一种基于 AdaBoost-SVM 的流量分类方法*

张　震，汪斌强，梁宁宁，程国振
（国家数字交换系统工程技术研究中心，郑州 450002）

摘　要：针对传统分类方法的缺陷，提出了一种基于 AdaBoost-SVM 的流量方法。该方法利用 K-L 变换从大量冗余流量特征中遴选出少量本征特征，有效降低了算法的处理复杂度；应用 AdaBoost 机制将一次分类过程等分成若干层基于支持向量机的弱分类器，使得分类方法简单、易于实现；通过分层组合和迭代权重的方法聚焦在困难分类的数据样本上，提高了分类器的准确性能。理论分析和实验结果表明：在降低计算复杂度的同时，AdaBoost-SVM 算法的准确性能够达到 95%。

关键词：流量分类；K-L 变换；支持向量机；AdaBoost；弱分类器
中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1001-3695(2013)05-1481-05
doi:10.3969/j.issn.1001-3695.2013.05.051

Internet traffic classification based on AdaBoost-SVM

ZHANG Zhen, WANG Bin-qiang, LIANG Ning-ning, CHENG Guo-zhen
(*National Digital Switching System Engineering & Technology R&D Center, Zhengzhou* 450002, *China*)

Abstract: Aiming at the deficiencies of traditional classified methods, this paper presented a novel scheme called Internet traffic classification based on AdaBoost-SVM. Herein, this method selected a few intrinsical flow characteristics using K-L transform from a great deal redundant ones. In order to make the process easily implemented, AdaBoost equally partitioned the whole classification into several layers. It constructed one non-linear support vector machine in each layer. Through stratified combinations and iterative weights, the algorithm focused on hard-classified data to improve the classifier's performance. Theoretical analysis and experimental results show that the algorithm based on AdaBoost-SVM can achieve the accuracy of 95% and better computational performance compared with traditional K-means and NBC methods.

Key words: traffic classification; K-L transformation; support vector machine; AdaBoost; weak classifier

0　引言

随着网络技术的不断发展，互联网已经由单一的数据传送网络发展成为数据、语音、图像和实时多媒体信息的综合传输网络。特别是近年来，网络业务呈现规模化、差异化的发展趋势，对传统粗放式的网络运营和维护模式提出了严峻的挑战，具体表现为：

a）网络应用不断衍生出新的业务形态，搜索引擎、即时通信、网络音乐、网络新闻、个人空间等业务得到广泛应用，微博应用爆发，商务团购、快速支付等新型业务的使用率快速攀升[1]。

b）P2P 应用吞噬网络带宽。P2P 流量已经取代了 HTTP 流量成为 Internet 流量的主体[2]。

c）非法信息泛滥。一些病毒、攻击等不良的流量常常导致网络瘫痪，造成极大的损失。当前网络基础设施不时面临着 DDoS 攻击、蠕虫传播、僵尸网络等安全威胁。

d）加密流量比例不断增加。信息安全日益成为大众关注的热点，网络安全协议在实际的网络应用中发挥了重要作用：IPSec 广泛应用于 VPN；在网上银行、电子商务等领域更是可见传输层加密协议 SSL 的身影。

面对网络业务多元化的发展趋势，迫切需要对网络流量进行精确识别和分类。流量分类是指在基于 TCP/IP 的互联网中，按照网络的应用类型（如 FTP、P2P、HTTP 等），将网络通信产生的双向 TCP 流或 UDP 流进行分类[3]。目前，网络流量分类多以具有相同 5 元组（源 IP、目的 IP、源端口、目的端口、协议类型）的报文序列（即 IP 流）为基本处理单元，同一流中的报文归为同一种业务。

1　流量分类相关算法

基于端口的流量分类是最简单和最原始的方法，它是根据互联网地址指派机构（Internet Assigned Numbers Authority，IANA）规定的端口映射表，将特定端口的网络流量划分到相应的网络应用。然而，随着 P2P 和被动 FTP 等新型网络业务的日益流行，随机端口技术和端口跳变技术被用于数据传输，导致这种基于端口的流量分类方法被迅速淘汰[4]。

针对新兴业务动态绑定端口的特点，应用深度报文检测（deep packet inspect，DPI）已成为业界公认比较成熟的技术。文献[5]基于 DPI 技术构建了 P2P 流的识别基准集，并通过匹

收稿日期：2012-09-17；**修回日期**：2012-10-23　　**基金项目**：国家“863”计划资助项目（2011AA01A103）；国家“973”计划资助项目（2012CB312901，2012CB312905）

作者简介：张震（1985-），男，山东济宁人，博士研究生，主要研究方向为宽带信息网络、网络管理（zhangzhenhigh@gmail.com）；汪斌强（1963-），男，教授，博导，主要研究方向为宽带信息网络与高速路由技术；梁宁宁（1982-），女，博士，主要研究方向为网络体系结构；程国振（1986-），男，博士研究生，主要研究方向为网络管理.

