doi:10.3969/j.issn.1000-436x.2013.10.016

基于主动学习和 SVM 方法的网络协议识别技术

王一鹏[1,2,3]，云晓春[1,3]，张永铮[3]，李书豪[3]
(1. 中国科学院 计算技术研究所，北京 100190；2. 中国科学院大学，北京 100049；3. 中国科学院 信息工程研究所，北京 100093)

摘 要：针对未知网络协议数据流的获取与标记工作主要依赖于领域专家。然而，样本数据量的增加会导致人工成本超过实际负荷。提出了一种新颖的未知网络协议识别方法。该方法基于主动学习算法，仅依靠原始网络数据流的载荷部分实现对未知网络协议的有效识别。实验结果表明，采用该方法设计的识别系统在保证识别准确率和召回率的前提下，能够有效地降低学习过程中标记的样本数目，更适用于实际的网络应用环境。
关键词：网络安全；网络协议识别；主动学习；网络数据流；支持向量机
中图分类号：TP393 **文献标识码：**B **文章编号：**1000-436X(2013)10-0135-08

Network protocol identification based on active learning and SVM algorithm

WANG Yi-peng[1, 2, 3], YUN Xiao-chun[1, 3], ZHANG Yong-zheng[3], LI Shu-hao[3]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Obtaining qualified training data for protocol identif ion generally requires domain experts to be involved, which is time-consuming and laborious. A novel approach for network protocol identification based on active learning and SVM algorithm was proposed. The experimental evaluations on real-world network traces show this approach can accurately and efficiently classify the target network protocol from mixed Internet traffic, and meanwhile display a significant reduction in the number of labeled samples. Therefore, this approach can be employed as an auxiliary tool for analyzing unknown protocols in real-world environment.
Key words: network security; protocol identification; active learning; network traces; support vector machine

1 引言

识别网络数据流中所承载的应用协议在网络与安全领域有着众多应用，例如入侵检测和防范系统(IDS/IPS)、网络测量、面向应用的缓存和路由机制、面向应用感知的负载均衡、流量分类和隧道检测等。以其在入侵检测和防范系统中的应用为例，入侵检测和防范系统通常依照已有的协议规范，通过对数据分组载荷部分的有效解析从而实现积极、有效的安全防护策略。然而，互联网中许多网络协议属于未知协议或者私有协议，这些网络协议没有公开可得到的协议规范文档，这给网络协议分类与识别带来新的挑战。根据 Internet2 NetFlow 组织对骨干网中流量的统计发现：超过 40%的网络数据流属于未知的应用协议[1]，其中恶意代码流量占有相当的比例。同时，传统以端口分配规则(IANA[2]规范)判定协议类别的流量分类方法也面临着诸多新的问题。例如，互联网中大量涌现的 Peer-to-

收稿日期：2013-04-29；**修回日期：**2013-07-18
基金项目：国家高技术研究发展计划(“863”计划)基金资助项目(2012AA012803, 2013AA014703)；国家科技支撑计划基金资助项目(2012BAH46B02)；国家自然科学基金资助项目(61303261, 61303170)
Foundation Item: The National High Technology Research and Development Program of China (863 Program) (2012AA012803, 2013AA014703); The National Science and Technology Support Program (2012BAH46B02); The National Natural Science Foundation of China (61303261, 61303170)

Peer(P2P)协议因其在服务质量(QoS[3])上的巨大优势，在文件分享和在线流媒体等领域中取得了广泛应用。然而，大多数 P2P 应用协议并不遵守 IANA 规范，通常采用动态端口等技术进行伪装，从而逃避网络服务提供者(ISP)的检测。当面临大量未知流量时，传统的检测方法或手段很难对相关未知应用协议做出正确识别。针对上述问题，设计合理、有效的未知网络协议识别方法给网络信息安全研究人员带来了新的挑战。

网络协议识别方法根据其研究对象的不同可划分为基于传输层端口、基于数据分组载荷[6~14]和基于网络流行为[4,5]3 种类别。目前，基于数据分组载荷的分析方法主要通过基于主机端的协议解析[6~9]和基于网络端的协议指纹[10~14]2 种方式构建所分析协议的分类特征。其中，基于协议指纹的分析方法又可划分为人工分析和自动分析 2 种。人工分析方法依照经验或先验知识获取协议指纹信息，这种分析过程通常耗时、费力。自动化的分析方法应用模式识别、机器学习等理论对网络数据流中的协议指纹信息进行自动提取，从而最大可能地减少人工成本开销。本文仅针对自动化的协议指纹提取工作展开相关讨论。

