

基于 K 均值和双支持向量机的 P2P 流量识别方法

郭伟¹, 王西闯^{1*}, 肖振久²

(1. 辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105; 2. 中国传媒大学 计算机学院, 北京 100024)

(*通信作者电子邮箱 xichuang1988@126.com)

摘要:针对目前常用于 P2P 流量识别的有监督机器学习方法普遍存在时间代价较高的现状,提出采用时间代价为标准支持向量机四分之一的双支持向量机来构建分类器,并采用 K 均值集成方法快速生成有标签样本集,组合有标签样本集构成双支持向量机的训练样本,最后利用构建好的双支持向量机分类模型进行 P2P 流量的识别。实验结果表明采用基于 K 均值集成结合双支持向量机的方法在 P2P 流量识别的时间代价、准确率和稳定性方面要远优于标准支持向量机。

关键词: P2P 流量识别; 有监督机器学习; 双支持向量机; K 均值集成; 时间代价

中图分类号: TP393.06 **文献标志码:** A

P2P traffic identification method based on K -means and twin support vector machine

GUO Wei¹, WANG Xichuang^{1*}, XIAO Zhenjiu²

(1. College of Software, Liaoning Technical University, Huludao Liaoning 125105, China;

2. School of Computer, Communication University of China, Beijing 100024, China)

Abstract: Most of the P2P traffic identification methods have the problem of high time cost. Therefore, it was proposed to use TWin Support Vector Machine (TWSVM) whose time cost was a quarter of the common Support Vector Machine (SVM) to build classifier. K -means ensemble was used to create labeled sample set and labeled sample set was combined as the training sample of the TWSVM. At last, the constructed classification model was used to identify P2P traffic. The experimental results show that the method based on K -means and TWSVM can significantly decrease time cost of the P2P traffic identification, and has a higher accuracy rate and better stability than the standard SVM.

Key words: P2P traffic identification; supervised machine learning; TWin Support Vector Machine (TWSVM); K -means ensemble; time cost

0 引言

近年来,随着网络应用的疯狂增加,网络流量急剧增长,网络运营商的成本也大幅增加。然而,在所有的网络流量中,P2P 流量几乎占到 70%,所以如何准确地识别出 P2P 流量是网络流量识别中面临的一个重大难题,同时也是网络安全、流量计费、应用趋势分析等领域所面临的一个重要问题。然而,由于 P2P 流量不同于传统 Web 流量,其应用端口往往是动态变化的或者对端口以及传输数据等信息进行加密等;而且绝大多数的 P2P 应用没有统一的网络协议标准,都是一些不公开的专有协议(比如迅雷、eDonkey 等),这些都为 P2P 流量的识别带来了很大困难。

1 相关研究

截止到目前,P2P 流量识别方法^[1]主要有基于端口的识别方法、基于深层数据包检测(Deep Packet Inspection, DPI)技术的识别方法、基于网络行为的识别方法和基于机器学习的识别方法。

基于端口的识别技术主要考虑到一些网络应用是固定的端口,所以可以通过(IP, port)对来实现。然而,随着 P2P 技术的发展,现在的 P2P 软件大都是动态随机分配端口或者进

行端口加密,所以这一方法已经基本失效了。

基于 DPI 技术常利用模式匹配算法搜索流量载荷中 P2P 协议的特征值,进而判断是否属于 P2P 流量。应用层负载特征的提取是确保 DPI 技术识别准确率的关键,而模式匹配算法是确保 DPI 技术性能的关键。虽然 DPI 技术在非加密 P2P 流量的识别准确率较高,然而由于需要创建动态的 P2P 特征库,进而通过模式匹配进行 P2P 流量的识别,其空间和时间代价比较高,还涉及到个人隐私,在实际应用中涉及侵权问题。

基于网络行为的识别方法^[2],主要是利用 P2P 网络异于其他网络的一些特点进行 P2P 流量识别(如网络直径长、端口连接率高、TCP 连接突增等)。基于网络直径的识别方法由于需要存储各节点相关信息,时间和空间代价较高,不适合高速网络中流量的识别;并且基于网络行为的识别方法易受网络动态性影响,难以找到较好的分类属性特征。

