t 对特征 F_i 的条件相关度的定义如式(6)所示：

$$\mathrm{Re}\ l(G_t|F_i)=\frac{H(G_t|F_i)}{H(G_t)}\mathrm{Re}\ l(G_t) \tag{6}$$

其中，G_t 属于已选择的特征集合 F。相关度和条件相关度之差即可定义为冗余度。

定义 3(冗余度)　一个特征 F_i 与已选特征 G_t 的冗余度如式(7)所示：

$$\mathrm{Re}\ d(F_i,G_t)=\mathrm{Re}\ l(G_t)-\mathrm{Re}\ l(G_t|F_i) \tag{7}$$

根据定义 1－定义 3 即可得到互信息中最大相关-最小冗余的特征评价标准，如式(8)所示：

$$UmRMR(F_i)=\mathrm{Re}\ l(F_i)-\max_{G_t\in F_{m-1}}\{\mathrm{Re}\ d(F_i,G_t)\}$$

或者

$$UmRMR(F_i)=\mathrm{Re}\ l(F_i)-\frac{1}{m-1}\sum_{G_t\in F_{m-1}}\mathrm{Re}\ d(F_i,G_t) \tag{8}$$

根据所得到的特征候选集以及每一个特征的取值情况，并结合式(2)－式(4)，即可得到在本文的实验环境下式(8)中各项的计算方法，分别如式(9)－式(12)所示。

$$\begin{aligned}H(F_i)&=-\sum_{j=1}^{2}p(F_{i_j})\log p(F_{i_j})\\&=-p(f_i)\log p(f_i)-p(\overline{f_i})\log p(\overline{f_i})\end{aligned} \tag{9}$$

$$I(F_i,F_t)=H(F_i)-H(F_i|F_t) \tag{10}$$

其中，

$$\begin{aligned}H(F_i|F_t)&=\sum_{j=1}^{2}p(F_{t_j})H(F_i|F_{t_j})\\&=p(f_t)H(F_i|f_t)+p(\overline{f_t})H(F_i|\overline{f_t})\\&=p(f_t)H(F_i|f_t)+p(\overline{f_t})H(F_i|\overline{f_t})\end{aligned} \tag{11}$$

其中，

$$H(F_i|f_t)=-p(f_i|f_t)\log p(f_i|f_t)-p(\overline{f_i}|f_t)\log p(\overline{f_i}|f_t) \tag{12}$$

计算 $\mathrm{Re}\ l(G_t)$、$H(G_t)$ 等与 G_t 有关的项的方法是一样的，只不过在求和时，G_t 所作用的集合是已选特征集合 F 以及其中的特征元素。

(4)聚类：每一种协议都具有不同类型的消息，该模块的目的是将相同类型的消息聚在一起。本文使用 weka 中自带的聚类算法 k means 来对数据帧进行聚类，该算法采用距离作为相似性的评价指标，即认为 2 个对象的距离越近，其相似度就越大。对于 2 个向量表示的数据帧 X,Y，其中 $X=(x_1,x_2\cdots x_m)$，$Y=(y_1,y_2\cdots y_m)$，两者之间的距离为：

$$D(X,Y)=\left(\sum_{i=1}^{m}(x_i-y_i)^2\right)^{\frac{1}{2}}。$$

3　系统框架

本文模型的输入是真实网络中特定的二进制协议数据帧，每一种协议都有不同类型的消息，通过 *n-gram* 的生成、筛选和拼接，特征选择算法和聚类算法的使用，可以将不同类型的消息区分开来。该方法的流程如图 3 所示。

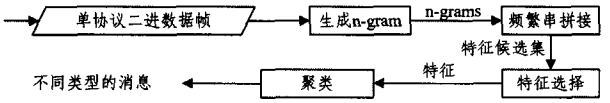

图 3　系统结构流程

3.1　*n-gram* 的生成

该模块的输入是在双方通信中所使用的单协议数据帧，输出是 *n-gram* 的集合，通过对 *n-gram* 进行拼接可以得到特征候选集。

首先对输入的数据帧，根据齐夫分布，选择曲线图中最接近直线的 n 值作为所求，例如在图 1 中，$n=1$。然后为了防止拼接过程中产生过多的冗余串，本文利用 *Jaccard* 参数对非频繁字节进行过滤，*Jaccard* 参数的计算公式如式(1)所示，在计算阈值之前需要对原始的数据帧进行预处理：将原始数据帧随机地平均分成 A,B 两部分，根据所选取的 n 值，对数据帧进行切分并排序，由此可以得到：