传统的网络协议识别方法大多属于非主动学习的机器学习方法。这类方法依照所获得的离线学习样本构建单一或者多种协议分类模型，从而实现对网络协议的准确识别。这类方法实验效果的优劣均依赖所分析的训练样本集合。然而，在实际分析过程中，未知协议网络数据流(如僵尸网络)的获取与标记工作严重依赖领域专家。这是一件费时且繁杂的工作。甚至在样本数据量过大时，人工标记已无法满足实际需求。因此，在复杂的网络环境中如何以最小的样本标记代价构建准确的协议识别模型，是目前网络协议识别领域的研究热点。

针对上述问题，本文提出了一种基于主动学习的未知网络协议识别方法，基于该方法设计并实现了 ProLearner 系统。该方法以网络数据流为输入，自动地从混杂网络流量中对所分析协议的网络数据流进行准确识别。该方法只分析 TCP/UDP 数据分组的载荷部分，不需要对程序的可执行代码进行逆向分析，也不依赖协议规范中的先验知识(如分隔符等)。同时，该方法可解决面向连接协议(如 TCP)和面向无连接协议(如 UDP)的识别问题，并可适用于文本和二进制类协议的分析。该方法的主要特点是通过采用主动学习算法，在学习过程中只选择最有价值的样本训练分类器。这种抽样策略使得学习效率(样本标记时间、学习训练时间等)得到大幅度的提高。在实践过程中，通过对训练样本的合理选取，在样本标记代价很小的前提下，同样可以保证很高的识别准确率和召回率。

2 相关研究工作

本文属于基于数据分组载荷的研究方法。下面对此部分相关工作进行介绍。

2.1 协议解析

2007 年，CABALLERO 等人[6]通过分析应用程序的可执行代码和数据分组载荷部分信息，采用协议逆向工程中的动态分析方法对协议的信息格式进行提取。2008 年，LIN 等人[7]和 WONDRACEK 等人[8]通过对可执行程序处理协议报文的工作流程进行观察分析，构建分析工具从而实现对协议格式信息的自动提取。2008 年，CUI 等人[9]提出并设计了 Tupni 系统，该系统利用逆向工程分析方法对输入数据流中的诸如记录序列、记录类别等格式信息自动地提取，从而实现有效的协议解析。

与以上工作不同，本文方法不需要对可执行程序进行逆向分析。

2.2 协议指纹提取

2005 年，HAFFNER 等人[10]提出了一种自动化的协议指纹构建方法——ACAS。该方法以所分析协议网络数据流中前 64byte 作为协议特征，应用机器学习算法构建协议分类模型。2006 年，KANNAN 等人[11]提出了一种基于泊松过程的协议格式特征挖掘方法。该方法是一个半自动化的方法，可有效识别属于同一会话的多个连接；但该方法只针对 TCP 数据流进行分析，并且利用了 TCP 数据分组中的 SYN、FIN 和 RST 等标志作为先验条件。2006 年，MA 等人[12]提出了一种基于无监督学习的协议推理方法，该方法通过对网络流的载荷部分进行分析，利用聚类方法实现对网络协议的自动识别。2007 年，CUI 等人[13]提出并设计了 Discoverer 系统。CUI 的方法利用统计学习和数据挖掘的研究方法，自动地从特定协议的网络数据流中提取该协议的格式信息。Discoverer 系统在分析文本类协议过程中需要依赖经验分隔符，在一定程度上限制了该系统的通用性。2010 年，FIANAMORE 等人提出了一个基于数据分组载荷特征的协议分类系统——

KISS[14]。该系统基于假设检验和机器学习理论，通过构建统计化的协议指纹信息实现对 UDP 网络数据流准确、快速的识别。该方法只针对 UDP 数据流分析建模，在适用范围上受到了一定的限制。

然而，上述研究均未涉及应用主动学习方法以减少在学习过程中样本的标记代价，从而实现协议分类模型的构建。

3　离线学习

本文提出并设计实现了一个基于主动学习的未知网络协议识别系统——ProLearner。如图 1 所示，ProLearner 系统由两阶段构成：离线学习和在线识别。离线学习是 ProLearner 系统的第一阶段。在这个阶段，通过对离线训练样本进行学习进而构建协议识别模型。离线学习阶段首先对网络数据分组的载荷部分进行数据分组建模，依照建模后得到的数据分组特征向量，采用面向 SVM 的主动学习算法构建协议分类模型。

3.1　数据分组建模

本文使用自然语言处理中的 *n*-gram 模型实现对网络数据分组的抽象建模。*n*-gram 模型可以将原始网络数据分组中的字节序列映射到新的特征维度空间。*n*-gram 模型的主要优点是：在并不依赖任何先验知识(如分割符等)的前提下，可以将任意的网络数据分组表示为 *n*-gram 序列的形式。因此，*n*-gram 模型适用于文本类和二进制类的协议分析。数据分组建模部分由 2 个部分组成，分别是数据分组 *n*-gram 序列化和数据分组向量化。