配负载内容和负载偏移量,设计了准确性和实时性较好的 P2P 流量分类器。文献[6]对传统 DPI 分类器进行了优化,采用了早期行为特征的思想,即业务流的前几个报文包含了大量的协议通联信息,负载特征最早出现在前几个包中的概率非常大,仅需要对连接开始阶段的数据包进行特征匹配就可以进行流量分类。以上方法均属于 DPI 技术的应用范畴,但是随着骨干网链路速率的增加,分析完整的应用层负载不仅计算开销较大,而且还可能带来不必要的用户隐私纠纷。另外,面对应用层负载加密类业务或者内容特征尚未公布的新型业务流,DPI 对此无能为力。为了识别加密流量和新型业务,基于流量统计特征的分类方法通过提取网络流的统计特征(如平均报文长度、流的持续时间、平均报文间隔等),将网络流抽象为由一组统计特征值构成的属性向量,实现由流量分类向机器学习的转换,使得基于机器学习的流量分类成为该领域一个新兴的研究方向。

文献[7]使用经典的 K-means 方法进行流量分类,该方法的中心思想是找到 K 个聚簇中心,使得每个样本流到该聚类中心点的平方和最小。K-means 的优势是模型简单,计算过程相对高效。但是,K-means 流量分类方法对初始 K 个聚类中心的选择敏感,并需要事先确定参数 K;不适合发现非凸形状的簇和形状差别较大的簇,且对异常数据点敏感;在每个类中,若样本流分布不规范或者数据偏差较大时,分类的准确性会大大降低。

另外一种应用广泛的方法是基于贝叶斯技术的流量分类。文献[8]设计了一种基于概率模型的朴素贝叶斯分类器(naïve Bayesian classifier,NBC),该方法要求参与分类的各项属性特征相互独立而且遵循高斯分布。然而在流量分类问题中,原始的网络流属性集合很难满足上述条件,因此,该方法的整体准确率只有 65% 左右。为了克服 NBC 分类器的缺陷,Moore 等人[9]采用基于关联的快速过滤机制(fast correlation-based filter,FCBF)和核估计(kernel estimation,KE)技术对朴素贝叶斯方法进行了改进,改进后的平均分类准确率达到了 95% 左右。但是,该方法存在以下严重的效率问题:使用信息熵和对称不确定性作为属性特征的相关性度量,计算变量取值概率和条件概率的复杂度较高;使用核密度估计时,由于使用的训练样本有限,难以模拟未知空间样本,无法保证分类结果的稳定性。

立足于分类算法的准确性和计算复杂度,本文提出一种基于 AdaBoost-SVM 的流量分类机制(traffic classification based on AdaBoost with SVM components, AdaBoost-SVM)。本文概述了 AdaBoost 组合提升的方法思想,给出了构造弱分类器的详细流程,详细介绍了 AdaBoost-SVM 算法的详细流程、准确性分析以及复杂度分析,最后应用 Moore 数据集,实验验证了 AdaBoost-SVM 的性能,并与 K-means、NBC、NBC + KE + FCBF 等分类算法进行了仿真对比。

2 AdaBoost 方法

设 $X=(x_1,x_2,\cdots,x_N)$ 为模式空间 R^K 的一个有限数据集,其中 $x_i(i=1,2,\cdots,N)$ 是由 K 维属性构成的向量空间中的一个数据点,每个数据点对应一个类别标签 $y_j\in Y,j\in[1,L]$。所谓分类,就是要训练学习得到一个分类模型 $f:X\rightarrow Y$,并基于此模型对网络流进行实时分类。Boosting 算法(即组合提升算法)通过对若干分类器的组合可以使分类算法的性能得到有效提高,其中 AdaBoost 是 Boosting 家族最具有代表性的算法[10]。作为一种自适应的 Boosting 算法,AdaBoost 分类算法的基本思想是:将 N 个数据样本的分类过程等分成 T 层弱分类器的组合叠加,其中每层抽样处理固定数量的 M 个样本;自适应迭代训练样本的权重,使得弱分类器聚焦在那些困难的数据样本上;利用分类能力一般的弱分类器,通过一定的方法进行组合叠加,并最终生成一个强分类器。理论证明[10]:只要每个弱分类器的分类能力比随机猜测的好,则当弱分类器的个数趋向于无穷时,强分类器的错误率将趋于 0。

