基于近邻传播学习的半监督流量分类方法

张 震 [1] 汪斌强 [1] 李向涛 [1] 黄万伟 [1]

摘 要 准确的流量分类是进行网络管理、安全检测以及应用趋势分析的基础. 针对完全监督和无监督分类的缺陷, 提出了一种基于近邻传播学习的半监督流量分类方法. 通过引入 "近邻传播聚类" 机制构建分类模型, 使得分类器实现过程简单、运行高效. 应用 "半监督学习" 的思想, 抽象出少量已标记样本流约束和流形空间先验信息, 定义了 "流形相似度" 的距离测度, 既降低了标记流量样本的复杂度, 又提高了流量分类器的性能. 理论分析和实验结果表明: 算法具有较高的分类准确性和较好的凝聚性.

关键词 流量分类, 半监督学习, 近邻传播聚类, 流形相似度

引用格式 张震, 汪斌强, 李向涛, 黄万伟. 基于近邻传播学习的半监督流量分类方法. 自动化学报, 2013, **39**(7): 1100−1109

DOI 10.3724/SP.J.1004.2013.01100

Semi-supervised Traffic Identification Based on Affinity Propagation

ZHANG Zhen[1] WANG Bin-Qiang[1] LI Xiang-Tao[1] HUANG Wan-Wei[1]

Abstract Accurate traffic identification is the keystone of network management, security diagnosis and application prediction analysis. Aiming at the deficiencies of supervised and unsupervised classified methods, we present a novel scheme called semi-supervised internet traffic identification based on affinity propagation (AP). In order to circumvent the problem of choosing initial points, the method introduces affinity propagation clustering to construct classification model simply and effectively. Based on the idea of semi-supervised learning, a few restrictions of labelled flows and priori manifold distribution of sampled space are abstracted. Also, manifold similarity is defined. Henceforth, the semi-supervised method can not only largely reduce the complexity of marking sampled flows, but also nicely improve the performance of the classifier. Theoretical analysis and experimental results show that the algorithm can achieve higher accuracy and better aggregation.

Key words Traffic identification, semi-supervised learning, affinity propagation (AP) clustering, manifold similarity

Citation Zhang Zhen, Wang Bin-Qiang, Li Xiang-Tao, Huang Wan-Wei. Semi-supervised traffic identification based on affinity propagation. *Acta Automatica Sinica*, 2013, **39**(7): 1100−1109

Everything over IP 思想推动了各种异构网络向互联网融合, 网络业务呈现规模化、差异化的发展趋势, 具体表现为: 网络业务不断推陈出新, 网络应用不断衍生出新的业务形态; P2P 业务的分布式组织结构和私有协议的封闭性导致大量非法、盗版、暴力等内容在 P2P 网络中肆意传播, 严重影响了用户的正常上网行为; 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; 网络安全协议在实际的网络应用中发挥了重要作用. 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 TCP 流或 UDP 流进行分类[1]. 它是网络 QoS 调度、安全检测、用户行为分析以及应用趋势分析的前提和基础, 通过对网络业务进行精细化分类和建模, 以便更好地掌握网络行为的基本特征, 保证网络的健康运行.

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, 然而, 随着 P2P 和被动 FTP 等新型网络业务的日益流行, 大量的随机端口被用于数据传输, 导致这种基于端口的流量分类方法被迅速淘汰[2]. 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 inspection, DPI) 已成为业界公认比

收稿日期 2012-06-15 录用日期 2012-09-19
Manuscript received June 15, 2012; accepted September 19, 2012
国家重点基础研究发展计划 (973 计划) (2012CB312901, 2012CB312905), 国家高技术研究发展计划 (863 计划) (2011AA01A103) 资助
Supported by National Basic Research Program of China (973 Program) (2012CB312901, 2012CB312905), National High Technology Research and Development Program of China (863 Program) (2011AA01A103)
本文责任编委 王聪
Recommended by Associate Editor WANG Cong
1. 国家数字交换系统工程技术研究中心 郑州 450002
1. National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002

较成熟的技术. Moore 等[3] 提出了基于特征字段的流量分类机制, 该方法通过分析载荷中的协议特征值, 来识别业务流的应用类型. 为了适应网络高速化的发展趋势, 文献 [4−5] 都不同程度对 DPI 算法进行了改进. 以上均属于 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.

基于流量统计特征的分类方法可以识别加密流量和新型业务, 它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化, 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 [6] 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 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 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, 且易受异常数据点影响; 在每个类中, 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基于贝叶斯技术的流量分类, 文献 [7] 设计了一种基于概率模型的朴素贝叶斯分类器 (Naive Bayesian classifier, NBC), 该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, 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[8] 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.

以上介绍的基于机器学习流量分类算法主要分为两大类:

1) 基于无监督学习的流量分类. 该方法只根据网络流的相似程度进行流量分类, 不必事先知道任何样本空间的先验信息, K-means 流量分类方法属于此范畴. 此类方法能识别出新型业务, 但是不能自动给出流量的类别标签, 且分类准确性相对较低.

2) 基于完全监督学习的流量分类. 该方法是在已知业务类别的样本空间中进行训练, 需要获知样本的类别标签, NBC 及其改进方法属于此范畴. 该方法虽然分类准确性较高, 但是不能识别新型业务, 并且获取足够多的类别标记数据是代价高昂的、难以实现的.

针对完全监督方法计算复杂度较高的缺陷, 文献 [9] 首次使用半监督学习方法分类网络流量. 该方法应用 K-means 算法将大量未标记样本和少量标记样本混合的训练集, 聚类成若干个不相交的簇, 然后, 使用标记的样本完成簇与类别之间的映射. 虽然半监督的 K-means 能够降低标记样本的复杂度, 但是其分类准确性在 70 % 左右. 为了提高分类算法的性能, 本文提出一种基于近邻传播学习的半监督流量分类机制 (Semi-supervised traffic identification based on affinity propagation, STI-AP). 该方法具有以下特点: 1) 相比 K-means、贝叶斯分类器, 基于 “近邻传播学习” 构建的分类器, 不仅能识别未知业务而且实现简单、处理高效; 2) 引入半监督学习机制, 利用大多数无标号样本流和少量有标号样本流, 大大降低了标记流量样本的计算复杂度; 3) 利用少量标记样本指导分类器的构建, 能够自动获知业务类别标签, 提高分类准确性; 4) 通过已标记样本约束和流形空间的先验信息假设, 改进了样本流之间的相似性度量, 能够识别具有复杂结构的数据流集合, 提高了流量分类器的性能.