3.1.1　数据分组 *n*-gram 序列化

数据分组 *n*-gram 序列化利用计算语言学和概率论中的 *n*-gram 模型对所分析的网络数据分组进行分析建模。数据分组 *n*-gram 序列化操作利用 *n*-gram 模型将网络数据分组转化为以 *n*-gram 元素为基本单元的网络数据分组。

每个含有特定信息的网络数据分组都含有一个或多个协议关键字。网络协议识别的目标就是从数据分组中提取出能有效对网络协议合理区分的协议关键字。从本质上说，协议关键字是任意长度的字节序列。例如，“250”和“OK”是 SMTP 的协议关键字。对于文本类协议，通常可以利用经验分割符(如空格符和制表符)对协议中的关键字进行有效划分。然而，除了要处理文本类协议(如 SMTP 和 HTTP)，本文方法同样也需要处理二进制类协议(如 SMB 和 RTP)。因此，在实践中较优的选择是不依赖经验知识，而将数据分组内容看作若干相同长度且由基本元素构成的组合序列。这样的分析方法对文本和二进制类协议都适用，从而可以大大提高系统在实际中的可扩展性和通用性。

在自然语言处理中，*n*-gram 模型已成功应用于解决相似问题，因此本文利用 *n*-gram 模型来对网络数据分组的载荷部分进行建模分析。在计算语言学与概率学领域中，*n*-gram 是给定序列的(至少为 *n* 个元素的序列) *n* 个连续元素的子序列。例如，如果将每个字节视为一个元素，那么 SMTP 协议报文——“MAIL_FROM”所产生的 3-gram 元素是：“MAI”、“AIL”、“IL_”、“L_F”、“_FR”、“FRO”和”ROM”。具体而言，给定一个网络数据分组，*n*-gram 产生模块将字节大小为 *m* 的网络数据分组序列 $b_1b_2\cdots b_m$ 分解为 *n*-gram($n\leq m$)序列：$b_1b_2\cdots b_n b_2b_3\cdots b_{n+1}\cdots b_{m-n+1}b_{m-n+2}\cdots b_m$。基于 *n*-gram 的序列化操作，网络数据分组中的每个字节都被映射到新的维度空间。值得注意的是，在 *n*-gram 分析中通常应尽量避免较大的 *n* 值，因为较大的 *n* 值会使得 *n*-gram 集合中元素的数目过于庞大，在实践中需要使用大量的训练数据集合才可以避免状态空间过于稀疏。

n 值的选择是 *n*-gram 序列化的一个关键问题，该值的选择应尽量保证系统对不同协议的分析效果都较优秀。通过对多种协议的实验分析发现，*n*-gram 元素的出现频率与其在频率表中的排名是在以 log-log 为刻度的坐标轴下近似的一条直线，如图 2 所示。其中，R^2 代表拟合系数，其越接近 1，表示拟合效果越好。这种现象满足了自然语言中齐夫定律(Zipf's law)[15]的分布特性，因此认为可以利

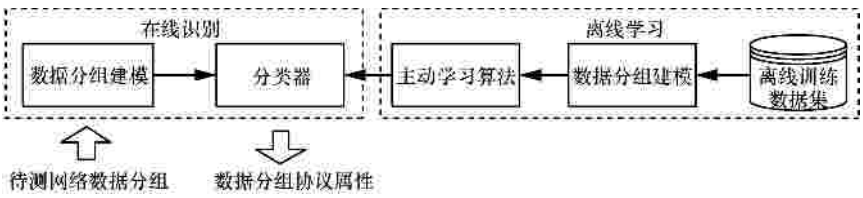

图 1　ProLearner 系统架构

用齐夫定律来近似地找到使得语言模型为最佳的拟合参数。对于网络协议数据分组的前16 byte，在n为不同取值的情况下，进行实验对比分析发现，当n=3时的拟合效果最好。

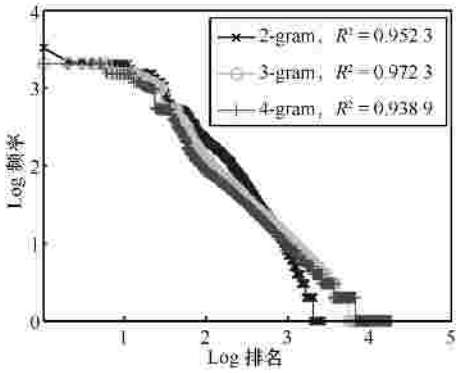

图2 SMTP协议中n-gram元素的概率分布

3.1.2 数据分组向量化

通过n-gram分析，数据分组中的字节序列被映射为n-gram序列。然而以n-gram序列形式描述的数据分组无法直接进行后续的运算处理。为解决这一问题，作者使用数据分组向量化操作对数据分组中的n-gram进行描述。通过该操作，每个数据分组都被表示为一个特征向量，特征向量的每个维度是唯一的n-gram元素，特征向量的分量数值代表n-gram元素在所分析数据分组中出现的次数。例如，假定有4种类型n-gram元素，分别为“HEL”、“ELO”、“DAT”和“ATA”。那么SMTP数据分组“DATA”向量化分析后的结果为(0, 0, 1, 1)。其中数值0代表所分析数据分组中没有该n-gram元素，数值1代表该n-gram元素出现1次。