算法 1 给出了 AdaBoost 组合提升的详细步骤。其中 $\delta(l^t(x_i)\neq l^{t-1}(x_i))$ 表示:若 $l^t(x_i)\neq l^{t-1}(x_i)$ 成立,则其值等于 1,否则等于 0。通过算法 1 可以看出,每个被抽中数据点的权重会得到不断调整,若上次分类结果和本次结果不同,认为其是困难样本,则增大其下一次的抽样权重;反之,则减小其下一次的抽样权重。最后,通过加权组合投票方法得到每个数据点的业务标签。

算法 1 AdaBoost 组合提升流程

1 输入:训练数据集合 $(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)$,迭代层数 T,初始化数据点的抽样权重 $w_i^1=1/N(i=1,2,\cdots,N)$。

2 for $t=1$ to T do

3 根据 w_i^t,计算各个点的抽样概率 $p_i^t=w_i^t/\sum_i w_i^t$,并进行不等概率抽样得到 M 个样本点;

4 基于抽样得到的 M 个样本点构造弱分类器 h_t;

5 计算 h_t 的分类误差 $\varepsilon_t=[\sum_{i=1}^{M}p_i^t\delta(l^t(x_i)\neq l^{t-1}(x_i))]$,$l^t(x_i)$ 为第 t 步中样本点 x_i 的业务标签;

6 if $\varepsilon_t>0.5$ then

7 重置 $w_i=1/N(i=1,2,\cdots,N)$,返回步骤 3;

8 end if

9 令 $\alpha_t=\varepsilon_t/(1-\varepsilon_t)$,更新每个样本的抽样权重为 $w_i^{t+1}=w_i^t\alpha_t^{1-\delta(l^t(x_i)\neq l^{t-1}(x_i))}$;

10 end for

11 输出:每个样本对应的类别标签。计算表达式为 $l(x_i)=\arg\max_{y\in Y}\sum_{t=1}^{T}(\log 1/\alpha_t)\delta(l^t(x_i)=y)$。

3 弱分类器构造

AdaBoost-SVM 算法在每一层需要训练弱分类器。本文首先遴选出具有代表性的业务特征,然后结合支持向量机(SVM)创建弱分类器。

3.1 流量统计特征的遴选

在流量分类预处理阶段,特征遴选直接关系到后续分类的准确性和复杂度。文献[11]提供了 249 种业务流特征,包括端口号、包长、包到达时间、TCP 三次握手包数、往返时间、TCP 拥塞窗口大小等,其中有些特征关联性较强,含有大量冗余信息。为了有效降低特征维数,AdaBoost-SVM 算法利用 K-L 变换,通过计算训练样本矩阵的本征值来选取最优的特征。将每个数据点看做一个由 K 维属性特征组成的列向量 $\boldsymbol{x}_i$,则 $\boldsymbol{X}_{k\times N}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$ 为训练样本矩阵。K-L 变换的目标是生成 J $(J<K)$ 个互不相关的特征,其主要步骤如下:

a)计算自相关流量矩阵 $\boldsymbol{R}_x=E(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$。若给定 N 个样本向量,则自相关流量矩阵 $\boldsymbol{R}_x\approx 1/N\cdot\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$。

b)计算协方差流量矩阵 $\boldsymbol{\Sigma}_X=\boldsymbol{R}_x-E(\boldsymbol{X})E(\boldsymbol{X})^{\mathrm{T}}$。

c)计算协方差流量矩阵 $\boldsymbol{\Sigma}_X$ 的本征向量 $\boldsymbol{a}_i(i=0,1,\cdots,K-1)$ 和本征值 $\lambda_i(i=0,1,\cdots,K-1)$,使得 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{A}=\boldsymbol{\Lambda}$,其中矩阵

A 为本征向量组成的正交矩阵。

d)将本征值的秩以降序排列 $\lambda_0 \geqslant \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{K-1}$,选择前 J 个特征值和对应的特征向量。

基于以上步骤,将原有 K 维属性特征表示的模式空间 R^K 映射到 R^J 中,一方面降低了数据表示的维数,能够大大降低后续分类算法的计算复杂度;另一方面,剔除了无关流量特征的影响,能进一步提高算法的分类准确性。

