基于K均值集成和SVM的P2P流量识别研究

刘三民[1,2] 孙知信[2,3,4] 刘余霞[5]
(安徽工程大学计算机与信息学院 芜湖 241000)[1]
(南京航空航天大学计算机科学与技术学院 南京 210016)[2]
(南京邮电大学计算机技术研究所 南京 210003)[3]
(南京大学计算机软件新技术国家重点实验室 南京 210093)[4]
(安徽工程大学电气工程学院 芜湖 241000)[5]

摘 要 提出基于K均值集成和支持向量机相结合的P2P流量识别模型,以保证流量识别精度和稳定性,克服聚类识别模型中参数值难以确定、复杂性高等缺点。对少量标签样本采用随机簇中心的K均值算法训练基聚类器,按最大后验概率分配簇标签,无标签样本与其最近簇标签一致;按投票机制集成无标签样本标签信息,并结合原标签样本训练支持向量机识别模型。该模型利用了集成学习稳定性和SVM在小样本集上的良好泛化性能。理论分析和仿真实验结果证明了方案的可行性。

关键词 流量识别,支持向量机,K均值,集成学习
中图法分类号 TP393 **文献标识码** A

Research on P2P Traffic Identification Based on K-means Ensemble and SVM

LIU San-min[1,2] SUN Zhi-xin[2,3,4] LIU Yu-xia[5]
(College of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China)[1]
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)[2]
(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[3]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[4]
(College of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China)[5]

Abstract A P2P traffic identification model was constructed by the combination of K-means ensemble and support vector machine. It owns high accuracy, stability and overcomes complexity of cluster model. Firstly, the three base clusterer was formed by few labeled sample, and then the each cluster's label was assigned by MAP. The unlabeled sample's label is the same with the closest cluster. Identification model based on SVM was built by new sample set. The model makes the best of ensemble learning's stability and SVM's generalization ability, theoretical analysis and result demonstrate its feasibility.

Keywords Traffic identification, Support vector machines, K-means, Ensemble learning

1 引言

网络应用层出不穷和新技术不断引入,导致基于端口和特征字段的流量识别方法效率较低[1]。在网络环境中获取大量"昂贵"样本标签费时、费力,而取得"廉价"无标签样本相对容易,所以基于半监督学习进行流量识别的研究方兴未艾。

文献[2]使用包统计特征和最大期望(Expectation Maximization, EM)算法对流量进行聚类分析,实验得到不同流量特性的聚类簇。Roughan等[3]利用流的持续时间和平均包大小特征,采用K-近邻(K-nearest Neighbor, KNN)方法处理流量识别问题,该方法适合低维空间同时要求较大的存储空间以便记住样本。Bernaille等[4]用TCP连接前5个数据分组长度序列构建流量样本向量,采用K均值和谱聚类等算法实现流量识别,该方案依赖于数据分组到达次序,稳定性难以保证。Erman等[5-7]详细分析有监督学习和无监督学习的异同点,在此基础上设计聚类流量识别方案,并对K-means、DBSCAN和AutoClass 3种无监督学习方法做对比分析。文献[8]利用混合高斯模型构建基于聚类的半监督流量识别模型,具有较好的识别能力。文献[9]设计一种基于最大方差属性的起始簇中心K均值方法实现流量识别,以保证各簇具有较好的分离性。文献[10]在K均值基础上设计出正常流量和僵尸网络流量识别模型。

到稿日期:2011-05-13 返修日期:2011-09-15 本文受国家自然科学基金(60973140),江苏省自然基金(BK2009425),江苏省高校自然科学基础研究项目(08KJB520005),安徽工程大学校青年基金项目(2008YQ041)资助。

刘三民(1978-),男,博士生,讲师,主要研究方向为模式识别、流量检测,E-mail:aqlsm@163.com;孙知信(1964-),男,教授,博士生导师,主要研究方向为计算机网络与安全、多媒体通信;刘余霞(1980-),女,硕士生,主要研究方向为自动检测。

基于聚类学习的流量识别模型稳定性不好，计算量和存储开销较大。本文提出基于K均值(K-means)集成与支持向量机(Support Vector Machine，SVM)相结合的流量识别方案，以保证流量识别的稳定性和高效率。本方案首先用少量的标签样本进行半监督学习，再对无标签样本的标签分配结果进行集成学习，得到稳定的标签样本集。本模型充分利用了集成学习优点和SVM在小样本学习中的良好泛化能力，实验结果证明了方案的可行性和稳定性。