2 近邻传播学习原理

已知类型标签集合 $L = \{L_1, \cdots, L_m\}$ 和网络流样本集合 $X = \{X_1, \cdots, X_n\}$, 其中, 网络流 X_i 是一个由 k 个网络流统计特征构成的属性向量 $(A_1^i, \cdots, A_k^i)^{\mathrm{T}}$. 基于近邻传播进行流量分类, 就是要训练学习得到一个多元分类模型 f: $X \rightarrow L$, 并以此对类型未知的网络流进行实时分类.

Frey 等[10] 于 2007 年在 *Science* 上首次提出近邻传播 (Affinity propagation, AP) 学习算法, 该算法是一种基于近邻信息传播的无监督聚类算法, 其目的找到最优的类代表点集合 (类代表点对应为某个样本点), 使得所有数据点到最近的类代表点的相似度之和最大. 与 K-means 相比, AP 算法具有以下优势: 1) 将所有的数据点作为候选类代表点, 避免了聚类结果受限于初始类代表点的选择; 2) 基于相似度信息的传播来优化目标函数, 实现简单、计算高效; 3) 对数据点的相似度没有对称性的要求, 扩大了 AP 算法的应用范围[11−13].

AP 算法是在 n 样本点的相似度矩阵上进行学

习聚类的. 首先, 将 n 样本点都视为候选类代表点(即潜在的聚类中心), 并为每个点建立与其他样本点的吸引程度 $s(i,j)=-\|X_i-X_j\|^2\,(i\neq j)$, 形成一个 $n\times n$ 的相似度矩阵 $S_{n\times n}$. 其中, $s(i,j)$ 表示数据点 X_j 适合作为 X_i 的类代表点的程度, $s(i,j)=0$ 代表 X_i 和 X_j 有最大相似性; 相反, $s(i,j)=-\infty$ 代表 X_i 和 X_j 属于不同类别. 另外, AP 算法要为每个点 X_i 设定其偏向参数 $s(i,i)$, 作为数据点 X_i 被选作类代表点的倾向性, $s(i,i)$ 的值越大, 对应数据点 X_i 被选中作为类代表点的可能性越大. AP 算法初始设定所有 $s(i,i)$ 为相同值, 通过改变 $s(i,i)$ 值来寻找合适的类数.

为了进行数据点的相似度信息传播, AP 算法引入了两个重要信息量参数, 分别定义为 "吸引度 $r(i,j)$" 和 "归属度 $a(i,j)$". $r(i,j)$ 是从 X_i 指向 X_j, 表示 X_j 适合作为 X_i 的类代表点的程度; $a(i,j)$ 是从 X_j 指向 X_i, 表示 X_i 选择 X_j 作为其类代表点的合适程度, 图 1 给出了 $r(i,j)$ 和 $a(i,j)$ 的信息量传播示意图, $r(i,j)$ 和 $a(i,j)$ 越大, X_j 作为 X_i 的类代表点的可能性越大.

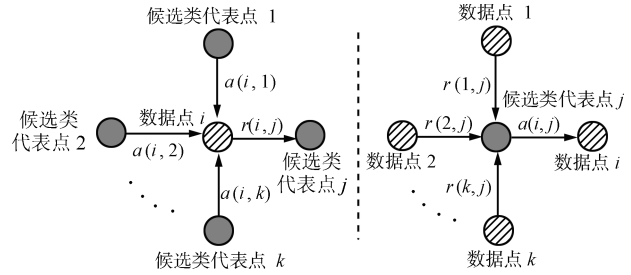


图 1 信息量传播示意图

Fig. 1 Sketch map of information propagation

AP 算法的核心步骤就是两个信息量 $r(i,j)$ 和 $a(i,j)$ 交替更新的过程, 算法初始阶段, $r(i,j)$ 和 $a(i,j)$ 都设为 0, 两个信息的更新过程如下:

$$r^{(t)}(i,j)\leftarrow \lambda r^{(t-1)}(i,j)+(1-\lambda)\{s(i,j)-\sum_{k\neq j}\max(a^{(t-1)}(i,k)+s(i,k))\} \tag{1}$$

$$a^{(t)}(i,j)=\begin{cases}\lambda a^{(t-1)}(i,j)+\\ \quad(1-\lambda)\min\{0,r^{(t-1)}(j,j)+\\ \quad\sum\limits_{i'\neq i,k}\max(0,r^{(t-1)}(i',j))\}\,,\ i\neq j\\ \lambda a^{(t-1)}(i,j)+(1-\lambda)\times\\ \quad\sum\limits_{i'\neq j}\max\left(0,r^{(t-1)}(i',j)\right),\ \ i=j\end{cases} \tag{2}$$

其中, 式 (1) 是吸引度 $r(i,j)$ 的表达式, 它等于数据点 X_i 和候选类代表点 X_j 的相似度减去其他点和 X_j 相似度的最大值, 其计算方式是数据点 X_i 从其他候选类代表点搜集的证据信息; 式 (2) 是归属度 $a(i,j)$ 的表达式, 它等于除 X_i 之外的其他点到 X_j 的吸引度之和, 其计算方式是候选类代表点 X_j 从其他数据点搜集的证据信息.

另外, AP 算法在信息更新步骤中引入了另一个重要的参数 λ, 称为阻尼因子 $(0\leq\lambda<1)$. 在每一次循环迭代中, $r(i,j)$ 和 $a(i,j)$ 的更新结果都是由当前迭代过程中更新的值和上一步迭代的结果加权得到的, 目的是改进算法收敛性、避免迭代过程中出现数值震荡. 迭代终止的条件满足以下其中之一即可: 超过某一迭代最大数目; 信息改变量低于某一固定阈值; 选择的类代表点在迭代过程中保持稳定. 迭代完成后, 对任意 X_i, 计算满足条件 $\arg\max_j\left(a^{(t)}(i,j)+r^{(t)}(i,j)\right)$ 的 X_j, 并将其作为 X_i 的类代表点.