3.2　SVM 弱分类器构造

SVM 建立在统计学习理论基础上,是一种针对线性和非线性数据的机器学习方法。它将训练数据非线性映射到较高的维度上,并在新的维度上搜索线性最优超平面,最终由该超平面决定样本的类别。由于 SVM 对复杂的非线性决策边界的建模能力是高度准确的,所以它已经成功应用于手写体识别、语音识别、图像分割等领域[12,13]。先考虑线性可分的两类问题:某区域内存在由 J 维属性构成的 N 个样本 $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$,其中类别标签 $y_j \in \{-1, +1\}$;存在超平面 $w^{\mathrm{T}} \cdot x + w_0 = 0$ 能将 N 个样本分为两类,其中 w 是权重向量,w_0 是标量。如图 1 所示,假设两个平行的侧面 $H_1: w^{\mathrm{T}} \cdot x + w_0 = 1$ 和侧面 $H_2: w^{\mathrm{T}} \cdot x + w_0 = -1$,最优超平面的边缘即为满足距离最大化的两个侧面 H_1 和 H_2。因此,SVM 的目标函数为最大化 H_1 和 H_2 的边缘距离 $2/\|w\|^2$,等同于最小化等式 $f(w) = \|w\|^2/2$。另外,所有属于类别 1 的数据 x_i 均满足 $w^{\mathrm{T}} \cdot x + w_0 \geqslant 1$;属于类别 -1 的数据 x_i 均满足 $w^{\mathrm{T}} \cdot x + w_0 \leqslant -1$。这两个不等式约束可以合并为 $y_i(w^{\mathrm{T}} \cdot x + w_0) \geqslant 1$。基于以上分析可得到 SVM 线性可分的形式化描述:

$$\text{最小化}: \|w\|^2/2 \tag{1}$$

$$\text{约束条件}: y_i(w^{\mathrm{T}} \cdot x + w_0) \geqslant 1 \quad i = 1, 2, \cdots, N \tag{2}$$

如图 2 所示,考虑 SVM 线性不可分的情况:

a)对于分离域以外的数据向量可以由式(2)正确处理;

b)在分离域内,正确分类的向量(如图 2 中黑色的点)满足不等式 $0 \leqslant y_i(w^{\mathrm{T}} \cdot x + w_0) < 1$;

c)在分离域内,错误分类的向量(如图 2 中灰色的点)满足不等式 $y_i(w^{\mathrm{T}} \cdot x + w_0) < 0$。

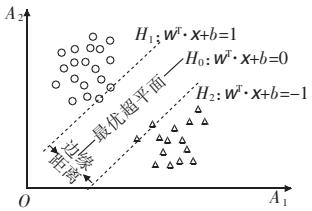

图 1　SVM 线性可分的情况

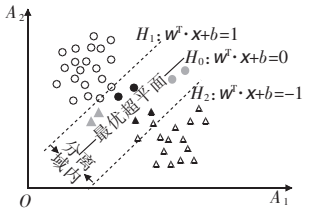

图 2　SVM 线性不可分的情况

通过引入松弛变量 ξ_i 将以上三种情况进行合并,即 $y_i(w^{\mathrm{T}} \cdot x + w_0) \geqslant 1 - \xi_i$。类似地,目标函数应该使得分类间隔尽量大,同时保持错误分类的点尽量少。基于以上分析,给出线性不可分的形式化描述:

$$\text{最小化}: \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} I(\xi_i) \tag{3}$$

$$\text{约束条件 1}: y_i(w^{\mathrm{T}} \cdot x + w_0) \geqslant 1 - \xi_i \quad i = 1, 2, \cdots, N \tag{4}$$

$$\text{约束条件 2}: \xi_i \geqslant 0 \quad i = 1, 2, \cdots, N \tag{5}$$

式(3)~(5)可以写成 Wolf 双重表达式[12]:

$$\text{最大化}: \sum_{i=1}^{N}\mu_i - \frac{1}{2}\sum_{i,j}\mu_i\mu_j y_i y_j x_i^{\mathrm{T}} x_j \tag{6}$$

$$\text{约束条件 1}: 0 \leqslant \mu_i \leqslant C \quad i = 1, 2, \cdots, N \tag{7}$$

$$\text{约束条件 2}: \sum_{i=1}^{N}\mu_i y_i = 0 \tag{8}$$

由式(6)~(8)可以看出,SVM 算法实现简单,对非线性数据分布具有较好的分类性能。因此,立足于 K-L 变换遴选出的 J 个流量统计特征,针对每一层抽样获取的 M 个数据,应用线性不可分的 SVM 进行流量分类(即将 SVM 抽象为 AdaBoost 机制中的弱分类器)。另外,由于 Internet 流量分类是一个典型的多分类问题,而原始 SVM 只处理二分类问题,为此本文采用了一对一的方法构建 $L(L-1)/2$ 个二元分类器来处理 L 元标签分类问题。