2 数学模型

2.1 支持向量机

支持向量机是以统计学习理论为基础，基于结构风险最小化原则实现的一种小样本学习模型，具有较好的泛化能力。

数据集 $X=\{(x_i,y_i)\mid i=1,\cdots,l\}$，其中 x_i 是 d 维向量。对于二分类问题，y_i 取1或−1。在空间寻找一个最优超平面使两类样本线性可分，即使两类样本到超平面距离最大化，其数学模型描述如下：

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i\\&\text{s.t. } y_i((w\cdot\phi(x_i))+b)\geqslant 1-\xi_i,i=1,\cdots,l\\&\quad \xi_i\geqslant 0,\quad i=1,\cdots,l\end{aligned}\tag{1}$$

式中，C 表示惩罚系数，ξ_i 表示松驰变量。根据拉格朗日函数和KTT条件，上式可以转化成如下对偶最优化问题：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(x_i\cdot x_j)-\sum_{j=1}^{l}\alpha_j\\&\text{s.t. }\sum_{i=1}^{l}y_i\alpha_i=0\\&\quad 0\leqslant\alpha_i\leqslant C,i=1,\cdots,l\end{aligned}\tag{2}$$

根据对偶问题最优解 $\alpha^*=(\alpha_1^*,\cdots,\alpha_l^*)$，即可求出最优超平面 $\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x)+b^*=0$。

在线性不可分的情况下，采用核技巧将其映射至高维空间进行线性求解，此时最优化问题如式(3)所示：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(\Phi(x_i)\cdot\Phi(x_j))-\sum_{j=1}^{l}\alpha_j\\&\text{s.t. }\sum_{i=1}^{l}\alpha_iy_i=1\\&\quad 0\leqslant\alpha_i\leqslant C,i=1,\cdots,l\end{aligned}\tag{3}$$

相应决策函数 $f(x)=\mathrm{sgn}(\sum_{i=1}^{l}y_i\alpha_i^*(\Phi(x_i)\cdot\Phi(x))+b^*)$

2.2 K均值聚类模型

K均值是一种无监督学习方法，通过划分思想按簇内相似、簇间相异原则把样本集划分成 K 个簇。通常相似性以欧氏空间中距离作为测度，距离近，相似度高。

数据集 $X=\{x_i\mid i=1,\cdots,l\}$ 和整数值 K，其中 x_i 是 d 维样本向量。寻找映射 $f:X\to\{1,\cdots,K\}$，保证数据集中每个样本 x_i 映射到某簇 $j(1\leqslant j\leqslant K)$ 中，并使准则函数 J 值最小：

$$J=\sum_{j=1}^{K}\sum_{i=1}^{N_j}\|x_i-z_j\|^2\tag{4}$$

式中，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。对准则函数求偏导数，并使其等于0，可求出最优解：

$$z_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_i\tag{5}$$

由式(5)可知最优解是各簇中心。

3 流量识别模型

3.1 聚类集成

K均值效率与初始中心和 K 值的选择有较大关系，对于孤立点和“噪声”数据敏感，稳定性较差。为提高聚类结果的稳定性，为SVM训练提供准确的标签样本，本文采用基于K-means集成的半监督学习方案，实现样本数据的预处理。

聚类集成[11]是用若干独立的基聚类器分别对原始数据进行聚类，然后对基聚类器结果进行组合，减弱噪声和孤立点的影响，增强聚类结果的稳定性和鲁棒性。基聚类器形成通常有4种[12]：一是不同特征集合视图的样本形成基聚类器；二是随机初始簇中心和簇数，如K-means；三是随机样本输入顺序；四是不同的聚类算法形成基聚类器。

本文方案采用随机聚类中心生成3个基聚类器，基于最大后验概率和硬划分方法分配簇标签，按最大相似性原则选择簇标签作无标签样本信息。在集成过程中采用投票机制给无标签样本赋予样本标签。数据集 $D=\{(x_1,y_1),\cdots,(x_l,y_l),x_{l+1},\cdots,x_{l+k}\}$，其中前 l 个样本表示标签样本，相应的 y_l 是标签代表类别，后 k 个样本是无标签样本，一般情况 $k\gg l$。

聚类集成算法思想描述如下：

(1)形成基聚类器。利用标签样本随机选择初始聚类中心和不同 K 值，形成基聚类器。

(2)建立簇与标签间映射关系。按最大后验概率进行簇标签分配。簇到标签映射概率分布为 $P(Y=y_j\mid C_k)$，其中 $j=1,2,\cdots,q$(q 是类别数)，$k=1,2,\cdots K$(K 是聚类数)。该概率分布值由极大似然 N_{jk}/N_k 估计，其中 N_{jk} 是被分到簇 k 标签为 j 的样本数量，N_k 是被分配到簇 k 中样本的总数量。