基于机器学习的方法不依赖于应用层负载信息,它利用流量统计特征建立机器学习分类模型识别 P2P 流量,所以如何构建良好的分类模型就成为一个非常重要的问题。机器学习方法主要分为有监督机器学习、无监督机器学习和半监督机器学习,目前常用于 P2P 流量识别的机器学习方法是有监督的机器学习^[2]方法,主要有朴素贝叶斯、贝叶斯神经网络、

收稿日期:2013-04-23;修回日期:2013-06-17。

基金项目:国家自然科学基金资助项目(61103199);北京市自然科学基金资助项目(4112052)。

作者简介:郭伟(1970-),女,辽宁阜新人,副教授,主要研究方向:P2P 流量控制;王西闯(1988-),男,河南许昌人,硕士研究生,主要研究方向:P2P 流量识别;肖振久(1968-),男,辽宁阜新人,副教授,主要研究方向:信息安全。

C4.5 决策树以及支持向量机 (Support Vector Machine, SVM)^[3]。然而有监督机器学习方法需要依赖大量有标签样本,并且目前应用于P2P流量识别中的大部分有监督机器学习方法时间代价较高,难以应用到高速网络中。本文提出采用时间代价为传统支持向量机的四分之一的双支持向量机方法,大大提高了识别效率。然后,采用K均值聚类集成方法快速获得有标签样本数据,利用有标签样本对双支持向量机进行离线训练,最后用训练好的分类模型进行P2P流量的识别和控制。

2 数学模型

2.1 双支持向量机

支持向量机方法是一种建立在统计学习理论 (Statistical Learning Theory, SLT) 基础之上的机器学习方法,目前该方法已被广泛应用于语音识别、人脸识别、文本分类和网络流量控制、信息安全等领域。

支持向量机主要基于结构风险最小化原理,能够实现用较小样本训练获得较高泛化能力的决策函数。已知有训练集 $T = \{(x_i, y_i)\} (i = 1, 2, \dots, l)$, 其中正输入集合表示为矩阵 A (即 $y_i = +1$ 类), 负输入集合表示为矩阵 B (即 $y_i = -1$ 类)。标准支持向量机的目的是寻找一对最大间隔的最优超平面能将这2类样本准确分开 (如图1所示), 数学模型描述如下:

$$\begin{aligned} & \min_{\omega_1, b_1, \xi} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l \xi_i \\ & \text{s. t. } y_i(\omega \cdot x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, l \end{aligned} \quad (1)$$

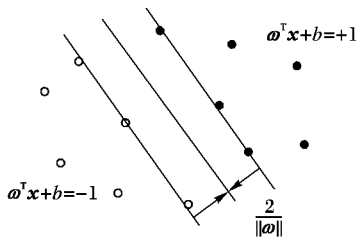


图1 SVM分类示意图

不同于标准支持向量机的是,双支持向量机 (Twin Support Vector Machine, TWSVM)^[4-5] 摒弃了平行约束的条件,通过构建两个非平行超平面,使得每一类样本离一个超平面尽可能近,而离另一个超平面尽可能远 (如图2所示)。

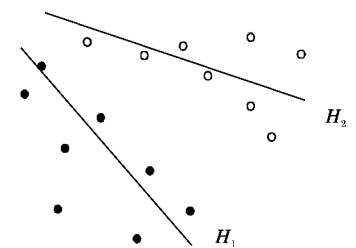


图2 TWSVM分类示意图

则双支持向量机分类器可以通过以下一对二次规划问题求解:

TWSVM1

$$\min_{\omega_1, b_1, \xi} \frac{1}{2} (A\omega_1 + e_1 b_1)^T (A\omega_1 + e_1 b_1) + c_1 e_2^T \xi \quad (2)$$

$$\text{s. t. } -(B\omega_1 + e_2 b_1)^T + \xi \geq e_2, \xi \geq 0$$

TWSVM2

$$\min_{\omega_2, b_2, \xi} \frac{1}{2} (B\omega_2 + e_2 b_2)^T (B\omega_2 + e_2 b_2) + c_2 e_1^T \xi \quad (3)$$