4　基于 AdaBoost-SVM 的流量分类

4.1　流量分类算法流程

AdaBoost-SVM 流量分类方法在构造 SVM 弱分类器的基础上,通过对各层分类结果的加权组合来提升算法的准确性。如图 3 所示,AdaBoost-SVM 算法主要包括两大步骤:AdaBoost-SVM 学习和在线流量识别。在学习阶段,根据数据包的五元组将报文划分为不同的流,然后获取每条流的统计特征,利用 AdaBoost-SVM 算法学习得到流量分类器。在线识别阶段,对真实流量进行同样的分流及特征统计,然后将其特征与学习阶段得到的流量分类器比较并输出识别结果,对于出现的新型业务类别,要进行反馈和重新学习。

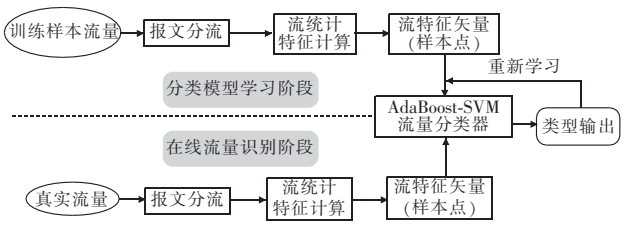

图 3　基于 AdaBoost-SVM 流量分类的逻辑流程

结合算法 1 给出的 AdaBoost 组合提升的具体步骤,图 4 给出了 AdaBoost-SVM 分类器的创建过程示意图:将整个流量数据的分类过程等分成若干层 SVM 弱分类器的组合叠加;基于 K-L 变换得到最优的 J 个流量统计特征,应用非线性 SVM 进行流量分类,其中每层抽样处理固定数量的样本;对于每层分类错误和正确的样本分别增加和减小其权重,使得 SVM 弱分类器聚焦在那些困难的数据样本上;利用分类能力一般的弱分类器,通过一定的方法进行组合叠加,最终生成一个强分类器。可以看出,非线性 SVM 分类器的目标是尽量减少分离域内被错误分类的流量样本,AdaBoost 方法通过分层组合和迭代权重的方法聚焦在困难分类的数据样本上,也就是分离域内容易被错误分类的样本上,AdaBoost-SVM 分类器结合了两者的优势,能有效提高分类算法的准确性。

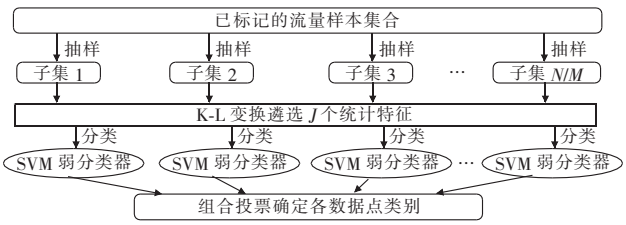

图 4　AdaBoost-SVM 分类器创建过程

4.2　算法的准确性分析

根据文献[12]的定理 6,AdaBoost-SVM 算法的分类错误

率满足不等式:$\varepsilon \leqslant 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t (1-\varepsilon_t)}$。其中 ε_t 为第 t 步的分类误差。如果每一步的分类误差小于随机猜测的概率 0.5,则可得 $\varepsilon_t = 0.5 - h_t$,其中 h_t 度量了比随机猜测更准确的程度。基于此,可进一步得到分类算法的错误概率满足

$$\varepsilon \leqslant \prod_{t=1}^{T} \sqrt{1-4h_t^2} \leqslant \exp\left(-2\sum_t h_t^2\right) \tag{9}$$

如果存在猜测变量 h,使得 $\forall h_t < h$ 成立,则可推出 $\varepsilon \leqslant \exp(-2T \cdot h^2)$。如图 5、6 所示,每层的分类准确性比随机猜测越好,则分类误差降低越快;且分类误差随迭代步骤的增大呈指数型递减,算法收敛也就越快。