簇标签的决策函数为：

$$label_k=\arg\max_{j=1,\cdots,q}P(Y=y_j\mid C_k)$$

(3)无标签样本分配标签。给定流样本 x_i，在各基聚类器簇标签已知的情况下，标签分配函数为：

$$label_i=\arg\min_k d(x_i,c_k)$$

式中，$d(,)$ 表示欧氏距离函数，c_k 代表簇 k 中心。即流样本 x_i 标签与其相邻最近的簇标签一致。

(4)标签合成。按少数服从多数合成3个基聚类器，对无标签样本标签分配结果。

3.2 识别模型

在上述思想指导下，识别模型如图1所示。该模型主要包含4大功能模块。

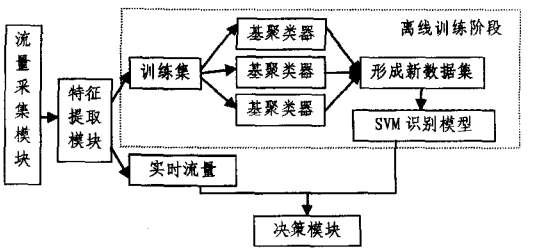

图1 流量识别模型

流量采集模块：负责从网络采集流量，形成相应的训练样本集和实时流量识别样本信息。

特征提取模块：对流样本进行特征选择，去除冗余或不相关的属性集合，以便降低计算复杂度，提高识别效率。

离线训练模块：该模块的目标是得到精准的分类模型。为减少无标签样本对聚类过程产生影响，由标签样本训练3个K-means基聚类器，按最大后验概率给簇分配标签，然后计算无标签样本相对各簇距离的远近来分配标签给样本，形成最后的标签数据集，进行SVM分类模型的训练。

决策模块：利用离线训练的识别模型进行网络流量的实时识别，同时可以根据预置策略采取相应的网络管理措施。

3.3 算法描述

根据流量识别模型，给出仿真实验相应算法：

(1)对带标签信息的流样本集合 D，进行特征抽取，实现降维。

(2)从 D 中随机选择样本集 $X=\{(x_i,y_i)\mid x_i$ 是样本向量，y_i 是标签信息$\}$和 $Y=\{x_j\mid j=1,\cdots,k,x_j$ 是去除标签后的样本$\}$。

(3)调用聚类算法 kmeans(X,K) 形成3个基聚类器。在基聚类器上，用标签决策函数对 Y 分配标签信息。

(4)按投票机制形成带标签样本集 Y'，并与 X 组合形成训练集。

(5)训练SVM分类模型，并在数据集 D 上进行测试。

聚类算法 kmeans(X,K)，X 为样本集，K 是簇数，输出簇标签 $lbl=\{lbl_1,lbl_2,\cdots,lbl_k\}$和簇中心 $c=\{c_1,c_2,\cdots,c_k\}$。算法步骤可作如下描述：

(1)随机选取 K 个样本作为簇中心 c，迭代变量 $iter$ 赋初值。

(2)while(ischange(c)&$iter<$maxIter)

a)计算样本与各簇中心的距离，将样本划分至最近簇，即最相似簇。

b)计算各簇样本平均值，作为新簇中心。

c)$iter=iter+1$。

(3)按簇标签决策函数分配簇标签。

4 实验及结果分析

4.1 实验准备

实验采用Moore数据集[13]。该数据集共10个子集，包含10种应用类型，各样本共有249个属性，较全面地反映了网络流量相关特征。

从数据集中随机抽取目前网络中两大类型流量WWW(代表非P2P流量)和P2P样本作为实验数据集。采用FCBF[14]过滤算法剔除冗余或不相关属性，采用排名前十的属性特征构成样本向量，实现降维。

标签样本提供簇分布信息，而无标签样本是为增加样本分布信息。在固定无标签样本前提下，设计两种方案验证预期目标：一是给定标签样本数挖掘 K 值对流量识别的影响；二是在相同 K 值下标签标本数对分类的影响。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。上述实验均在Matlab7.1和软件LibSVM中完成，支持向量机相关参数均是默认设置。

4.2 结果分析

经上述实验方案后，可得如下相关实验数据。

图2反映模型准确率、标签样本数和聚类簇数3者的关系。从横轴 X 可得，随着加入的标签样本的增多，模型的准确率显著提高。因为随着样本的增加，经聚类集成的样本分布信息越全面、准确。在相同标签样本数的情况下，聚类簇数设置较小，其精度明显偏低。从图中可看出，在标签样本数大于400时，簇数值越大精度越高。在样本标签过少的情况下，簇数值较大反而会制约模型精度，这是因为其未能全面反映分布信息。