从 AP 算法的学习过程可以看出: AP 算法是一种连续优化的过程, 不受初始点选择的困扰; 只需要进行简单的局部计算, 能够在更短的 CPU 运算时间里达到更好的聚类效果. 但是, AP 是一种无监督的学习方法, 没有考虑应用的先验信息和背景知识, 分类性能有待提高; 比较适合处理超球形结构的数据聚类问题, 当数据集分布松散或结构复杂时, 不能给出理想的聚类结果.

3 获取样本流先验信息

STI-AP 算法通过构造先验信息, 对近邻传播学习进行指导训练, 创建更加符合实际网络环境、准确率更高的流量分类器. STI-AP 应用的先验信息主要包括: 少量的已知标签 IP 流约束, 即属于同种类型的已知标签 IP 流有较大的相似性, 反之, 则相似度应达到最小; 流形空间的先验假设, 即流量样本空间的整体分布结构具有流形分布特性.

3.1 已知标签样本流约束

STI-AP 算法将已知类别标签的 IP 样本流转化为等价的成对点约束信息, 其中, 成对点约束分为两种[14]: Must-link, 即两个 IP 样本流 X_i, X_j 必须属于同一类, 表示为 $(X_i,X_j)\in$ Mustlink; Cannot-link, 限制规定两个 IP 样本流 X_i, X_j 不属于同一类别, 表示为 $(X_i,X_j)\in$ Cannotlink. STI-AP 将成对点约束转化为样本相似度的核心过程如下: 当约束对 $(X_i,X_j)\in$ Mustlink, 则认为样本流 X_i, X_j 具有很高的相似度, 因此, 令 $s(i,j)=0$; 当约束对 $(X_i,X_j)\in$ Cannotlink, 则认为样本流 X_i, X_j 具有最低相似度, 并令 $s(i,j)=-\infty$. 除了将已知的成对点约束转化为样本流相似度外, STI-AP 还利用

约束的二值传递关系[15], 对其他点进行调整, 进而得到更多的同种类型和不同类型的 IP 流约束信息. 相似度矩阵的具体调整步骤如下:

步骤 1. 对先验信息中已有的 Must-link 约束 $(X_i, X_j) \in$ Mustlink, 对应的两个样本点的相似度做出如下调整:

$$(X_i, X_j) \in \text{Cannotlink} \Rightarrow \\ s(i,j) = -\infty,\ s(j,i) = -\infty$$

由已知的 Must-link 约束扩展得到新的约束, 即若存在样本流 X_k, 满足 $(X_i, X_k) \in$ Mustlink 约束, 则:

$$\begin{cases} (X_i, X_j) \in \text{Mustlink} \\ (X_i, X_k) \in \text{Mustlink} \end{cases} \Rightarrow \\ (X_j, X_k) \in \text{Mustlink} \Rightarrow \\ s(j,k) = 0,\ s(k,j) = 0$$

步骤 2. 对与先验信息中已有的 Cannot-link 约束 $(X_i, X_j) \in$ Cannotlink, 对应的两个样本点的相似度做出如下调整:

$$(X_i, X_j) \in \text{Cannotlink} \Rightarrow \\ s(i,j) = -\infty,\ s(j,i) = -\infty$$

由已知的 Cannot-link 约束扩展得到新的约束, 即若存在样本流 X_k, 满足 $(X_i, X_k) \in$ Mustlink 约束, 则:

$$\begin{cases} (X_i, X_j) \in \text{Cannotlink} \\ (X_i, X_k) \in \text{Mustlink} \end{cases} \Rightarrow \\ (X_j, X_k) \in \text{Cannotlink} \Rightarrow \\ s(j,k) = -\infty,\ s(k,j) = -\infty$$

基于上述调整, 样本流的相似度矩阵 $S_{n\times n}$ 将会得到很大改善. 经过 AP 算法的信息量迭代, 使得同类样本流之间的吸引力最大而被尽可能划归为一类, 不同类别样本流之间的吸引力最小而被强制拆开, 进而能够得到更加符合实际的、性能更好的分类器.

3.2　流形空间的先验信息

传统的欧氏距离仅能反映样本空间的局部分布特征, 不能反映复杂形状数据集的全局一致性特征. 如图 2 所示, 在欧氏距离测度下, 相比样本点 C, 样本点 A 和点 B 更近, 但是, A 和 B 却不在同一空间分布中. 而流形假设则是从样本空间的整体分布来发现数据集的内在分布规律, 流形假设来源于 "流形理论", 其主要目标是根据样本的分布来发现隐含其中的流形结构. 文献 [16] 首先从认知上讨论了流形学习, 认为人的认知过程是基于稳定状态的流形和拓扑连续性的视觉记忆的过程, 也有大量文献将流行概念用于数据分析, 对非线性流形的线性结构进行挖掘, 然后, 利用线性分析技术对非线性进行分析, 从而降低原问题的学习难度[17−21].

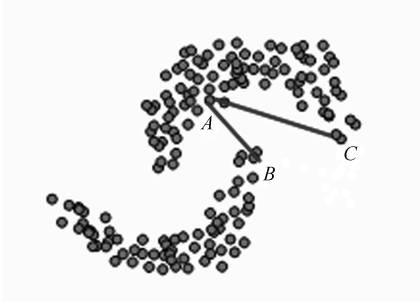


图 2　基于欧氏距离的样本空间分布示意图

Fig. 2　Sketch map of space distribution based on Euclid distance

由于不同的特征提取方法会造成在表示空间中离得很近的两个点, 并不一定是在事物本身所处空间中离得很近, 所以引入流形假设是非常必要的. 在半监督流量分类中, 由于样本空间中占绝对比重的是未知标签 IP 流, 所以流形假设实质上体现的是未知类别 IP 流的空间分布先验信息. 下面给出流形假设的定义:

定义 1 (流形假设)[22]. 在流形中互相靠近的点最可能属于同一类别. 即同一类别内的数据趋向于分布在一个密度比较高的区域, 而不同类别之间存在一个数据分布稀疏的低密度区域.