$$\text{s. t. } (A\omega_2 + e_1 b_2)^T + \xi \geq e_1, \xi \geq 0$$

对上述优化问题,利用拉格朗日变换转换为求两个对偶问题,然后根据KKT条件,对于待分类样本 x , 可以根据下式确定它的类别

$$x^T \omega_l + b_l = \min_{i=1,2} |x^T \omega_i + b_i|$$

其中 $| \cdot |$ 为 x 到平面 $x^T \omega_l + b_l = 0 (l = 1, 2)$ 的垂直距离。所以双支持向量机的分类函数为

$$f(x) = \arg \min_{k=1,2} |\omega_k x + b_k| \quad (4)$$

仿照标准支持向量机在处理线性分类时采用非线性映射把输入样本映射到高维空间中,也可以将核函数运用到双支持向量机中,求解过程与非线性情况类似,这里引入线性核函数 $K(X^T, C^T) = X^T C$ 。此时,最后的决策函数为:

$$f(x) = \arg \min_{k=1,2} |K(x^T, C^T) \omega_k + b_k| \quad (5)$$

其中 $C = [A \ B]^T$ 。

综上所述,TWSVM可以看成是标准支持向量机的分解,TWSVM中的每一个二次规划问题都类似于SVM,所以TWSVM算法可以看成是解决一对二次规划问题,而SVM则是解决一个二次规划问题。如果样本数相等TWSVM的时间复杂度为 $2 * (t/2)^3 = t^3/4$, 而SVM的时间复杂度为 t^3 , 可见双支持向量机的时间复杂度要远低于标准支持向量机的时间复杂度,大大提高了算法的效率。

2.2 K均值聚类模型

聚类是采用划分思想,按照簇内相似、簇间相异原则将数据对象划分为多个簇或类,衡量指标通常采用欧氏距离,距离近表明相似度高^[6-7]。

K均值是一种应用广泛的聚类算法,具有理论完善、算法简洁、收敛速度快等特点,能有效处理大数据集。K-means首先随机选定K个数据对象作为初始聚类中心,然后计算每个数据对象与K个聚类中心的距离,将数据对象划分到距离它最近的一类;然后更新聚类中心,重新计算每个簇的平均值作为新聚类中心;重复迭代至聚类中心不再变化或准则函数收敛为止。

以下为数学模型描述:

已知,数据集 $X = \{x_i | i = 1, 2, \dots, l\}$ 和整数K,其中 x_i 是d维样本向量,K-means方法的目标是寻找映射 $f: X \rightarrow \{1, 2, \dots, k\}$, 使得X中每个样本 x_i 能够映射到某簇 $j (1 \leq j \leq k)$ 中,并使准则函数J值最小:

$$J = \sum_{j=1}^k \sum_{i=1}^{N_j} \|x_i - z_j\|^2$$

其中: N_j 表示簇j的样本数, x_i 表示簇j所包含的样本, z_j 是簇j的中心。接下来对准则函数求偏导数,然后令偏导数为零,求出最优解:

$$z_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \quad (6)$$

式(6)最优解即为各簇中心,然后根据式(6)更新聚类中心,重新计算每个簇平均值作为新的聚类中心;重复以上步骤至聚类中心不再变化或准则函数收敛。

3 K-TWSVM模型

3.1 聚类集成

K均值聚类算法的效率与初始簇中心和K值的选择有很大关系,往往需要根据实际经验指定合适的K值,并且初始聚类中心的选择也对聚类结果有一定影响,对孤立点和“噪声”

数据也较敏感,稳定性差。为了提高 K 均值聚类方法的稳定性,采用 K -means 集成的半监督学习方法^[7]进行样本数据的预处理,为双支持向量机训练提供准确的标签样本。

聚类集成采用若干独立基聚类器分别对原始数据进行聚类,然后通过组合处理基分类器的结果来达到削弱噪声和孤立点的影响,增强结果的稳定性和鲁棒性。本文采用的是随机簇中心和 K 值的 K -means 方法。

