DOI: 10.13382/j.jemi.2014.04.006

基于半监督的网络流量分类识别算法*

周文刚[1,2] 陈雷霆[1] Lubomir Bic[2] 董 仕[3]
(1. 电子科技大学计算机科学与工程学院 成都 610054; 2. University of California, Irvine USA 92617;
3. 东南大学计算机科学与工程学院 南京 211189)

摘 要: 近年来,许多机器学习的方法被广泛应用于网络流量分类识别的问题中,结合有监督学习与无监督学习的特点,提出一种基于半监督学习的流量分类识别方法,该方法改进 K 均值聚类算法中初始簇中心的选取,通过基于密度因子的相似性函数来满足聚类数据的全局一致性要求以获取更适合的初始簇中心,并通过最大似然估计方法标记聚类结果实现与相关应用类型或协议的对应匹配过程,实验结果表明,该算法提升了网络流量分类识别结果的准确性和分类识别效率,能够有效满足流量分类识别的应用需求。

关键词: 网络流量; 半监督学习; 分类识别; 聚类中心点
中图分类号: TP391 文献标识码: A 国家标准学科分类代码: 510.4050

Algorithm for network traffic classification and identification based on semi-supervised learning

Zhou Wengang[1,2] Chen Leiting[1] Lubomir Bic[2] Dong Shi[3]
(1. School of Engineering and Computer Science, UESTC, Chengdu 610054, China;
2. School of Information and Computer Science, University of California, Irvine 92617, USA;
3. School of Engineering and Computer Science, Southeast University, Nanjing 211189, China)

Abstract: In recent years, many machine learning methods have been widely used in network traffic classification issues. Combining the characteristics of supervised learning and unsupervised learning, this paper proposes a network traffic classification method based on semi-supervised learning, which can improve the selection of initial cluster centers of K-means clustering algorithm. It is chosen by the similarity factor based on density function to meet the requirement for global consistency of clustering process, and more suitable initial cluster centers are selected. Then clustering results are tagged through the maximum likelihood estimation method and achieved with the correspondence matching process of relevant application or protocol type. The experimental results show that the algorithm can improve the accuracy and efficiency of network traffic classification and identification, which can effectively meet the application requirements of traffic classification.

Keywords: network traffic; semi-supervised learning; traffic classification; clustering center

1 引 言

近年来,许多机器学习的方法广泛应用于网络流量分类识别问题的处理中,这一思想通常根据网络流的特征行为和机器学习的原理,通过构建流量分类识别模型,进而判别网络流量的类型[1]。众所周知,不同的网络应用有各不相同的传输目标和特征行为,从而在网络流中体现出各不相同的网络行为模式,例如:与实时的短会话相比,使用 FTP 协议

收稿日期: 2013-09 Received Date: 2013-09
* 基金项目: “十一五”国家科技支撑计划重点资助项目“国际贸易区域经贸合作与流通促进关键支撑技术研究”(2009BAH46B03)

进行文件传输时，网络特征行为将体现为 1 个相对较长的连接周期和相对较大的数据传输规模。同样，与 FTP 数据传输相比，P2P 的传输通常是双向的[5]，而 FTP 传输一般是单向的，常用的传输层统计特征还包括：数据包的大小、数据包的传输方向、TCP 的窗口尺寸及 TCP 标记位等。通常这些网络应用的特征行为属性也作为流量测度辅助信息应用于基于机器学习的流量分类识别方法中，以提升系统分类的准确性和分类效率。

在常见的分类模式下，基于机器学习理论的有监督分类识别方法，可用于识别已知类型的网络流量，而对未知流量识别存在缺陷；基于机器学习的无监督分类识别方法可以在分类识别过程中发现新的网络应用协议类型[9]，其聚类所形成的类或簇需要与实际的网络应用类型或协议进行对应标识，这一对应过程在实际应用识别中还有提升的空间[9]；而基于机器学习的半监督分类识别方法介于上述两者方法之间，结合两者的长处，可以发现未知的流量类型，并且在结果对应匹配问题上做出相应改进，以提升系统总体分类识别性能。

对基于 2 个阶段的半监督网络流量分类识别方法提出改进，首先利用有效载荷分析方法对网络流进行识别并标记，然后对未标记的网络流量使用聚类算法进行分类识别，该方法采用 K 均值聚类算法并针对其算法中初始簇中心选取的问题，利用基于密度因子的相似性函数来选择最适当的初始簇中心，并通过最大似然估计方法标记聚类结果实现与相关应用类型或协议的对应匹配过程，实验结果表明，该算法提升了网络流量分类识别结果的准确性和分类识别效率。