3.2 主动学习方法

经过数据分组建模分析后，ProLearner得到数据分组的特征向量。依照这些分类特征，ProLearner利用面向SVM的主动学习算法进行训练并得到所分析协议的分类模型。下面将详细介绍本文所使用的面向SVM的主动学习方法。

3.2.1 主动学习方法概述

有监督机器学习(supervised machine learning)领域根据对学习样本处理方式的不同，可将分类模型划分为两大类别：被动学习模型和主动学习模型。被动学习模型随机地从训练数据中选取样本进行分类模型构建。然而，训练集本身经常包含许多信息量太少的样本，甚至可能是冗余或者噪音样本。这些样本的出现不仅使得样本标记工作大大增加、训练时间大幅度提高，而且还有可能导致分类器泛化能力的下降。与被动学习模型不同，主动学习模型采取主动的策略，选择最有利于提高分类器性能指标的训练样本，并只对这些样本进行标记。这样的选择策略有效地避免了学习模型本身对于重复的、无意义的样本的学习，使标记训练样本的代价大幅度降低，同时减少了训练过程的时间开销。

在主动学习中，主动抽样策略的出现是主动学习模型与传统被动学习模型最大的不同。根据抽样策略对未标记样本处理方式的不同，可将主动学习算法划分为：成员查询综合(membership query synthesis)算法[16]、基于流(stream-based)的主动学习[17]和基于池(pool-based)的主动学习[18]。其中，基于池的主动学习算法是目前研究最充分、使用最广泛的一类策略。按照选择未标记样本标准的不同，pool-based算法又可分为基于不确定性的抽样(UBS, uncertainty based sampling)策略、基于委员会投票的选择(QBC, query by committee)策略和基于估计误差缩减(EER, estimated error reduction)的抽样策略等几种策略[19]。本文采用了基于不确定性的抽样策略。

评价主动学习算法相对于被动学习算法性能的提升通常可以从2个角度进行考虑。一个角度是：在给定性能指标的前提下，主动学习方法相对于被动学习方法训练样本数量的精简比例。另一角度是：在给定训练样本数目的前提下，主动学习方法相对于被动学习方法性能的提升比例。在实际执行过程中，主动学习方法通常需要使用一定的终止策略。一方面可以控制样本标记的数目，从而减少总的执行时间；另一方面可以控制主动学习方法的学习效果，如准确率和召回率。

3.2.2 面向SVM的主动学习方法

由于SVM分类器在高维度和小样本数据分类中的突出表现，本文采用以SVM作为基准分类器的主动学习算法来构建协议分类模型。

给定未标记样本集合U的情况下，面向SVM的主动学习算法l主要由分类器f、查询函数q和已标记样本集合X 3部分组成[20]，如图3所示。主动学习算法是一个迭代训练分类器f的过程。其中，查询函数q是主动学习模型与被动学习模型最大的不同。

在面向SVM的主动学习算法中，查询函数q所采取的查询策略为：每次选取距离分类超平面最近的一个或几个样本，并将这个或这些样本提交给领域专家标记，这种查询策略也被称为不确定抽样

(UBS)。这些新的已标记好的训练样本将被加入到已标记样本集合 X 中，用于重新对分类器 f 的训练。不确定抽样选择策略总是选择不确定性最大的样本进行学习，因为这些样本最有可能是 SVM 分类器的支持向量。这样的选择策略可以有效避免 SVM 分类器对无意义、重复样本的学习，从而大大减少标注时间，提高学习效率。

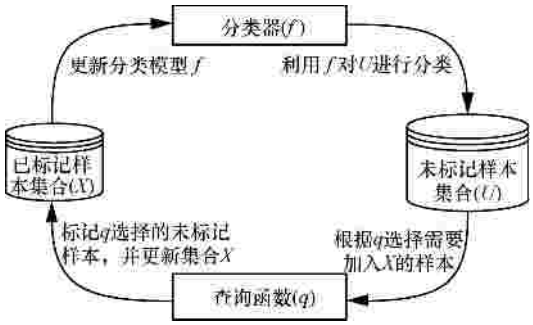

图 3　主动学习算法过程

主动学习的过程如图 3 所示，其具体算法过程如下。

输入：已标记训练样本集合 X 中少量已标记样本 x(至少包含一个正样本和一个负样本)，未标记测试样本集合 U 中(包含有正样本和负样本)大量未标记样本 u。r 为主动学习终止条件。