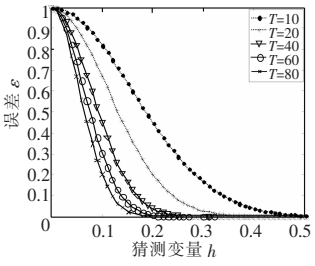

图 5 分类误差与每层精度的关系曲线

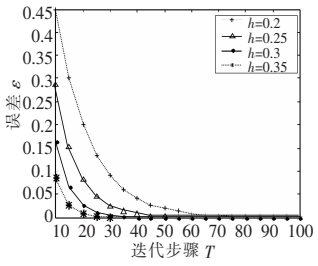

图 6 分类误差与迭代步骤的关系曲线

4.3 算法的计算复杂度分析

由 SVM 弱分类器的学习过程可知[12,13],单纯 SVM 分类的计算复杂度主要用于计算二次规划的最优化上,需要的计算复杂度为 $O(N^3)$。而 AdaBoost-SVM 将数据进行了分层处理,将一次大规模数据的分类转换为多次小规模的分层迭代处理。设定 $T = N/M$,即要进行 N/M 步的迭代分类,每一步分类数据样本的个数等于 M,则 AdaBoost-SVM 算法的计算复杂度为 $O(M^3 \cdot T) = O(M^2 \cdot N) < O(N^3)$。如果设定 $M = \sqrt{N}$,则 AdaBoost-SVM 的计算复杂度可以降为 $O(N^2)$。

5 实验结果及分析

5.1 算法评价指标

实验中采用以下三种评价指标:

定义 1 整体准确率。对于任意聚簇 $C_i \in C = \{C_1, \cdots, C_q\}$,其检测准确率为

$$P_i = \frac{\text{被正确分类为 } C_i \text{ 的样本流总数}}{\text{被分类为 } C_i \text{ 的样本流总数}} = \frac{N'_i}{N_i}$$

则分类器的整体准确率为

$$P_{\text{all}} = \frac{\text{所有被正确分类的样本流总数}}{\text{样本流总数}} = \sum_{i=1}^{q} N'_i / \sum_{i=1}^{q} N_i$$

定义 2 计算复杂度。针对某一数据集,分类算法收敛所需要的时间。

5.2 实验数据说明

为了便于仿真对比,本文采用文献[9]中的 Moore 数据集。Moore 数据集包含 10 个子数据集,每个数据集采自一天中的不同数据时间段,且每个数据集包含 28 min 内经过被测网络出口的所有完整 TCP 双向流(满足正常三次握手的 TCP 流)。Moore 数据集数据格式为 ARFF,可以使用 Weka 数据分析软件打开。如表 1 所示,Moore 数据集共包含 377 526 个网络流样本,10 种业务类型。Moore 数据集中每条网络流样本都是从一条完整的 TCP 双向流抽象而来,包含 249 项属性,即 249 个流的统计特征,如流的持续时间、每报文时间间隔等,具体描述可查阅文献[11]。

表 1 Moore Set 数据集详细信息

流量类型	具体应用	IP 流个数
WWW	HTTP,HTTPS	328 091
MAIL	IMAP,POP2/3,SMTP	28 567
BULK	FTP	11 539
SERVICES	X11,DNS,IDENT,LDAP,NTP	2 099
P2P	KaZaA,BitTorrent,GnuTella	2 094
DATABASE	POSTGRES,SQLNET,Oracle,INGRES	2 648
ATTACK	Internet worm and virus attacks	1 793
MULTIMEDIA	Windows Media Player,Real	1 152
INTERACTIVE	SSH,KLOGIN,RLOGIN,Telnet	110
GAMES	Half-Life	8

5.3 算法仿真比较

1)整体准确率仿真比较

在整体准确率的实验中,应用第一个 Moore 数据集(即 Moore Set 1)作为训练样本集,剩余九个(Moore Set 2 ~ 10)作为检测样本集,来评估分类结果。实验过程如下:针对每种流量类型,从 Moore Set 1 中随机抽取 N 个 IP 数据流作为训练样本,则样本集的总大小为 $10N$;在分类测试中,使用剩余的数据集 Moore Set 2 ~ 10 作为测试样本,每个数据集分别进行分类评估测试,可得到九个整体准确率,进行平均计算后,得到最后结果。其中,使用 Moore Set 2 ~ 10 作为测试样本进行 9 次评估,主要原因是 Moore Set 2 ~ 10 采自一天中的九个不同时段,能够在一定程度上反映不同算法对流量的依赖程度。