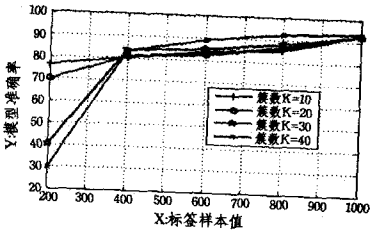

图2 模型准确率

定义统计量平均值(Med)和偏差(Dev)来衡量给定标签样本数实验结果的稳定性，其中 x_i 表示在不同 K($K=10$，20,30,40)值下5次实验精度的平均值。

$$Med=\frac{1}{4}\sum_{i=1}^{4}x_i$$

$$Dev=\frac{1}{4}\sum_{j=1}^{4}|x_i-Med|$$

表1表示在5种标签样本数的情况下，4种不同 K 值实验环境下的平均结果及偏差。

表1 模型精度均值与偏差

	200	400	600	800	1000
平均值	53.9	81.9	85.1	88.2	92.8
偏差	19.2	2.2	2.02	1.42	0.33

分析聚类结果偏差值可知，当标签样本适量时，本文方案具有较好的稳定性，簇数设置影响较小，避免了基于聚类半监督识别模型 K 值影响较大的缺点。

结束语 基于K均值集成和SVM相结合的P2P流量识别模型精度高、稳定性好；克服了现有基于聚类学习的流量识别模型中 K 值难于确定、流量识别过程计算较复杂的缺点。由于无标签样本组成比例是随机的，导致在SVM训练过程中出现样本不平衡现象，本文未进行深入分析；在进行MAP形成簇标签时，可以通过设置阈值来发现新应用类型，这些将是后续工作。

参考文献

[1] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]// Proc the PAM 2005. LNCS 3431, Heidelberg: Springer-Verlag, 2005: 41-54

[2] McGregor A, Hall M, Lorier P, et al. flow clustering using machine learning techniques[C]// Proc of 5th Passive Active Measurement Workshop. Apr 2004, 3015: 205-214

[3] 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 ACM SIGCOMM Internet Measurement Conf. Taormina, 2004: 135-148

[4] Bernaille L, Teixeira R, Salamatian K. Early application identification[C]// Proc of 2006 ACM CoNEXT Conf. Lisboa, 2006: 1-12

[5] Erman J, Mahanti A, Arlitt M. Internet traffic identification using machine lear-ning techniques[C]// Proc of the 49th IEEE GLOBECOM. San Francisco, 2006: 1-6

(下转第74页)

R_{BS})是否与可信中心发送的相等，即可验证信息是否为可信中心发送来的，是否被篡改。

没有人可以假冒买方：水印申请阶段，买方计算本次水印申请使用的口令 $P_i=H(N_i \| ID_{B_i})$，其中的 N_i 为秘密随机数，水印中心可以验证一次性口令来对买方进行认证。交易阶段，买方计算 $H(ID_{B_i} \| R_{BS})$，其中 R_{BS} 为秘密随机数，发送给卖方，卖方通过自己存储的 R_{BS}，计算 $H(ID_{B_i} \| R_{BS})$ 是否与发送来的一致，以对买方进行认证。

没有人可以冒充卖方：交易阶段中，卖方将 $(C, H(C \| R_{BS}))$ 发送给买方，买方根据自己存储的 R_{BS}，即可对卖方进行认证。

3.2.2 匿名性

在注册阶段，水印中心利用买方的真实身份和随机数 N_1，计算买方的匿名身份，买方在第一次申请水印时使用该身份，第一次申请水印的过程中，水印中心在返回水印的同时，也安全返回了随机数 N_2，用于买方计算第二次申请水印时的临时身份，这样每次买方的临时身份都会不一样，不会被恶意跟踪者追踪。

3.2.3 抵抗伪造攻击

交易阶段卖方将 $(C, H(C \| R_{BS}))$ 发送给买方，买方根据自己存储的 R_{BS}，既可以对卖方的身份进行认证，在认证的同时也验证了数字产品的正确性与完整性。

3.2.4 不可诬告性

协议采用了同态加密技术，买方的指纹水印是使用其公钥加密的，只有买方自己能得到水印，卖方不知道买方的具体水印，交易过程中填写了订单，卖方不可能将买方的水印应用到其他的数字产品中，恶意诬告买方。