$$A=\{T1_1:F1_1,\cdots,T1_i:F1_i,\cdots,T1_n:F1_n\}$$

$$B=\{T2_1:F2_1,\cdots,T2_i:F2_i,\cdots,T2_n:F2_n\}$$

其中，$T1_i$ 和 $F1_i$ 分别表示 A 中的第 i 个字节及其在 A 中出现的频率，$T2_i$ 和 $F2_i$ 分别表示 B 中的第 i 个字节及其在 B 中出现的频率。最后通过阈值进行过滤得到所需要的 *n-gram* 集合。

3.2　特征候选集的拼接

该模块的输入是 3.1 节中产生的 *n-gram* 集合，输出是由拼接产生的长串所组成的特征候选集。

特征候选集的拼接依据是 *n-gram* 在数据帧中是否连续

出现，如果有 2 个或 2 个以上的 *n-gram* 在一帧数据中是连续出现的，就将它们拼接在一起形成频繁长串，作为下一步特征选择的输入。例如筛选出来的 1-*gram* 集合为{00，01，41，10}，对于数据帧片段“00 01 4e 45 46 41 41”，根据拼接原则就可以得到特征候选集 0001 和 4141。

3.3 特征选择

假设 3.2 节中所产生的特征候选集为 $U=\{F_1,F_2,\cdots,F_n\}$，在特征选择中将集合 U 作为输入，根据最大相关-最小冗余的特征选择标准，得到能更好地表征该协议不同类型消息的特征，以此作为特征向量用于聚类中。

在实际的实验中，假设特征候选集为 $U=\{F_1,F_2,\cdots,F_n\}$，实验中选择出来的特征用集合 F 表示，其中含有 m 个特征。

在选择第一个特征时，只需考虑特征之间的最大相关性即可，即取使式(5)中 $\mathrm{Re}\,l(F_i)$ 取得最大值的特征。在选择第 m 个重要特征时，除了要考虑最大相关之外，还要考虑要选择的特征是否与集合 F 中的特征达到了最小程度的冗余，即取使式(8)得到最大值的特征。

3.4 聚类

为了把不同的消息甄别开来，本文采用聚类算法将具有相同格式的消息聚在一起。在聚类之前要将所有的数据帧通过特征向量进行量化，假设 3.3 节中经过特征选择得到的集合为 $F=\{F_1,F_2,\cdots,F_m\}$，对于每一个数据帧 X，量化的结果为 $X=(x_1,x_2,\cdots,x_m)$，x_i 为 1 或者 0，分别表示特征 F_i 在数据帧 X 中是否出现。然后采用 weka 中自带的 k means 算法来实现聚类，根据前提假设，双方通信中所使用的协议只有 M 种消息，在进行聚类时，将 k 的值取为 M，算法运行结束得到的 M 个簇即为该模块的输出。

4 实验结果和分析

4.1 实验条件

本文针对双方简单通信的场景，提出了一种基于特征串的单协议分类方法。其中数据帧的 *n-gram* 切分、拼接以及特征选择算法都是通过 JAVA 代码实现的，聚类算法是通过 weka 中自带的 k means 算法实现的，运行环境为 WIN7 操作系统，Pentium(R) Dual-Core CPU E5800，2GB 内存。

4.2 实验数据集

本次实验中所使用的数据集是来自林肯实验室的 DARPA 数据集。数据集分为 inside. tcpdump 和 outside. tcpdump 2 部分。本文选用 inside. tcpdump 作为实验数据集，其中共包含 1130829 条数据帧，提取其中的 714 条 ICMP 协议进行实验，该协议含有 2 种类型的报文：请求应答报文和差错传播报文。

4.3 评价指标

为了评价该模型的性能，本文采用精确率(precision)和召回率(recall)2 种评价指标，在给出这 2 种指标的定义之前，先定义 3 种集合。

(1)True Positive(TP)：在类 X 中，真正属于 X 的数据帧所组成的集合；

(2)False Negative(FN)：本身属于类 X 的数据帧错分给其它类，这些数据帧所组成的集合；

(3)False Positive(FP)：在类 X 中不属于 X 的数据帧所组成的集合。

在此基础上，准确率(precision)和召回率(recall)的定义如式(13)和式(14)所示。

(1)精确率(Precision)：即类 X 中正确分类的成员和总成员的比例：

$$Precision=\frac{TP}{TP+FP}\times 100\% \tag{13}$$

(2)召回率(Recall)：即类别 X 的成员被正确分给 X 的比例：

$$Recall=\frac{TP}{TP+FN}\times 100\% \tag{14}$$

4.4 实验结果

首先根据齐夫分布曲线(见图 1)和 *Jaccard* 参数与阈值的关系曲线(见图 2)，得到 ICMP 的频繁字节为：“00”，“41”，“10”，“01”，“45”，“ac”，“08”，“03”，“46”，“7b”，“33”，“38”共 12 个。