在 $N = 100, 200, 300, 400, 500$ 时,表 2 给出了四种算法整体准确率的比较,可以看出:四种算法随着训练样本集的增大,其准确性均有增长趋势;NBC + KE + FCBF、AdaBoost-SVM 算法的整体准确率要明显高于 K-means 和 NBC,并且 AdaBoost-SVM 算法的分类准确率最高。其主要原因:K-means 方法不依赖任何先验信息,且对初始 K 个聚类中心和异常数据点敏感,导致分类的准确性会处于较低的水平;NBC 方法要求参与分类的各项属性特征相互独立且遵循高斯分布,这与实际情况差距较大,导致分类性能下降;NBC + KE + FCBF、AdaBoost-SVM 两种方法均经过了特征过滤,剔除了不相关特征的影响,提高了算法的准确性;另外,AdaBoost-SVM 算法结合了 AdaBoost 和 SVM 算法的优势,对容易错误分类的样本数据进行聚焦处理,能进一步改善分类器的性能。

表 2 各分类算法整体准确率的比较 /%

样本集大小	K-means	NBC	NBC + KE + FCFB	AdaBoost-SVM
100 个 IP 流/类型	55.91	60.39	71.64	70.58
200 个 IP 流/类型	59.27	56.85	80.03	81.55
300 个 IP 流/类型	61.34	67.72	75.39	86.73
400 个 IP 流/类型	66.08	74.52	84.21	90.17
500 个 IP 流/类型	70.42	78.37	90.50	95.46

2)计算复杂度仿真

在 $N = 100, 200, 300, 400, 500$ 时,图 7 仿真了四种流量分类算法的计算复杂度。可以看出 AdaBoost-SVM 与 K-means 相当,NBC + KE + FCBF 计算复杂度最高。这主要是因为:K-means 以牺牲算法的准确性来换取模型的简单易用和计算过程相对高效;AdaBoost-SVM 将一次大规模数据的分类转换为多次小规模的分层迭代处理,能有效降低算法的计算复杂度;NBC + KE + FCBF 采用 FCBF 过滤机制和核估计技术对朴素贝叶斯方法计算变量取值概率和条件概率,复杂度较高。

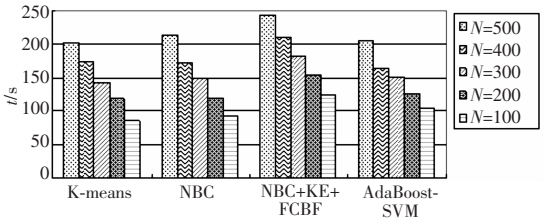

图 7　四种算法计算复杂度仿真

6　结束语

基于机器学习的流量分类方法利用业务流的统计特征进行识别,无须关心数据包的内容,使得基于机器学习的流量分类成为该领域一个新兴的研究方向。立足于算法的计算复杂度和准确性,本文提出一种基于 AdaBoost-SVM 流量分类器。首先,该方法利用 K-L 变换从大量冗余特征中遴选出本征特征;然后基于 AdaBoost 将一次分类等分成若干层弱分类过程,其中每层构造了非线性 SVM 弱分类器,并通过分层组合和迭代权重的方法聚焦在分离域内容易被错误分类的样本上;最后基于 Moore 数据集对算法进行了仿真实验,验证了 AdaBoost-SVM 既保持了 K-means 算法计算简单的优势,又提高了分类器的性能。另外,AdaBoost-SVM 算法依赖全部已标记的样本流,这在实际操作中需要大量的人工操作。本文下一步将结合半监督学习机制,研究如何利用少量已标记样本提高流量分类的性能。

参考文献:

[1] CNNIC. 中国互联网络发展状况统计报告 [EB/OL]. [2012-09-01]. http://www. cnnic. cn.

[2] Cisco. Networking solutions white paper [EB/OL]. [2012-09-01]. http://www. cisco. com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827//white_paper_c11-481360_ns827_Networking_Solutions_White_Paper. html.

[3] 杨家海,吴建平,安常青. 互联网络测量理论与应用 [M]. 北京:人民邮电出版社,2009.

[4] KARAGIANNIS T, BROIDO A. Transport layer identification of P2P traffic [C]//Proc of the 4th ACM SIGCOMM Internet Measurement Conference. 2004:121-134.

[5] SUBHABRATA S, SPATSCHECK O. Accurate, scalable in-network identification of P2P traffic using application signatures [C]//Proc of the 13th International World Wide Web Conference. 2004: 512-521.