在流形假设的前提下, 需要定义一种更加合理的相似度来表达样本流之间的关系: 将每个样本 IP 流看作是一个加权无向图 $G = (V, E)$ 的顶点 V, 边集合 E 表示样本流之间的相似度, 如果两个样本由一条穿过高密度区域的路径相连接, 则这两个样本 IP 流将被赋予较高的相似度, 否则, 将被赋予较低的相似度. 基于此原则, 给出如下定义:

定义 2 (ε-近邻距离). 在样本流量空间中, 构建一个加权无向图 $G = (V, E)$, 其中, V 是顶点集, 每个顶点对应一个样本 IP 流, E 是边集. 给定任意样本流 X_i, 并定义其 ε-近邻距离为

$$D_\varepsilon(i,j) = \begin{cases} \theta_1^{\text{dist}(i,j)} - 1, & \text{dist}(i,j) \le \varepsilon \\ \theta_2^{\text{dist}(i,j)} - 1, & \text{dist}(i,j) > \varepsilon \end{cases} \tag{3}$$

其中, $\text{dist}(i,j)$ 表示两个样本 IP 流 X_i 和 X_j 的欧氏距离, θ_1 和 θ_2 是密度调节因子. 计算某样本流小于 ε 距离的近邻点时, 设定 $\theta_1 < 1$, 相反, 对与距离

较远的样本点, 令 $\theta_2 > 1$; 减去常数 1, 目的是满足当 $\mathrm{dist}(i,j) = 0$ 时, $D_\varepsilon(i,j) = 0$.

定义 ε-近邻距离的目的是: 缩小靠近高密度区域的样本点之间的距离, 放大那些穿过低密度区域的样本点之间的距离. 但是, 若只用 ε-近邻距离来衡量两个样本点的相似度, 有可能导致处于同一流形的两个样本点的距离被拉伸, 即: 若 X_i, X_j, X_k 处于同一流形中 (即同一类中), 并同时满足 $\mathrm{dist}(i,j) \le \varepsilon$, $\mathrm{dist}(j,k) \le \varepsilon, \mathrm{dist}(i,k) > \varepsilon$, 用式 (5) 计算, 使 X_i, X_k 的 ε-近邻距离被拉伸, 可能会得到结果 $\mathrm{dist}(i,j) + \mathrm{dist}(j,k) \le \mathrm{dist}(i,k)$, 导致 X_i, X_k 会被误分为两种类别. 因此, 需要进一步可以给出流形相似度的定义:

定义 3 (流形相似度). 在加权无向图 $G = (V, E)$ 中, 令 $P(i,j)$ 表示连接样本流 X_i, X_j 的所有路径集合, p 为其中任意一条路径, p_k 为路径上的第 k 个样本流. 则根据最短路径来度量流形相似度:

$$S(i,j) = -\min_{p \in P(i,j)} \sum_{k=1}^{|p|-1} D_\varepsilon(p_k, p_{k+1}) \qquad (4)$$

流形相似度是沿着流形上的最短路径计算得到的, 使得位于同一高密度区域的样本 IP 流可用许多较短的边相连, 而位于不同密度的点要用穿过低密度区域的较长边相连. 其最终目的是: 最大可能地将同一流形中互相靠近的样本 IP 流判为同一类别, 不同流形的样本 IP 流被判为不同类别. 如图 3 所示, 在流形相似度下, 图 2 的非凸空间分布变得更加紧致和规律, 同一空间结构上的点相似性更大, 因此, 基于流形相似度的测度, 真实地反映了空间分布的先验信息.

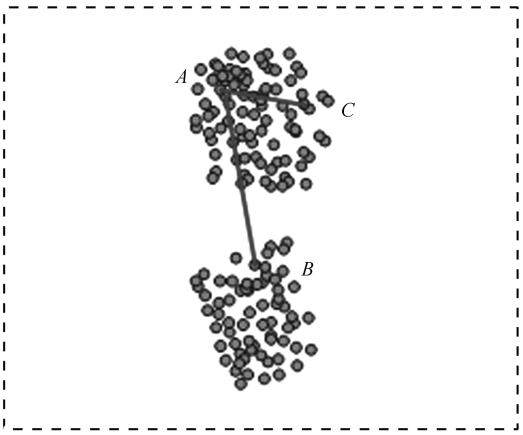


图 3 转化为 "流形相似度" 后的空间分布示意图

Fig. 3 Sketch map of space distribution based on manifold similarity

4 基于 AP 的半监督流量分类

4.1 构造半监督流量分类器

步骤 1. 基于先验信息的 AP 聚类学习

为了得到流形相似度矩阵 $S_{n\times n}$, 首先, 要计算任意两个 IP 流的欧氏距离 $\mathrm{dist}(i,j) = \sqrt{\|X_i - X_j\|^2}$, 并存储在 $n \times n$ 的矩阵 E 中, 另外, 令对角线元素 $E_{ii} = w$, w 为偏向参数; 基于第 3.1 节中的已知样本流约束对矩阵 E 进行调整, 得到矩阵 $F_{n\times n}$; 根据流形先验约束, 由式 (3) 和 (4) 计算任意两点的 ε-近邻距离和流形相似度, 并最终得到相似度矩阵 $S_{n\times n}$. 图 4 给出了计算相似度矩阵的流程.

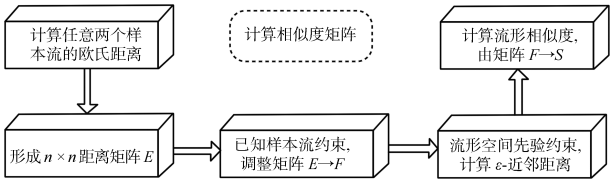


图 4 计算相似度矩阵的流程

Fig. 4 Flow chart of computing similarity matrix

最后, 基于 AP 算法进行信息量迭代学习, 迭代终止的条件为: 信息改变量 $\Delta r(i,j) = r^{(t+1)}(i,j) - r^{(t)}(i,j)$ 和 $\Delta a(i,j) = a^{(t+1)}(i,j) - a^{(t)}(i,j)$ 低于某一固定阈值.