输出：分类器 f 和已标注样本集合 X。

1) 根据已标记训练样本集合 X 中少量已标记样本 x，训练构造 SVM 分类器 f，从而样本集合 X 被划分为 2 类，$f: X ? \{-1, 1\}$。

2) 利用已构造的 SVM 分类器 f 对 U 中所有未标记的样本 u 进行分类。

3) 根据分类器 f 的分类结果，查询函数 q 判定未标记样本集合 U 中每个样本的信息量，将信息量最大、最不确定(距离超平面最近)的样本交由领域专家进行标记。

4) 将经由领域专家标记好的样本加入到已标记样本集合 X 中，根据更新后的已标记样本集合 X 对分类器模型 f 进行评估。

5) 若达到终止条件 r 时，则算法终止，返回分类器 f；否则重复步骤 1)~步骤 4)。

4　在线识别

在线识别阶段通过利用离线学习阶段得到的分类器实现对实时网络流量的协议判别。在线识别阶段的输入是待测网络数据分组。通过对数据分组载荷部分进行 n-gram 建模，得到该数据分组的特征向量。依照其分类特征和离线学习阶段训练得到的分类器 ProLearner 对所分析网络数据分组的协议属性进行判别。在线识别阶段的输出结果为 2 类：一类是属于目标协议的网络数据分组，另一类是非目标协议的网络数据分组。

5　实验结果和分析

为了验证 ProLearner 系统在实际网络环境中的有效性，作者使用真实的网络数据流量对该系统进行验证。在实验阶段，假定几种协议的网络数据流是未知的，从而模拟对未知网络协议识别的整个交互过程。

5.1　数据集

对于特定协议网络数据流的获取通常可以采用以下 2 种方法：1) 在可控环境下运行所分析协议的可执行程序代码，从而获得该协议的网络数据分组(例如 GT 方法[21])；2)对于使用非可变传输层端口的协议，在可控环境下进行端口监听的方式。2 种方法都可在已获得程序可执行代码的前提下，实现对未知协议网络流量的获取。然而，在已获得的网络数据流中，可能同时混杂有其他非所分析协议的网络数据流。因此，在这个环节中需要领域专家的参与，从而对训练样本进行有效、合理的区分。在本文中，作者采用第 2 种方法实现对数据集合的构建。

本文选取了 4 种协议对 ProLearner 系统的性能进行测试，分别是 SMTP、DNS、XUNLEI 和 CIFS/SMB 协议，这些协议中既包含有面向连接的协议(如 TCP)，也包含面向无连接的协议(如 UDP)，同时，也包含文本和二进制类协议。SMTP 协议通常用于电子邮件通信，CIFS/SMB 协议主要用来提供共享的文件访问策略。为了获得这 2 种协议的网络数据流，作者采用 TCP 端口过滤的方法进行数据采集，SMTP 端口号为 25，CIFS/SMB 端口号为 445。DNS 协议是域名系统，其主要用于将域名信息转化为数字化的网络地址，XUNLEI 协议是当今十分流行的 P2P 文件共享应用，其在国内骨干网流量中占有相当的比例。为了获得这 2 种协议的网络数据流，作者使用 UDP 端口过滤的方法对数据进行采集，DNS 端口号为 53，XUNLEI 端口号为 15 000。非目标协议的网络数据流(负样本)则通过在以上指定之外的端口来进行捕获。在每种协议的数据集合中随机提取 10000 条所分析协议的网络数据分组和 2 000 条非目标协议的数据分组。采用模 5(即 5-fold)

交叉验证的方法，重复实验10次取平均值。

5.2 评价指标体系

给定ProLearner系统要分析的某种未知协议，首先定义以下3种数据集合。

1) true positives (TP): 被ProLearner系统识别为某协议的网络数据分组，且确实是属于该协议的网络数据分组集合。

2) false positives (FP): 被ProLearner系统识别为某协议的网络数据分组，但并不属于该协议的网络数据分组集合。

3) false negatives (FN): 被ProLearner系统识别为非某协议的网络数据分组，但其实是属于该协议的网络数据分组集合。

基于以上3种数据集合，本文采用机器学习领域中通常使用的准确率(precision)、召回率(recall)和F-Measure 3种评价指标来对ProLearner系统的有效性和可靠性进行评价。3种评价指标定义如下。

$$precision=\frac{TP}{TP+TP}$$

$$recall=\frac{TP}{TP+FN}$$

$$F\text{-}Measure=2\times\frac{precision\cdot recall}{precision+recall}$$

由于准确率与召回率分别描述系统性能的2个方面，单一使用准确率和召回率作为评价指标具有局限性，因此，本文选用F-Measure指标将这2个指标进行综合考虑，从而选择最优方案。