仿照数据挖掘中构建组合分类器^[8]的方式构建 K -means 基聚类器。首先,采用随机簇中心生成 4 个聚类器,采用最大后验概率和硬划分方法^[9]给数据对象分配簇标签,根据最大相似原则选择簇标签作为无标签样本信息;最后采用投票机制为无标签样本分配标签。已知数据集 $S = \{(x_1, y_1), \dots, (x_l, y_l), x_{l+1}, \dots, x_{l+k}\}$, 其中前 l 个样本是有标签样本,后 k 个样本为无标签样本,通常 $k \gg l$, K -means 聚类集成步骤如下:

1) 生成基聚类器,利用有标签样本随机选择不同初始 K 值和簇中心,构成基聚类器。

2) 建立簇与标签间映射关系,根据最大后验概率分配簇标签。设簇到标签映射的概率分布 $P(Y = y_j | C_k)$, 其中 $j = 1, 2, \dots, m$ (m 为类别数), $k = 1, 2, \dots, K$ (K 为聚类数), 该概率分布值由极大似然 N_{jk}/N_k 估计, 其中 N_{jk} 是被分配到簇 k 中且标签为 j 样本数, N_k 是被分配到簇 k 中样本总数。

簇标签的决策函数为

$$f(l_k) = \arg \max_{j=1,2,\dots,m} P(Y = y_j | C_k) \quad (7)$$

3) 为无标签样本分配标签。给定样本 x_i , 在各基聚类器簇标签已知前提下, 标签分配函数为

$$f(l_i) = \arg \min_k d(x_i, c_k) \quad (8)$$

其中: $d()$ 为欧氏距离函数, c_k 代表簇 k 中心。

4) 按平均投票策略合成 4 个基聚类器的结果, 为无标签样本分配标签结果。

3.2 TWSVM 分类模型

因为实际应用中的问题几乎都是非线性的, 所以这里只讨论非线性情况下双支持向量机决策函数的求解和分类模型的建立过程^[10-11]。

已知有以下非线性训练样本集 $X = \{(x_1, y_1), \dots, (x_l, y_l) | i = 1, 2, \dots, l, y_i \in \{1, -1\} \text{ 为类别标识}\}$, 矩阵 A 和 B 分别表示 1 类和 -1 类样本点, 引入线性核函数 $K(X^T, C^T) = X^T C$, 其中 $C = [A \ B]^T$, 此时最终的决策函数变为

$$f(x) = \arg \min_{k=1,2} |K(x^T, C^T) \omega_k + b_k|$$

此时, 需要求解的优化问题为:

TWSVM1

$$\min_{\omega_1, b_1, \xi_2} \frac{1}{2} |K(A, C^T) \omega_1 + e_1 b_1|^2 + c_1 e_2^T \xi_2$$

$$\text{s. t. } -(K(B, C^T) \omega_1 + e_2 b_2) + \xi_2 \geq e_2$$

$$\xi_2 \geq 0$$

TWSVM2

$$\min_{\omega_2, b_2, \xi_1} \frac{1}{2} |K(B, C^T) \omega_2 + e_2 b_2|^2 + c_2 e_1^T \xi_1$$

$$\text{s. t. } (K(A, C^T) \omega_2 + e_1 b_2) + \xi_1 \geq e_1$$

$$\xi_1 \geq 0$$

转化为求解如下两个对偶问题:

DTWSVM1

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2} \alpha^T R (S^T S)^{-1} R^T \alpha$$

$$\text{s. t. } 0 \leq \alpha \leq c_1$$

DTWSVM2

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2} \gamma^T L (N^T N)^{-1} L^T \gamma$$

$$\text{s. t. } 0 \leq \gamma \leq c_2$$

其中 $R = [K(B, C^T) \ e_2]$, $S = [K(A, C^T) \ e_1]$, $L = S$, $N = R$; 令 $Z_1 = [\omega_1, b_1]$, $Z_2 = [\omega_2, b_2]$ 。