步骤 2. 对聚簇进行标签映射

AP 聚类学习得到的结果是未知类别的聚簇集合 $C = \{C_1, \cdots, C_q\}$ 和对应的类代表点集合 $\gamma = \{\gamma_1, \cdots, \gamma_q\}$. 流量识别过程需要确定聚簇集合 C 和类别集合 $L = \{L_1, \cdots, L_m\}$ 的映射关系. STI-AP 算法采用条件概率 $P(l = L_j | C_i)$ 来表示在聚簇 C_i 中任意样本点属于类别标签 L_j 的概率, 并且利用训练样本集对其进行估计, 其中, $i \in [1,q], j \in [1,m]$.

令 N_i^j 表示在聚簇 C_i 中属于类别 L_j 的样本流总数, N_i 表示聚簇 C_i 中所有的样本流总数, 则根据最大似然估计, 可得条件概率的估计表达式为: $\hat{P}(l = L_j | C_i) = N_i^j / N_i$. 基于此, 可得到聚簇的标签映射函数为

$$l = \arg\max_{j=1,\cdots,m} \hat{P}(l = L_j | C_i) \qquad (5)$$

映射函数就是将聚簇中包含最多样本流的类别标签赋给该簇. 若某聚簇中, 不包括任何已标签样本, 则认定其为 "未知流量类型".

步骤 3. 重构流量分类器

流量识别阶段, 对于任意到达的真实流 X_i, 将其分配给距离最近的类代表点所在聚簇. 即:

$$\gamma_j = \arg \min_{j=1,\cdots,q} \text{dist}\,(X_i, \gamma_j) \tag{6}$$

其中, $\text{dist}\,(X_i, \gamma_j)$ 代表业务流 X_i 与类代表点的欧氏距离. 为了检测新型业务, 需要在识别过程中对应创建新的子簇. 设定最大聚簇半径 $R_{\max} = \max\{R_1, R_2, \cdots, R_j, \cdots, R_q\}$, 其中, R_j 表示聚簇 C_j 的半径. 若对任意类代表点 γ_k, 有 $\|X_i - \gamma_k\| > R_{\max}$ 成立, 则为业务流 X_i 创建新的聚簇. 当出现较多新的聚簇或者业务流时, 表明现有的网络环境出现了一些新的流量类型, 需要对分类器进行重新学习, 以适应网络流量模式的动态变化. STI-AP 算法采用的重构流量分类器原则是: 新型业务流总数大于某一阈值 TH.

4.2　流量分类算法流程

如图 5 所示, STI-AP 算法主要包括两个步骤: 近邻传播学习和流量实时识别. 在近邻传播学习阶段, 针对已标记训练样本集合, 利用 AP 算法学习得到流量分类器. 实时识别阶段, 将新到的 IP 流特征与学习阶段得到的流量分类器比较并输出识别结果, 对出现的新型业务类别, 要进行反馈和重新学习.

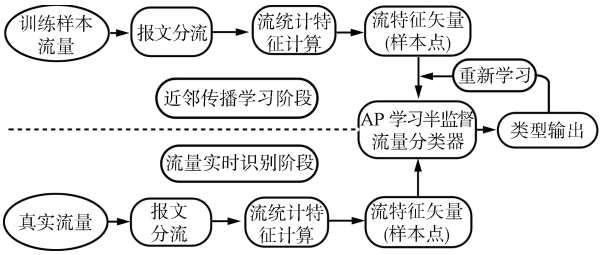


图 5　基于 AP 的半监督流量分类模型

Fig. 5　Model of semi-supervised traffic classification based on affinity propagation

图 6 给出了 STI-AP 分类器的学习过程示意图, 其中, STI-AP 算法的半监督学习主要体现在: 利用少量的已标记样本流转化为成对点约束, 修改距离测度; 获取绝大多数未知样本的流形空间分布先验假设, 构成流形相似度矩阵; 利用已知样本流标签, 进行聚簇的标签映射. 另外, 需要判断新型业务流的个数是否大于阈值 TH, 若大于 TH, 返回重新获取新的样本数据进行学习; 否则, 创建新的聚簇.

4.3　分类器凝聚性分析

首先考虑样本质心 $\hat{\mu}$ 的凝聚性. 设某聚簇对应的类代表点为 F^*, 该聚簇对应的样本流集合为 $F = \{F_1, \cdots, F_i, \cdots, F_l\}$, 并令质心 $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} F_i$, 由中心极限定理可得引理 1.

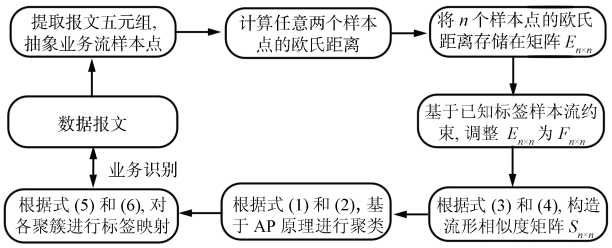


图 6　STI-AP 分类器的学习过程

Fig. 6　Learning prodedure of STI-AP classifier

引理 1.　当样本流的个数 l 趋于非常大时, $\hat{\mu} = \frac{1}{l}\sum_{i=1}^{l} F_i$ 服从正态分布 $\text{N}(\mu, \sigma^2/l)$, 其中, $\text{E}(F_i) = \mu$ 为均值, $\text{Var}\,(F_i) = \sigma^2$ 为方差.

引理 2.　令 $\varepsilon = F_i - \hat{\mu}$, 变量 ε 服从正态分布 $\text{N}\,(0, \sigma^2 - \sigma^2/l)$.

证明.　由引理 1, 易得均值 $\text{E}\,(\varepsilon) = \text{E}\,(F_i) - \text{E}\,(\hat{\mu}) = \mu - \mu = 0$. 另外可得方差: $\text{Var}\,(\varepsilon) = \text{Var}\,(F_i - \hat{\mu}) = \text{Var}\,(F_i) - 2\text{Cov}\,(F_i, \hat{\mu}) + \text{Var}\,(\hat{\mu})$.