5.3 实验结果

ProLearner系统的整个过程需要确定以下几个参数：

1) 每个数据分组分析的字节数l；

2) SVM模型中核函数的参数C和$?$；

3) 集合X中已标记样本个数的初始值T。

5.3.1 每个数据分组分析的字节数l

本文以网络数据分组的载荷部分作为研究对象。在实验分析过程中，ProLearner系统选取每个数据分组的前l byte进行分析。通常而言，分析过程中l选取的越大，其越能准确地反映出所分析协议载荷部分的特征，从而使得ProLearner系统在识别过程中的分类准确率越高。然而，分类方法的计算复杂性和内存开销同样也随着l的变大而相应增加。在验证实验中，作者选择了一个折中数值，设l为16。

5.3.2 SVM模型中的参数

本文SVM模型中所选用的核函数为径向基核函数(radial basis function)。径向基核函数可以将样本映射到一个更高维的空间，从而对复杂特征的样本集进行有效分类。径向基核函数中有2个可调节参数分别为C和$?$，其中，C为惩罚因子，$?$为核参数。通过交叉验证进行对比发现，模型中参数的最佳取值为C=1 000，$?$=0.5。

5.3.3 召回率和准确率

本节对数据集中的4种协议在集合X中已标记样本个数的初始值T为10、30和50这3种不同取值的情况下分别进行实验。在对所获得的网络协议数据集合进行分析时，传统方法通常采用随机抽样的方法来构建学习集合。因此，在验证对比实验中，作者分别对每种协议进行训练和测试，并比较其在主动学习策略和随机抽样方法下的准确率、召回率和F-Measure 3种评价指标。

DNS和XUNLEI 2种协议的实验结果如图4和图5所示。对于这2种协议，在T为不同取值的情况下，通过对比实验发现：当已标记样本集合的数目达到100时，主动学习方法和随机抽样方法的召回率均已达到约100%。然而，通过观察发现当已标记样本数量为100、150、200、250和300时，主动学习方法的准确率接近100%，而随机抽样策略的准确率大约在85%左右。主动学习方法的实验效果明显优于随机抽样策略。图4和图5同时表明，对于DNS和XUNLEI这2种协议，随着T值的增大，主动学习方法达到识别最佳实验效果的收敛速度将减慢。

图6为CIFS/SMB协议的实验结果。通过改变T值进行实验对比发现：当已标记样本集合的数量达到100时，主动学习和随机抽样方法2种策略的识别准确率基本达到一致。从图6中可以发现，主动学习方法的召回率已达到约100%，且不随T值的增大而改变。而随机抽样策略的召回率相对于主动学习方法则比较低，且随样本数量的增加而缓慢提高。对于CIFS/SMB协议，随机抽样策略的召回率在小样本的情况下与主动学习方法在实际效果上有一定的差距。

SMTP协议的实验结果如图7所示。通过改变T值进行对比实验发现：主动学习方法的F-Measure指标在大多数情况下均高于随机抽样策略。当T=50，已标记样本数量增加到300时，主动学习方