其次对得到的 12 个字节，首先对一些不必要的串进行删除，将特征候选集的维度从 133 降低到了 25。然后基于特征选择算法得到的前 5 个、前 10 个和前 11 个特征，如表 1 所列。11 个特征的含义如下：0303 是差错传播报文的标志位，0800 是请求报文的标志位，00107b384633 是请求报文的物理地址，0800450000 是请求应答报文中的特征串，其中的 0800 表示该协议类型是 IP 协议，45 表示报头的长度为 20 个字节，00000000000000000000000000000000000000000000 也是请求应答报文中的特征串，从左到右依次表示 ICMP 协议的验证符、序列号、10 个字节的数据以及最后的结束符。00000000450000 是差错传播报文中的特征串，其中 45 表示 ICMP 协议中所封装的 IP 协议的版本为 4，报头长为 20 个字节，00100001000000000000 和 1414141414141414141414141414141414141414141414141414141414141410000 都是差错传播报文中的特征串，表示 ICMP 协议中封装的 nbns 协议的会话服务和名字服务。由此可见，本文中采用的基于互信息的特征选择算法能够找到协议中各类消息的特征串。

表 1 不同个数的特征分布

前 5 个特征	00107b384633080045,0001,00000000450000,0010,0303
前 10 个特征	00107b384633080045,0001,00000000450000,0010,0303,00100001000000000000,41414141414141414141414141414141414141414141414141414141414141410000,0800450000,00000000000000000000000000000000000000000000000000000000,00107b3846330800450000
前 11 个特征	00107b384633080045,0001,00000000450000,0010,0303,00100001000000000000,41414141414141414141414141414141414141414141414141414141414141410000,0800450000,00000000000000000000000000000000000000000000000000000000,00107b3846330800450000,0800

最后根据所得到的特征对原始的数据帧进行聚类，在实验过程中考虑到特征个数对聚类准确率的影响，分别将特征个数取为 5 个、10 个，以及 11 个进行聚类实验，实验结果分别如表 2－表 4 所列，为了形象地表示不同的特征个数对聚类效果的影响，本文将 cluster0 和 cluster1 在不同特征个数下的性能指标表示成了图 4 和图 5。

表中的 cluster0 和 cluster1 分别表示请求应答报文和差错传播报文。

表 2　前 5 个特征的实验结果

	TP	FP	FN	Precision	Recall
cluster0	261	19	0	0.932	1
cluster1	434	0	19	1	0.958

表 3　前 10 个特征的实验结果

	TP	FP	FN	Precision	Recall
cluster0	261	35	0	0.882	1
cluster1	418	0	35	1	0.923

表 4　前 11 个特征的实验结果

	TP	FP	FN	Precision	Recall
cluster0	261	11	0	0.960	1
cluster1	442	0	11	1	0.976

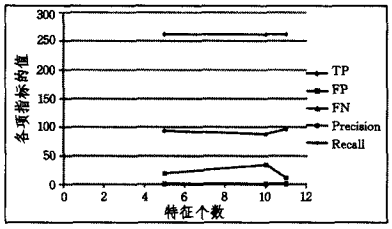

图 4　cluster0 的各项指标

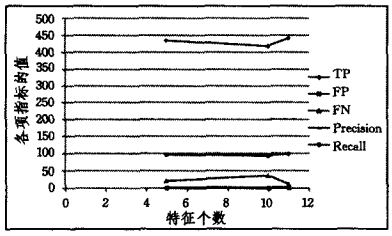

图 5　cluster1 的各项指标

本文在协议识别过程中利用互信息选择的特征串作为协议帧的聚类特征，将来自 DARPA 数据集中的 714 条 ICMP 数据帧作为数据源进行实验，数据源含有 2 种类型的报文，分别为差错传播报文和请求应答报文。