又 $\text{Var}\,(F_i) = \sigma^2$, $\text{Var}\,(\hat{\mu}) = \sigma^2/l$, $\text{Cov}\,(F_i, \hat{\mu}) = 2\,(\text{E}\,(F_i\hat{\mu}) - \mu^2)$, 其中

$$\begin{aligned}
&\text{E}\,(F_i\hat{\mu}) = \text{E}\left(F_i \frac{1}{l}\sum_{i=1}^{l} F_i\right) = \\
&\frac{1}{l}\left[\text{E}\left(F_i^2\right) + \sum_{i=1, i\neq j}^{l} \text{E}\,(F_i)\,\text{E}\,(F_j)\right] = \\
&\frac{1}{l}\left(\mu^2 + \sigma^2 + (l-1)\mu^2\right) = \\
&\mu^2 + \frac{\sigma^2}{l}
\end{aligned}$$

因此, $\text{Cov}\,(F_i, \hat{\mu}) = 2\sigma^2/l$, $\text{Var}\,(\varepsilon) = \sigma^2 - \sigma^2/l$. □

引理 3.　对于任意大于 0 的常数 a, 有以下概率不等式成立: $P_\varepsilon\,(a) = P\,(|\varepsilon| \geq a) \leq (\sigma^2 - \sigma^2/l)/a^2$.

证明.　根据引理 2, 由契比雪夫不等式[23] 可得: 对与任意的 $a > 0$, 不等式 $P\,(|\varepsilon| \geq a) \leq (\sigma^2 - \sigma^2/l)/a^2$ 成立, 其中, $|\varepsilon|$ 表示样本流 F_i 到质心 $\hat{\mu}$ 的距离. □

当样本数量 l 趋于非常大时, 类代表点 F^* 趋于质心 $\hat{\mu}$, 即 $P\,(|F_i - F^*| \geq a) \rightarrow P_\varepsilon\,(a) = P\,(|F_i - \hat{\mu}| \geq a)$. 因此, 可以用概率 $P_\varepsilon\,(a)$ 来衡量聚簇的凝聚性. 引理 3 给出了 $P_\varepsilon\,(a)$ 的上界, 可以看出: 聚簇的凝聚性与 a^2 成反比, 大部分的样本点能够以较高的概率围绕在 F^* 的周围. 因此, 类似于 K-mean 算法的质心, STI-AP 的类代表点也能很好地诠释整个聚簇.

5 实验结果及分析

基于图 5 的流量分类模型, STI-AP 算法的性能评价将在训练样本流量和检验样本流量中进行: 在训练样本流量中构造分类器; 在检验流量中运行流量分类模型, 并根据运行结果来评估分类器的性能. 在实验仿真中, 分别选取无监督和完全监督分类方法中应用广泛的 K-means 方法[6]、NBC[7] 及 "NBC+KE+FCBF"[8], 并结合应用单纯的 AP 算法与 STI-AP 算法进行仿真对比.

5.1 算法评价指标

实验中采用以下三种评价指标:

定义 4 (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=N_i'/N_i, 其中, N_i' 是被正确分类为 C_i 的样本流总数, N_i 等于被分类为 C_i 的样本流总数. 则分类器的整体准确率为 $P_{\text{all}} = \sum_{i=1}^{q} N_i' / \sum_{i=1}^{q} N_i$.

定义 5 (F-measure 指标). F-measure 指标沿用了检索领域的准确性指标, 是常用的分类器评价指标. F-measure 是将 "准确率" 和 "召回率" 两个指标组合形成的, 其中, "召回率" 的定义如下: $R_i = N_i'/N_{C_i}$, 其中 N_{C_i} 为未分类前属于 C_i 的样本流总数. 则聚簇 C_i 的 F-measure 指标定义为 F-measure$(i) = 2P_iR_i/P_i + R_i$, 经平均后, 得到分类器的 F-measure 指标为 F-measure= $\frac{1}{q}\sum_{i=1}^{q}$ F-measure(i). F-measure 的取值范围是 $[0, 1]$, F-measure 值较大时表示 "准确率" 和 "召回率" 都比较高, 算法也就越准确.

定义 6 (误差平方和). 误差平方和常作为构建分类器的目标函数, 常用来表示分类器的失真度或凝聚性. 误差平方和等于所有样本流到其类代表点的距离平方和, 即: $SSE = \sum_{k=1}^{q} \sum_{X_i \in C_k} \text{dist}\,(X_i, \gamma_k)^2$.

5.2 实验数据说明

为了便于对比仿真, 本文采用文献 [8] 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时间段, 且每个数据集包含 28 分钟内, 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 数据集可以直接从网站[24] 中下载获取, 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.

5.3 算法仿真比较

利用 Moore set 数据集, 分别从整体准确率、F-measure 指标、误差平方和以及计算复杂度对算法进行实验仿真. 针对 K-means 方法、NBC、NBC+KE+FCBF、AP 方法, 采用 Weka 中已有的软件包直接进行仿真, STI-AP 算法则使用 Matlab 软件编程实现. 在 K-means 算法中, 初始化聚类中心的个数 $K = 10$; 在算法 STI-AP 中, 初始化信息量 $r\,(i, j) = a\,(i, j) = 0$, ε-近邻距离的伸缩因子 $\theta_1 = 1/2$、$\theta_2 = 2$, 阻尼系数 $\lambda = 0.9$, 重构分类器阈值 $TH = 200$. 设定偏向参数 w 等于所有样本流的流形相似度的平均值, 即 $w = \sum_{i=1}^{n} \sum_{j=1}^{n} S\,(i, j)/n^2$.

表 1 Moore set 数据集详细信息

Table 1 Detailed information of Moore set

流量类型	具体应用	IP 流个数
WWW	HTTP, HTTPS	32 8091
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

1) 整体准确率仿真比较

在整体准确率的实验中, 具体过程如下: 针对每种流量类型, 从 “Moore set 1” 中随机抽取 α 个 IP 数据流作为训练样本, 则样本集的总大小为 10α; 在分类测试中, 使用剩余的数据集 “Moore set 2 ∼ 10” 作为测试样本, 每个数据集各进行分类评估测试, 可得到 9 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 9 个不同时段, 能够一定程度上反映不同算法对流量的依赖程度.