由 $Z_1 = -(S^T S)^{-1} R^T \alpha$, $Z_2 = -(N^T N)^{-1} L^T \gamma$ 可以求出决策函数中的 ω 和 b 。可以求得非线性样本下双支持向量机的分类决策函数为

$$f(x) = \arg \min_{k=1,2} |K(x^T, C^T) \omega_k + b_k|$$

最后, 根据以上决策函数确定待测样本类别。

3.3 P2P 流量识别模型

根据上面的数学模型, 图3给出了 P2P 流量的识别模型。

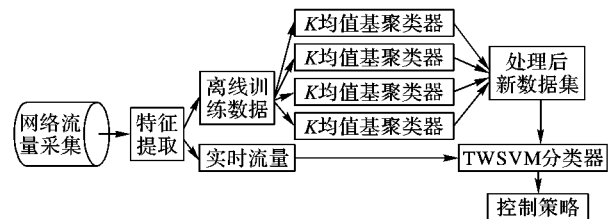


图3 P2P 流量识别模型

图3中 P2P 流量识别模型包含四个模块: 流量采集模块、特征提取模块、离线训练模块以及控制决策模块。其中: 流量采集模块负责采集网络流量, 生成离线训练数据和实时流量数据; 特征提取模块负责对采集到的流量数据进行特征选取, 去除冗余和不相关属性; 离线训练模块首先采利用有标签样本训练 4 个 K 均值基聚类器, 根据最大后验概率确定簇标签, 然后通过计算无标签样本和各簇之间的欧氏距离来给无标签样本分配标签, 最后将得到的标签数据集用于双支持向量机分类器进行训练; 控制决策模块利用离线训练的 TWSVM 分类模型进行 P2P 流量识别和采用相应的网络控制策略。算法描述如下:

K -means 聚类集成算法 $F(X, K)$ 如下:

已知 X 为样本集, K 为簇数, 输出簇标签 $L_l = (L_{l1}, L_{l2}, \dots, L_{lk})$ 和簇中心 $C = (c_1, c_2, \dots, c_k)$ 。

1) 随机选取 K 个样本作为簇中心 C , i 初始化。

2) 当簇中心没有改变且迭代变量未达到阈值进行如下操作:

a) 计算样本与各簇中心欧氏距离 $d(x_i, c_k)$, 依据最大相似性原则划分样本到最近簇;

b) 计算各簇样本均值, 作为新簇中心, 同时 $i = i + 1$;

c) 根据簇标签决策函数(7)分配簇标签。

识别模型算法如下:

1) 对带标签数据集 S 进行特征提取处理。

2) 从 S 中随机选择样本集 $X = \{(x_i, y_i) | x_i \text{ 为样本向量}, y_i \text{ 为标签信息}\}$ 和 $Y = \{x_j | j = 1, 2, \dots, k; x_j \text{ 为去除标签后样本}\}$ 。

3) 调用聚类集成算法 $F(X, K)$ 形成四个基聚类器, 然后根据标签决策函数(8)对 Y 分配标签。

4) 根据投票决策机制融合四个基聚类器结果生成带标签样本集 Y' , 组合 Y' 与 X 生成新训练集。

5) 用新生成的训练集训练 TWSVM 分类器, 在数据集 S 上进行 P2P 流量的识别。然而, 由于 K -TWSVM 增加了 K 均值聚类过程, 需要存储聚类结果, 空间代价增加, 所以要尽可能

能地精简样本特征集来减少空间消耗(文中采用 FCBF 特征选择算法实现)。

4 实验分析

4.1 实验环境和数据集

实验环境:新西兰怀卡托大学基于 Java 开发的开源数据挖掘平台 weka3.6、一台装有 Windows 7 操作系统和 Matlab 的个人 PC。实验数据采用 Moore 等使用的剑桥大学计算机实验室公开的网络数据集,记为 Moore_set^[12]。

Moore_set 中共 377 526 条网络流量,包括 10 种应用类型(WWW、MAIL、FTP-CONTROL、FTP-PASV、ATTACK、P2P、DATABASE、FTP-DATA、MULTIMEDIA、SERVICES、INTERACTIVE、GAMES)和 10 个数据集,每条网络流量都有 249 个属性,去除和 P2P 流量识别无关的 DATABASE、INTERACTIVE 和 GAME 流量,同时也去除每个数据集中的源端口和目的端口属性,记除 P2P 以外的剩余 6 种流量为非 P2P 流量。采用 FCBF 算法剔除冗余和不相关特征,采用排名前 10 的特征构成样本集,然后采用分层抽样从中抽取 2 000 条数据作为实验样本,记为 set-x。