[6] CASCARANO N, CIMINIERA L, RISSO F. Improving cost and accuracy of DPI traffic classifiers [C]//Proc of ACM Symposium on Applied Computing. 2010:641-646.

[7] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning [C]//Proc of the 30th IEEE Conference on Local Computer Networks. 2005: 15-17.

[8] ROUGHAN M, SEN S, SPATSCHECK O, *et al*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification [C]//Proc of the 4th ACM SIGCOMM Internet Measurement Conference. 2004:135-148.

[9] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques [C]//Proc of SIGMETRICS. 2005:50-60.

[10] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1):119-139.

[11] MOORE A W, ZUEV D. Discriminators for use in flow based classification, RR-05-13 [R]. London: Queen Mary University, 2005.

[12] HAN Jian-wei, KAMBER M. 数据挖掘概念与技术 [M]. 北京:机械工业出版社,2007.

[13] THEDORIDIS S, KOUTROUMBAS K. Pattern recognition [M]. 3rd ed. Beijing: Publishing House of Electronics Industry, 2010.

(上接第 1476 页)

4　结束语

本文在 BA 无标度网络的基础上,考虑了网络增长时节点度数增加的随机性,通过引入了度数增长服从泊松分布的改进模型,修改了新节点度数增长的方式,并利用率方程法分析了改进后模型的度分布的解析解。利用 MATLAB 仿真验证了理论的正确性,并发现只有在度分布 k 较大时才服从幂律分布,而在度分布 k 较小时总是有规律的弯曲,这和 BA 模型明显不同。文献[4,8]表明度分布弯曲现象在一些现实网络中也存在,因此改进后的网络模型既保留了度分布服从幂律分布这一现实网络特性,又比 BA 模型更好地拟合现实网络,利用它来刻画和分析现实中的复杂网络具有一定的研究价值。同时该模型也存在不足:a)模型中仅考虑了边数增加的随机性,而没有考虑节点数增加的随机性,考虑影响度分布的因素过于单一;b)模型中虽然体现了度分布弯曲现象,但是没有揭示现象背后的本质原因。因此,在下一步的工作中,将对网络模型作进一步修改,增加节点数的随机性使之更加符合现实网络,同时进一步分析度分布的弯曲现象,并试图揭示背后的原因。

参考文献:

[1] ERDÖS P, RÉNYI A. On the evolution of random graphs [J]. Publications of the Mathematical Institute of the Hungarian Academy of the Sciences, 1960(5):17-61.

[2] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks [J]. Nature, 1998, 393:440-442.

[3] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439):509-512.

[4] ALBERT R, BARABASI A L. Statistical mechanics of complex networks [J]. Review of Modern Physics, 2002, 74(1):47- 97.

[5] CLAUSET A, SHALIZI C R, NEWMAN M E J. Power-law distributions in empirical data [J]. SIAM Review, 2009, 51(4):661-703.

[6] BARABASI A L. Scale-free networks: a decade and beyond [J]. Science, 2009, 325(5939):412-413.

[7] 林兵,郭文忠,陈国龙,等. 标度网络中基于最短路径免疫策略的病毒传播研究 [J]. 计算机科学,2012,39(6):136-138.

[8] 窦炳琳,李澍淞,张世永. 基于结构的社会网络分析 [J]. 计算机学报,2012,35(4):741-752.

[9] 何敏华,张端明,王海艳,等. 基于无标度网络拓扑结构变化的舆论演化模型 [J]. 物理学报,2010,59(8):5175-5180.

[10] 王建伟,荣莉莉,于凯. 基于节点批量生长机制的无标度网络演化模型 [J]. 系统工程学报,2010,25(5):581-585.

[11] 陈琴琴,陈丹青. 基于二项分布随机增长的无标度网络模型 [J]. 数学研究,2010,43(2):185-191.

[12] LIU Zong-hua, LAI Ying-cheng, YE Nong, *et al*. Connective distribution and attack tolerance of general networks with both preferential and random attachments [J]. Physics Letters A, 2002, 303(5-6):337-344.

[13] BIANCONI G, BARABASI A L. Competion and multiscaling in evolving networks [J]. Europhysics Letters, 2001, 54(4):436-442.

[14] DOROGOVTSEY S N, MENDES J F F. Effect of the accelerating growth of communication networks on their structure [J]. Physical Review, 2001, 63(2):025101.

[15] SHI Ding-hua, CHEN Qing-hua, LIU Li-ming. Markov chain-based numerical method for degree distribution of growing networks [J]. Physical Review, 2005, 71(3):036140.