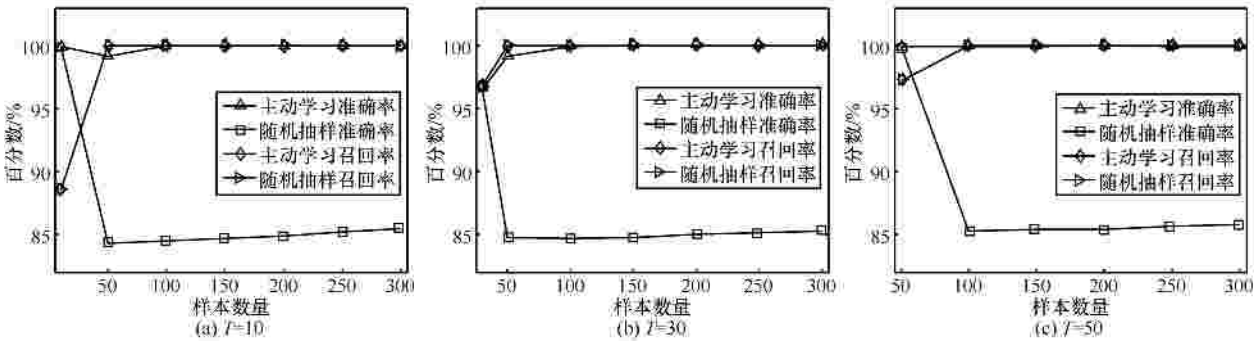

图 4　主动学习方法与随机采样方法在 DNS 中的对比实验

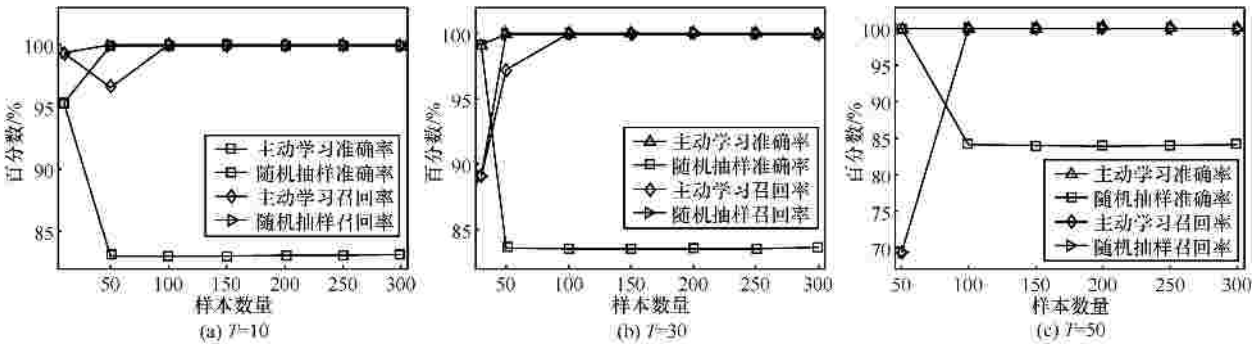

图 5　主动学习方法与随机采样方法在 XUNLEI 中的对比实验

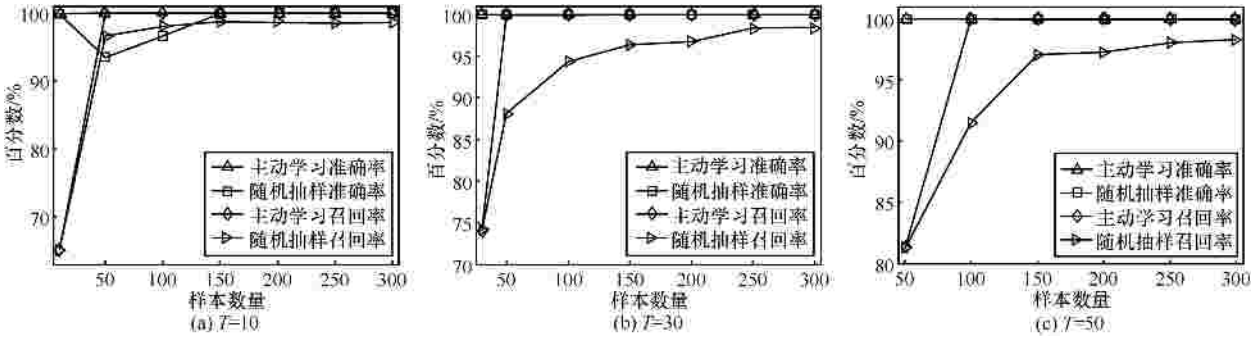

图 6　主动学习方法与随机采样方法在 CIFS/SMB 中的对比实验

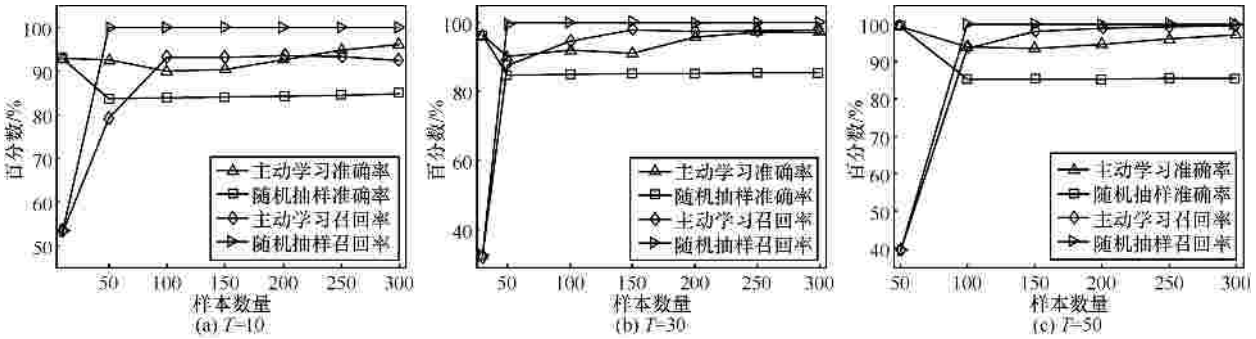

图 7　主动学习方法与随机采样方法在 SMTP 协议中的对比实验

法达到最佳的识别效果，其 F-Measure 指标为 0.98。此时随机抽样策略的 F-Measure 指标为 0.92。从 3 组对比实验可以发现，对于 SMTP 协议，需要将已标记样本集合中初始的样本数目适当增大，从而获得较优的识别效果。

5.4　总结与讨论

在实验验证阶段，作者使用包含文本和二进制的真实网络数据流对 ProLearner 系统的有效性进行

测试。多种协议的实验结果表明，在已标记样本量很少的情况下，ProLearner 系统能够从混杂的网络流量中准确地识别所分析的网络协议。在针对未知网络协议的识别过程中，相对于随机抽样策略，主动学习方法可以使用较少的已标记样本达到较优的学习效率(高准确率和高召回率)，从而有效地降低了学习过程中标记的样本数目。然而，本文的方法并不适用于解决加密网络数据流的识别问题，这也是本文方法的局限性。