首先对比标准支持向量机(SVM)、双支持向量机(TWSVM)和本文提出 K-TWSVM 在相同样本情况下的准确率和时间代价, K 值为 10;然后分别验证给定标签样本数下 K 值对 P2P 流量识别的影响和 K 值一定时有标签样本数对识别结果的影响。为验证稳定性,分别在各种情况下进行 5 次实验,取其平均值作为实验结果。

4.2 实验结果与分析

首先对比分析标准支持向量机 SVM、双支持向量机 TWSVM 和 K-TWSVM 在同一样本下的准确率和时间代价,抽取 set-x 的 10%、20%、40%、60%、80% 作为实验样本(分别记为 set-1 ~ set-5),进行 10 折交叉验证,实验结果如表 1、2 所示。

表 1 set-1 ~ set-5 上 P2P 流量分类准确率 %

数据集	SVM	TWSVM	K-TWSVM
set-1	67.90	71.40	75.40
set-2	78.60	79.24	88.42
set-3	80.14	81.32	90.68
set-4	88.12	88.38	92.56
set-5	89.88	90.18	92.72

从表 1 可以看出,随着训练样本数量的增加各种方法对 P2P 识别的准确率不断提高。其中 TWSVM 对 P2P 流量识别的准确率要略高于标准支持向量机 SVM,本文提出的 K-TWSVM 方法对 P2P 流量识别的准确率要远高于标准支持向量机和双支持向量机。

表 2 set-1 ~ set-5 上模型训练时间 ms

数据集	SVM	TWSVM	K-TWSVM
set-1	270	60	68
set-2	280	67	72
set-3	490	106	112
set-4	580	110	126
set-5	640	124	158

从表 2 中可以看出,TWSVM 训练模型建立时间要远小于 SVM,并且随着样本数量的增加,TWSVM 训练模型在时间花费上的优势越来越明显。K-TWSVM 训练模型建立时间要略高于 TWSVM,但远远小于 SVM 的训练时间,且 K-TWSVM 的

识别准确率也要远高于 SVM 和 TWSVM。所以,相比之下, K -TWSVM 方法要更适合应用于高速网络中的 P2P 流量识别。

综合表 1 和表 2 数据得到在实验样本 set-1 ~ set-5 上几种方法的平均识别准确率和平均时间代价如表 3 所示。

表 3 P2P 识别平均准确率和平均时间代价

方法	平均准确率/%	平均时间代价/ms
SVM	80.928	452.0
TWSVM	82.104	93.4
K-TWSVM	87.956	107.2

分析表 3 的结果进一步说明, K -TWSVM 无论在识别准确率和时间代价上都要优于 SVM。为了进一步分析在实验样本数不确定的情况下几种方法对 P2P 流量识别的准确率和时间代价的稳定性。

定义统计平均值为

$$Average_value = \frac{1}{n} \sum_{i=1}^n x_i$$

其中: x_i 表示不同情况下实验结果, n 为实验次数。

定义稳定性衡量指标偏差为

$$Dev = \frac{1}{n} \sum_{j=1}^n |x_j - Average_value|$$

统计分析几种方法在实验样本 set-1 ~ set-5 上对 P2P 流量识别准确率偏差和时间代价的偏差,如表 4 所示。

表 4 P2P 流量识别准确率偏差和时间代价偏差

方法	准确率偏差/%	时间代价偏差
SVM	6.46	141.60
TWSVM	5.74	23.92
K-TWSVM	3.02	21.76

由表 4 可以看出, K -TWSVM 方法在实验样本数不确定情况下对 P2P 流量识别准确率偏差要明显低于标准支持向量机和 TWSVM,这表明在不同实验样本下 K -TWSVM 方法对 P2P 流量识别准确率的稳定性要远好于标准支持向量机,也优于双支持向量机;在时间代价方面, K -TWSVM 方法多了 K 均值聚类对训练样本的处理过程,然而,由于 K 均值聚类算法的聚类速度很快,时间代价要稍多于双支持向量机方法,在时间代价的稳定性方面 K -TWSVM 却好于 TWSVM 和标准 SVM 方法。