在 $\alpha = 100, 200, 300, 400, 500$ 时, 表 2 给出了 4 种算法整体准确率的比较, 可以看出: K-means、AP、STI-AP 算法随着训练样本集的增大而增大; NBC 以及 “NBC + KE + FCBF” 方法的整体准确率出现了抖动; “NBC + KE + FCBF”、STI-AP 算法的整体准确率要明显高于 K-means、NBC 和 AP, 并且 STI-AP 算法的分类准确率最高. 其主要原因: K-means 和 AP 属于无监督的分类方法, 没有考虑先验信息和背景知识, 当样本流分布较为复杂时, 分类准确性较低; NBC 和 “NBC + KE + FCBF” 均属于完全监督学习算法, 对流量的变化较为敏感, 所以准确率出现了抖动; STI-AP 合理地利用了少量先验信息, 对样本集进行半监督指导训练, 避免了随流量的变化而导致的准确率下降.

2) F-measure 指标仿真比较

针对每种流量类型, 从 Moore set 1 中随机抽取固定的 α 个 IP 数据流作为训练样本, 则样本集的总大小为 10α; 在分类评估中, 使用 Moore set 2 作为测试样本, 分别得到 “准确率” 和 “召回率”, 并最终求得 F-measure 值. 图 7 给出了 α 取不同值时, F-measure 的仿真曲线图, 可以看出: “NBC + KE + FCBF” 通过使用 “核估计” 和 “特征过滤” 的方法, 其分类器性能比 NBC 有了一定的提升; K-means、AP 以及 NBC 分类性能相对较低; STP-AP 算法通过利用仅有的少量先验信息, 对近邻学习进行指导训练, 构造了更加实际网络环境的、准确率更高的流量分类器.

3) SSE 仿真比较

类似于 F-measure, 针对每种流量类型的不同样本数量, 图 8 仿真了 4 种分类算法的误差平方和 SSE. 从图 8 中可以看出: 随着样本数量的增加, 4 种算法的 SSE 都有所降低; K-means、NBC 的失真

表 2　整体准确率的仿真 (%)

Table 2　Simulation of integral accuracy (%)

样本集大小	K-means	NBC	NBC + KE + FCBF	AP	STI-AP
100 个 IP 流/类型	55.91	60.39	71.64	56.33	71.70
200 个 IP 流/类型	59.27	56.85	80.03	60.28	80.69
300 个 IP 流/类型	61.34	67.72	75.39	63.87	85.43
400 个 IP 流/类型	66.08	74.52	84.21	69.74	88.76
500 个 IP 流/类型	70.42	69.37	90.50	74.98	92.92

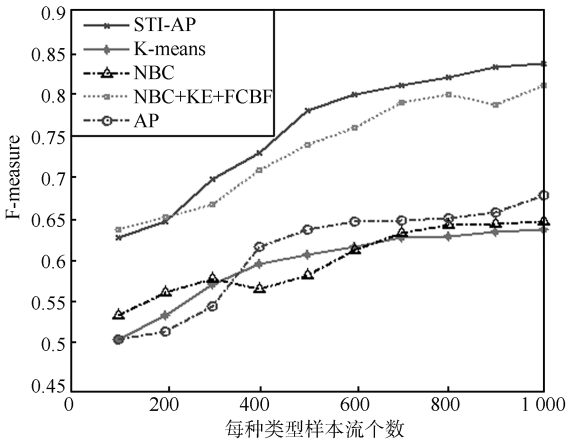


图 7　F-measure 仿真图

Fig. 7　Simulated figure of F-measure

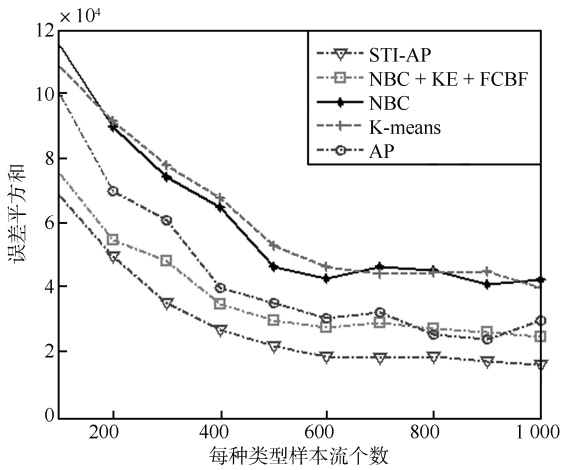


图 8　误差平方和仿真曲线图

Fig. 8　Simulated figure of SSE

度较高, AP 和 "NBC + KE + FCBF" 有了一定改善, STI-AP 的失真度最小, 这表明基于近邻传播学习构造的半监督分类器具有很好的凝聚性能, 即聚簇中的类代表点能较好地代表整个类.

6 结论

针对完全监督和无监督学习分类器的缺陷, 本文提出一种基于近邻传播学习的半监督流量分类器. 通过引入 AP 算法, 使得聚类过程实现简单且不受初始点选择的困扰. 同时, 应用 "半监督学习" 的思想, 利用少量已标记样本流, 抽象出成对点约束, 修改样本流之间的距离测度; 通过 ε-近邻距离的伸缩机制, 获取了大多数无标号样本流的空间分布先验信息, 构造了流形相似度的距离测度, 提高了分类器的准确性和凝聚性能. STI-AP 算法使用了网络流样本集合中的 249 种网络流统计特征, 增加了分类模型的计算开销, 如何从现有属性集合中快速提取具有代表性的属性特征, 提高分类模型的计算性能将是未来的主要工作.