3.2.5 不可否认性

在数字产品中，嵌入了变换后买方的指纹水印，买方得到数字产品后不可能从中提取出自己的指纹水印，若发现非法副本，其中必定含有买方的指纹水印，买方无法否认。

3.2.6 可追踪性

在数字产品中，嵌入了唯一的只与本次交易相关的版权水印，如果恶意买方非法得到数字产品，或将其公开发布出来，卖方发现非法副本后，能过通过版权水印找到叛逆者，再根据数字产品中的指纹水印确定以上信息后，将自己存储的信息提交给仲裁机构进行仲裁，并要求 CA 给出恶意买方的真实身份，对其进行起诉。

结束语 本文提出了一种基于移动设备的匿名的、可追踪版权保护协议，它借鉴了已有方案中的同态加密技术嵌入指纹水印和版权水印，做到了可追踪性。该协议在保证数字版权管理系统的基本要求的同时，进一步解决了用户的身份隐藏问题，设计了计算量小的高效认证方案，亦即将一些计算量大的计算从移动用户转嫁给了计算能力较强的可信中心，适应了移动设备计算能力和存储能力相对较弱的特点。

参 考 文 献

[1] 李润峰，马兆丰，杨义先，等. 数字版权管理安全性评测模型研究[J]. 计算机科学，2011，38(3)：24-27

[2] Fan Chu-ni, Chen Ming-te, Sun Wei-zhe. Buyer-seller watermarking protocols with off-line trusted parties[C]// MUE'07. International Conference on Multimedia and Ubiquitous Engineering. Secul: Kaoksiurg, 2007(4): 1035-1040

[3] Phan R C-W, Goi B-M, Poh G-S, et al. Analysis of a Buyer-Seller Watermarking Protocol for Trustworthy Purchasing of Digital Contents [J]. Wireless Pers Commun, 2011, 56: 73-83

[4] 胡玉平，张军. 用于盗版追踪的数字水印协议的研究[J]. 计算机科学，2010，37(1)：91-94

[5] Chen Chin-ling. A secure and traceable E-DRM system based on mobile device[J]. Expert Systems with Applications, 2008, 35: 878-886

[6] Chen T-h, Horng G. A lightweight and anonymous copyright-protection protocol[J]. Computer Standards & Interfaces, 2007, 29: 229-237

[7] Li M-J, Juan J S-T, Tsai J H-C. Practical electronic auction scheme with strong anonymity and bidding privacy[J]. Information Sciences, 2011, 181(12): 2550-2570

[8] Zhao Xing-wen, Zhang Fang-guo. A New Type of ID-based Encryption System and Its Application to Pay-TV Systems[J]. International Journal of Network Security, 2011, 13(3): 161-166

[9] Lin Shi-guo, Chen Xi. Secure and tracing multimedia distribution for convergent Mobile TV services [J]. Computer Communications, 2010, 23(3): 1664-1673

[10] Tomas-Buliart J, Fernandez M, Soriano M. Traitor tracing over YouTuBe video service-proof of concept[J]. Telecommun Syst, 2010(45): 47-60

(上接第 48 页)

[6] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms [C]// Proc of the SIGCOMM Workshop on Mining Network Data. Pisa, Italy, 2006: 281-286

[7] Erman J, Mahanti A, Arlitt M, et al. Offline/realtime traffic classification using semi supe-rvised learning[C]// 26th International Symposium on Computer Performance, Modeling, Measurements, and Evaluation. 2007, 64(9-12): 1194-1213

[8] Qian Feng, Hu Guang-min, Yao Xing-miao. Semi-supervised Internet network traffic classification using a Gaussian mixture model[J]. Electronics and Communications, 2008(62): 557-564

[9] Lin Guan-zhou, Xin Yang, et al. Network traffic classification based on semisupervised clustering [J]. China University of Posts and Telecommunications, 2010(17): 84-88

[10] Lu W, Rammidi G, Ghorbani A A. Clustering botnet communication traffic based on n-gram feature selection[J]. Computer Communications, 2010: 1-13

[11] 唐伟，周志华. 基于 Bagging 的选择性聚类集成[J]. 软件学报，2005，16(4)：496-502

[12] Strehl A, Ghosh J. Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Machine Learning Research, 2002(3): 583-617

[13] Moore A W, Zuev D, Crogan M. Discriminators for use in flow-based classification [R]. RR-05-13. London: Queen Mary University of London, 2005

[14] Yu Lei, Liu Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]// Proc of the 20th International Conference on Machine Learning. 2003: 1-8