接下来,验证给定标签样本数下 K 值对 P2P 流量识别的影响和 K 值一定情况下有标签样本数对识别结果的影响,如图 4 所示。

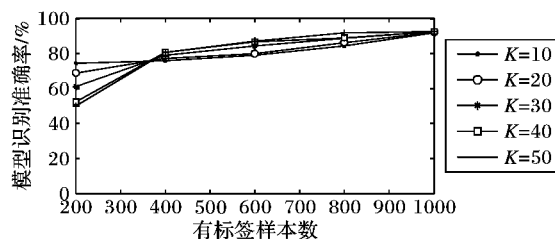


图 4 K -TWSVM 模型识别准确率

由图 4 可知,随着有标签样本数的增加, K -TWSVM 模型对 P2P 流量识别的准确率逐渐提高。当有标签样本数较少时, K 值越大反而会降低模型的识别准确率,主要原因是其未能全面反映样本分布信息;当有标签样本数达到 400 时, K 值越大, K -TWSVM 模型的 P2P 识别准确率越高。

最后,为了验证不同有标签样本数和不同 K 值下本文提

出 K -TWSVM 方法的稳定性, 统计分析五种含标签样本数和五种 K ($K = 10, 20, 30, 40, 50$) 值情况下, K -TWSVM 模型对 P2P 流量识别的平均准确率和偏差, 如表 5 所示。

表 5 K -TWSVM 模型 P2P 流量识别准确率和偏差

标签样本数	识别准确率/%	准确率偏差/%
200	61.36	8.32
400	78.85	3.62
600	83.54	2.76
800	88.01	1.54
1000	92.49	0.28

分析表 5 结果可知: 当含标签样本数适量时, 在不同 K 值条件下, 本文提出的 K -TWSVM 模型能有较高的 P2P 流量识别准确率, 同时也能保持极小的偏差。即当有标签样本数适量时, 针对不同 K 值, K -TWSVM 方法也能有较好的稳定性。

5 结语

本文提出的基于 K 均值和双支持向量机的 K -TWSVM 方法有较高的 P2P 流量识别准确率, 并且克服了传统 K 均值聚类方法受 K 值影响较大的缺点, 拥有很好的稳定性; 同时 K -TWSVM 的训练时间远小于标准支持向量机, 并且随样本数量的增加, 优势更明显。所以, 相比标准支持向量机, K -TWSVM 方法更适合应用于高速网络中 P2P 流量的识别。不过, 由于文中所用样本数量较少, 难以全面反映现实网络中的流量信息, 所以 K -TWSVM 对实时网络流量中 P2P 流量识别的效果还待进一步验证。此外, 本文仅分析了不同 K 值和含标签样本数对识别结果的影响, 未考虑 K -means 中不同初始聚类中心对实验结果的影响; 只是粗粒度分类 P2P 流量与非 P2P 流量, 并未确定具体的应用。所以, 讨论不同初始聚类中心对 P2P 流量识别结果的影响、如何根据 P2P 流量确定具体的

P2P 应用和如何选取适当的初始聚类中心是下一步的研究工作。

参考文献:

- [1] 吴敏. P2P 网络流量控制管理若干关键技术研究[D]. 南京: 南京邮电大学, 2011.
 - [2] 邬书跃, 余杰, 樊晓平. 基于流量与行为特征的 P2P 流量识别模型[J]. 计算机工程, 2012, 38(16): 182-184.
 - [3] 徐鹏, 刘琼, 林森. 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.
 - [4] YE Q L, ZHAO C X, YE N. Least squares twin support vector machine classification via maximum one-class within class variance[J]. Optimization Methods and Software, 2012, 27(1): 53-69.
 - [5] 王震. 基于双重支持向量机的分类算法研究[D]. 长春: 吉林大学, 2010.
 - [6] 毕晓君, 宫汝江. 一种结合人工蜂群和 K -均值的混合聚类方法[J]. 计算机应用研究, 2012, 29(6): 2040-2042.
 - [7] 郑丹, 王潜平. K -means 初始聚类中心的选择算法[J]. 计算机应用, 2012, 32(8): 2186-2188.
 - [8] STEINBACH M. Introduction to data mining[M]. 2 版. 范明, 范宏建, 译. 北京: 人民邮电出版社, 2011.
 - [9] 刑迪, 葛洪伟. 半监督 FSVM 在羽绒菱节识别中的应用[J]. 计算机工程与应用, 2013, 49(1): 242-244.
 - [10] KHEMCHANDANI J R, CHANDRA S. Twin support vector machines for pattern classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910.
 - [11] 谢娟英, 张兵权, 汪万紫. 基于双支持向量机的偏二叉树多类分类算法[J]. 南京大学学报: 自然科学版, 2011, 47(4): 354-363.
 - [12] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]// Proceedings of the 2005 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems. New York: ACM, 2005: 50-60.
- (上接第 2718 页)
- 信号的准确度对于系统性能至关重要。本文基于最大最小准则, 利用似然比估计中继接收信号质量, 提出了一种自适应协作选择方案。本文的难点在于确定合适的似然比门限, 以保证中继信道在无法正确解码时仍可以提供增益, 为此, 本文通过推导近似的误比特率闭合解来对门限进行优化。仿真结果表明, 本方案及其最优门限能有效避免错误传播, 具有一定的性能优势。需要说明的是, 本文只考虑了各节点均为单天线的情况, 因此在未来的工作中, 需要进一步研究多天线下情况的信号质量建模, 并且设计相应的协作方式转换算法。
- ## 参考文献:
- [1] LANEMAN J N, TSE D N C, WORNELL G W. Cooperative diversity in wireless networks: Efficient protocols and outage behavior[J]. IEEE Transactions on Information Theory, 2004, 51(12): 3062-3080.
 - [2] OGGIER F, RCKAYA G, BELFIORE D, et al. Perfect space-time block codes[J]. IEEE Transactions on Information Theory, 2006, 52(9): 3885-3902.
 - [3] BLETSAS A, KHISTI A, REED D P, et al. A simple cooperative diversity method based on network path selection[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(3): 659-672.
 - [4] KRIKIDIS I. Relay selection for two-way relay channels with MABC DF: a diversity perspective[J]. IEEE Transactions on Vehicular Technology, 2010, 59(9): 4620-4628.
 - [5] BERE E, ADVE R S. Selection cooperation in multi-source cooperative networks[J]. IEEE Transactions on Wireless Communications, 2008, 7(1): 118-127.
 - [6] JING Y, JAFARKHANI F. Single and multiple relay selection schemes and their achievable diversity orders[J]. IEEE Transactions on Wireless Communications, 2009, 8(3): 1414-1423.
 - [7] QIANG L, TING S H, PANDHARIPANDE A, et al. Adaptive two-way relaying and outage analysis[J]. IEEE Transactions on Wireless Communications, 2009, 8(6): 3288-3299.
 - [8] LEE I H, KIM D. Outage performance of opportunistic cooperation in amplify-and-forward relaying systems with relay selection[J]. IEEE Communications Letters, 2012, 16(2): 224-227.
 - [9] HARBIAN G, GHAYEB A, HASNA M, et al. Threshold-based relaying in coded cooperative networks[J]. IEEE Transactions on Vehicular Technology, 2011, 60(1): 123-135.
 - [10] ZHANG X H, GHAYEB A, HASNA M. On relay assignment in network-coded cooperative systems[J]. IEEE Transactions on Wireless Communications, 2011, 10(3): 868-876.
 - [11] ZHANG X H, HASNA M, GHAYEB A. Performance analysis of relay assignment schemes for cooperative networks with multiple source-destination pairs[J]. IEEE Transactions on Wireless Communications, 2012, 11(1): 166-177.
 - [12] GRADSHTEYN I S, RYZHIK I M. Table of integrals, series and products[M]. New York: Academic Press, 2007.
 - [13] KIM S W, KIM E Y. Optimum receive antenna selection minimizing error probability[C]// Proceedings of 2003 IEEE Wireless Communications and Networking Conference. Washington, DC: IEEE Computer Society, 2003: 441-447.
 - [14] PROAKIS J G. Digital communications[M]. New York: McGraw-Hill, 2001.