通过分析 4.4 节的实验结果可知，特征个数为 5 时聚类准确率和召回率已经达到 90%以上，而且与真实数据帧比较已经找到了真正的特征串。当特征个数增加到 10 时，聚类准确率反而有所下降，这是特征中的一些串对于聚类效果的干扰造成的，可见特征的个数并不是越多越好。虽然特征个数为 11 个时聚类准确率最高，达到了 96%，但是考虑到特征数量越多，在处理大数据量时所消耗的资源更多，本文选择特征个数为 5 时的特征集合作为识别 ICMP 帧数据的特征。

结束语　本文提出了一种利用关键字对未知二进制协议进行帧识别的系统模型，该模型可以自动地将输入数据帧按照不同的类型进行聚类，并采用无监督的互信息的特征选择方法，将 *n-gram* 和现有的熵值计算结合在一起。通过对 ICMP 进行实验，消息识别的准确率和召回率均可达到 90%以上。下一步对协议识别的研究可以考虑加入更多的信息，比如关键字的位置信息等，这样会使协议识别获得更精确的效果，并进一步减少冗余特征。

参 考 文 献

[1] 牟乔. 准确高效的应用层协议分析识别方法[J]. 计算机工程与程序，2010，32(8)：39-45
Mou Qiao. A Suite of Precise and Effcient Analyzing Techniques for Application Protocols[J]. Computer Engineering and Science，2010，32(8)：39-45

[2] IANA [OL]. http://www.iana.org/assignments/port—numbers

[3] Liu R T，Huang N F，Chen C H，et al. A fast string-matching algorithm for network processor-based intrusion detection system [J]. ACM Transactions on Embedded Computing Systems，2004，3(3)：614-633

[4] IANA. Internet Assigned Numbers Authority [OL]. http://www.iana.org/assignments/port-numbers

[5] Kim M S，Won Y J，Hong J W K. Application-level traffic monitoring and an analysis on IP networks[J]. ETRI Journal，2005，27(1)：22-42

[6] Chen C C，Wang S D. An efficient multicharacter transition string-matching engine based on the Aho-Corasick Algorithm [J]. ACM transactions on architecture and code optimization，2013，10(4)：1-22

[7] Wright C，Monrose F，Masson G M. HMM profiles for network traffic classification[C]// Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security. New York，USA，ACM，2004：9-15

[8] Wright C，Monrose F，Masson G M. Towards better protocol identification using profile HMMs：JHU-SPAR051201 [R]. 2005：325-328

[9] Bernaille L，Teixera R，Akodkenou I，et al. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review，2006，36(2)：23-26

[10] Zander S，Nguyen T，Armitage G. Self-learning IP traffic classification based on statistical flow characteristics[M]// Passive and Active Network Measurement. Heidelberg，Germany：Springer，2005

[11] Peltola H，Tarhio J. String matching with lookahead [J]. Discrete applied mathematics，2014，163(1)：352-360

[12] Giaquinta E，Fredriksson K，Grabowski S，et al. Motif matching using gapped patterns [J]. Theoretical Computer Science，2014，548：1-13

[13] Deyoung M E. Dynamic protocol reverse engineering：a grammatical inference approach [D]. Air Force Institute，2008

[14] Nohl K，Evans D，Starbug S，et al. Reverse-Engineering a Cryptographic RFID Tag[C]// USENIX Security Symposium. San Jose，California，USA，2008：185-194

[15] Wang Y，Zhang N，Wu Y，et al. Protocol Specification Inference Based on Keywords Identification[M]// Advanced Data Mining and Applications. Springer Berlin Heidelberg，2013：443-454

[16] Sen S，Spatscheck O，Wang D. Accurate，scalable in-network identification of p2p traffic using application signatures[C]// Proceedings of the 13th international conference on World Wide Web. New York，USA，ACM，2004：512-521

[17] Wang Y，et al. A semantics aware approach to automated reverse engineering unknown protocols[C]// ICNP 2012：20th IEEE International Conference on Network Protocols. Austin，TX，USA，IEEE，2012：1-10

[18] Kang H J，Kim M S，Hong J W K. A method on multimedia service traffic monitoring and analysis [M]// Self-Managing Distributed Systems. Heidelberg，Germany：Springer，2003

[19] Van Der Merwe J，Caceres R，et al. Mmdump：A tool for monitoring Internet multimedia traffic[J]. ACM SIGCOMM Computer Communication Review，2000，30(5)：48-59