References

1 Yang Jia-Hai, Wu Jian-Ping, An Chang-Qing. *Internet Measurement Theory and Its Applications*. Beijing: Post & Telecom Press, 2009. 383−408
(杨家海, 吴建平, 安常青. 互联网络测量理论与应用. 北京: 人民邮电出版社, 2009. 383−408)

2 Karagiannis T, Broido A, Faloutsos M, Claffy K C. Transport layer identification of P2P traffic. In: Proceedings of the 4th ACM SIGCOMM on Internet Measurement. New York, USA: ACM, 2004. 121−134

3 Moore A W, Papagiannaki K. Toward the accurate identification of network applications. In: Proceedings of the 2005 Passive and Active Network Measurement. Boston, MA: Springer, 2005: 41−54

4 Antonello R, Fernandes S, Sadok D, Kelner J. Characterizing signature sets for testing DPI systems. In: Proceedings of the 2011 IEEE GLOBECOM Workshops. Houston, TX: IEEE, 2011. 678−683

5 Santos A, Fernandes S, Antonello R, Szabo G, Lopes P, Sadok D. High-performance traffic workload architecture for testing DPI systems. In: Proceedings of the 2011 IEEE Global Telecommunications Conference (GLOBECOM 2011). Houston, TX: IEEE, 2011. 1−5

6 Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning. In: Proceedings of the 30th IEEE Conference on Local Computer Networks. Sydney, Australia: IEEE, 2005. 250−257

7 Roughan M, Sen S, Spatscheck O, Duffield N. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification. In: Proceedings of the 4th ACM SIGCOMM Internet Measurement Conference. Taormina, Sicily, Italy: ACM, 2004. 135−148

8 Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. In: Proceedings of the 2005 Internet Traffic Classification Using Bayesian Analysis Techniques (SIGMETRICS). Alberta, Canada: ACM, 2005. 50−60

9 Erman J, Mahanti A, Arlitt M, Cohen I, Williamson C. Offline/realtime traffic classification using semi-supervised learning. *Performance Evaluation*, 2007, **64**(9−12): 1194−1213

10 Frey B J, Dueck D. Clustering by passing messages between data points. *Science*, 2007, **315**(5814): 972−976

11 Zhang J, Tuo X G, Yuan Z, Chen H F. Analysis of fMRI data using an integrated principal component analysis and supervised affinity propagation clustering approach. *IEEE Transactions on Biomedical Engineering*, 2011, **58**(11): 3184−3196

12 He Y C, Chen Q C, Wang X L, Xu R F, Bai X H, Meng X J. An adaptive affinity propagation document clustering. In: Proceedings of the 7th International Conference on Information and System. Cairo, Egypt: IEEE, 2010. 1−7

13 Liu H W. Community detection by affinity propagation with various similarity measures. In: Proceedings of the 4th International Joint Conference on Computational Sciences and Optimization. Yunnan, China: IEEE, 2011. 182−186

14 Wagstaf K, Cardie C. Clustering with instance-level constraints. In: Proceedings of the 17th International Conference on Machine Learning. Stanford, USA: Morgan Kaufmann Publishers, 2000. 1103−1110

15 Bilenko M, Basu S, Mooney R J. Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the 21st International Conference on Machine Learning. New York, USA: ACM, 2004. 81−88

16 Seung H S, Lee D D. The manifold ways of perception. *Science*, 2000, **290**(5500): 2268−2269

17 Liu Sheng-Lan, Yan De-Qin. A new global embedding algorithm. *Acta Automatica Sinica*, 2011, **37**(7): 828−835
(刘胜蓝, 闫德勤. 一种新的全局嵌入降维算法. 自动化学报, 2011, **37**(7): 828−835)

18 Zhang S W, Lei Y K. Modified locally linear discriminant embedding for plant leaf recognition. *Neurocomputing*, 2011, **74**(14−15): 2284−2290

19 Yang W K, Sun C Y, Zhang L. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition*, 2011, **44**(8): 1648−1657

20 Zhang J P, Wang X D, Krger U, Wang F Y. Principal curve algorithms for partitioning high-dimensional data spaces. *IEEE Transactions on Neural Networks*, 2011, **22**(3): 367−380

21 Yan De-Qin, Liu Sheng-Lan, Li Yan-Yan. An embedding dimension reduction algorithm based on sparse analysis. *Acta Automatica Sinica*, 2011, **37**(11): 1306−1312
(闫德勤, 刘胜蓝, 李燕燕. 一种基于稀疏嵌入分析的降维方法. 自动化学报, 2011, **37**(11): 1306−1312)

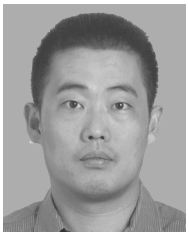
22 Thedoridis S, Koutroumbas K. *Pattern Recognition (3rd edition)*. Beijing: Publishing House of Electronics Industry, 2010. 389−407

23 Mitzenmacher M, Upfal E. *Probability and Computing: Randomized Algorithm and Probabilistic Analysis*. Cambridge, U. K.: Cambridge University Press, 2005. 44−45

24 Moore A W. Moore Set [Online], available: http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papsers/sigmetrics/index.html. 2012

张 震 国家数字交换系统工程技术研究中心博士研究生. 主要研究方向为网络测量, 网络安全, 网络管理. 本文通信作者.
E-mail: zhangzhenhigh@gmail.com
(**ZHANG Zhen** Ph. D. candidate at the National Digital Switching System Engineering and Technological Research and Development Center. His research interest covers internet measurement, network security, and network management. Corresponding author of this paper.)

汪斌强 国家数字交换系统工程技术研究中心教授. 主要研究方向为宽带信息网络. E-mail: wbq@ndsc.com.cn
(**WANG Bin-Qiang** Professor at the National Digital Switching System Engineering and Technological Research and Development Center. His main research interest is broadband network.)



李向涛 国家数字交换系统工程技术研究中心讲师. 主要研究方向为网络管理. E-mail: lxt@ndsc.com.cn
(**LI Xiang-Tao** Lecturer at the National Digital Switching System Engineering and Technological Research and Development Center. His main research interest is network security.)

黄万伟 国家数字交换系统工程技术研究中心讲师. 主要研究方向为交换技术. E-mail: hww@ndsc.com.cn
(**HUANG Wan-Wei** Lecturer at the National Digital Switching System Engineering and Technological Research and Development Center. His research interest covers routing and switching technologies.)