6 结束语

随着网络流量的日益复杂多样，有效的未知网络协议识别方法成为信息安全领域一个重要的研究方向。由于未知协议网络数据流的获取与标记工作通常需要高昂的人工成本，因此需要研究可适用于实际网络环境的样本标记方法。本文从网络数据流出发，构建了一整套基于主动学习的未知网络协议识别方法。该方法仅依靠原始网络数据流实现对未知网络协议的有效识别。实验结果表明，本文方法对面向连接和面向无连接的协议均可取得较好的实验结果。在保证识别准确率和召回率的前提下，有效地降低了学习过程中所需的标记样本数量。因此，本文的方法可直接应用于实际的网络环境。

参考文献：

[1] Internet netflow statistics[EB/OL]. http://netflow.internet2.edu, 2010.

[2] TCP and UDP port numbers[EB/OL]. http://www.iana.org/assignments/port-numbers, 2008.

[3] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[A]. Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement[C]. Taormina, Sicily, Italy, 2004.135-148.

[4] ZHANG J, CHEN C, XIANG Y. An effective network traffic classification method with unknown flow detection[J]. IEEE Transactions on Network and Service Management, 2013, 10(1):1-15.

[5] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[J]. SIGCOMM Computer Communication Review, 2005, 35(4):229-240.

[6] CABALLERO J, YIN H, LIANG Z, *et al*. Polyglot: automatic extraction of protocol message format using dynamic binary analysis[A]. Proceedings of the 14th ACM Conference on Computer and Communications Security[C]. Virginia, USA, 2007.317-329.

[7] LIN Z, JIANG X, XU D, *et al*. Automatic protocol format reverse engineering through context-aware monitored execution[A]. Proceedings of the 15th Network and Distributed System Security Symposium[C]. California, USA, 2008.1-17.

[8] WONDRACEK G, MILANI P, KRUEGEL C, *et al*. Automatic network protocol analysis[A]. Proceedings of the 16th Network and Distributed System Security Symposium[C]. California, USA, 2008.1-18.

[9] CUI W, PEINADO M, CHEN K, *et al*. Tupni: automatic reverse engineering of input formats[A]. Proceedings of the 15th ACM Conference on Computer and Communications Security[C]. Virginia, USA, 2008. 391-402.

[10] HAFFNER P, SEN S, SPATSCHECK O, *et al*. ACAS: automated construction of application signatures[A]. Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data[C]. Pennsylvania, USA, 2005.197-202.

[11] KANNAN J, JUNG J, PAXSON V, *et al*. Semi-automated discovery of application session structure[A]. Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement[C]. New York, USA, 2006.119-132.

[12] MA J, LEVCHENKO K, KREIBICH C, *et al*. Unexpected means of protocol inference[A]. Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement[C]. Rio de Janeriro, Brazil, 2006. 313-326.

[13] CUI W, KANNAN J, WANG H J. Discoverer: automatic protocol reverse engineering from network traces[A]. Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium[C]. Boston, MA, 2007.313-326.

[14] FINAMORE A, MELLIA M, MEO M, *et al*. KISS: stochastic packet inspection classifier for UDP traffic[J]. IEEE/ACM Tra tions on Networking, 2010, 18(5):1505-1515.

[15] MANNING C, SCHUTZE H. Foundations of Statistical Natural Language Processing[M]. MIT Press, 1999.

[16] ANGLUIN D. Queries and concept learning[J]. Machine Learning, 1988, 2(4):319-342.

[17] COHN D, ATLAS L, LADNER R. Improving generalization with active learning[J]. Machine Learning, 1994, 15(2):201-221.

[18] LEWIS D, GALE W. A sequential algorithm for training text classifiers[A]. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. New York, USA, 1994.3-12.

[19] SETTLES B. Active learning literature survey[EB/OL]. http://research.cs.wisc.edu/techreports/2009/TR1648.pdf.

[20] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2:45-66.

[21] GRINGOLI F, SALGARELLI L, DUSI M, *et al*. GT: picking up the truth from the ground for internet traffic[J]. SIGCOMM Computer Communication Review, 2009, 39(5):12-18.

作者简介：

王一鹏（1985-），男，河北遵化人，中国科学院博士生，主要研究方向为网络信息安全、网络协议识别。

云晓春（1971-），男，黑龙江哈尔滨人，博士，中国科学院研究员、博士生导师，主要研究方向为网络信息安全。

张永铮[通信作者]（1978-），男，黑龙江哈尔滨人，博士，中国科学院副研究员、博士生导师，主要研究方向为网络安全态势感知。E-mail: zhangyongzheng@iie.ac.cn。

李书豪（1983-），男，山西文水人，博士，中国科学院助理研究员，主要研究方向为网络与信息安全、恶意代码分析与防范。