2 相关研究

常见分类模式下，基于机器学习的流量分类识别方法大致可分为基于有监督学习的流量分类识别方法，基于无监督学习的流量分类识别方法和基于半监督学习的流量分类识别方法。基于有监督学习的流量分类识别方法需要提前建立分类器，然后对数据进行分类，其中典型的算法有 Navie Bayes[5]，c4.5 决策树方法[6]、支持向量机方法[7]、CBA(classification-based association rules)[8]等，而基于划分的 CLARANS 算法[9]、K-means 算法[10-11]，基于密度的 DBSCAN 算法[12]，AutoClass 算法[13]等通常应用于基于无监督学习的流量分类识别方法，以有效处理加密流量或新的未知网络应用。而基于半监督学习[14]的流量分类识别方法则是介于两者之间，所谓半监督是指首先获取少量的已知类型的数据流信息作为监督信息，然后将待识别网络数据集基于聚类方法划分成不同的类或簇后，利用前面已标记的数据信息，发现特定的对应规则，完成相关类簇的标记工作以明确网络流所承载的特定网络类型或网络协议。实际上，在一些其他的研究领域，如：图像分割[15]、视频检索[16]中，人们也引入了半监督学习的思想，将一些先验信息用于辅助和改善聚类结果，Berkhin. P 等人[17]还将 k-means 算法扩展到了分布式聚类领域。在网络流量分类识别方面，Erman 和 Mahanti 等人[18]利用基于 k-Means 的半监督学习分类识别方法对数据流进行分类，实验结果表明该方法能够达到 70% ~ 90% 的总体识别准确率，但算法中初始聚类中心的随机选择的好坏对分类结果的精度有较大影响。对于在基于无监督学习和半监督学习的分类识别问题中广泛应用的 k-means 算法，该算法力图找出 k 个聚类中心 $c_1, c_2, \ldots, c_k$，在计算每个数据与 k 个中心点的距离后，找到距离最小的聚类中心 c_v，并将其归并到该中心点所属的聚类划分，最终使得 k 个类簇的内部具有较高的相似度，而 k 个类簇之间的数据点相异度较大。这里，通常将平方偏差定义作为准则函数，且 k 值代表聚类数，其取值与最终分类识别效果密切相关。

3 算法描述

所提出半监督的分类识别算法，首先通过辅助端口信息和有效载荷分析标记部分流量类型，其所获取的标记流量用于后续聚类类簇的对应匹配过程，由于网络流量具有自相似性、长相关性和周期性等特性，其中，自相似性[19]意味着流量数据局部的结构和总体的结构在时间维度和空间维度上都具有某种程度的一致性，而长相关性[20]则表示自相似过程中的持续现象，即从过去和现在的信息中也能获得未来的相关信息。基于网络流量的自身特性，在本文中，定义数据的 2 个一致性特征。

定义 1：局部一致性：在空间和时间维度上相邻数据点之间具有较高的相似性；

定义 2：全局一致性：在同一流形上的数据点

具有较高的相似性；

接下来，利用流量数据的这些特性来辅助聚类过程，由于在基于聚类的流量识别过程中，同一类型的数据往往分布于 1 个相对比较密集的区域，不同类别之间则往往存在 1 个数据分布相对比较稀疏的区域，因此，为了充分利用网络流量数据的相关特性，引入新的密度因子来改进 k-means 算法的初始中心点的选择。在一般基于贪心算法的初始聚类中心点的搜索过程中，可能存在把孤立点当着中心点的问题，实际上，除了希望中心点在距离维度上尽量分散外，也需要考量这些中心点周围的密度分布，希望其紧密且具有一定的代表性。采用密度敏感的相似性函数可以满足聚类过程中的全局一致性要求，并依次从数据集中选取代表中心点，直到达到数值 k 为止，其具体定义和过程如下：

定义 3：已知流量样本数据 $D = \{d_1, d_2, \ldots, d_n\}$，其中对象 d_i 的密度记为 $dens(x_i)$，存在：

$$dens(d_i) = \sum_{i=1}^{n} \frac{1}{\min\limits_{l \in L_{ij}} \sum\limits_{i=1}^{|d|-1} \delta^{dist(d_i, d_{i+1})} - 1} \quad (1)$$

式中：边 $(d_i, d_{i+1}) \in E, 1 \leqslant i \leqslant |v|$，$L_{ij}$ 为连接点 d_i 与 d_j 的所有路径集合，$dist(d_i, d_j)$ 表示数据点 d_i 和 d_j 之间的欧式距离。

定义 4：邻域 ∂，对于某一数据集样本 $d(d \in D)$，对象 d 的邻域记为：

$$\partial_d = \{d \mid 0 < dist(d, d_j) \leqslant r\} \quad (2)$$

即以 d 为中心，r 为半径的圆形区域，

其中，$r = \dfrac{average(Dist)}{n^{adjustr}}$ (3)

$average(Dist)$ 表示全部样本点间距离的均值，n 表示样本点的数目，$adjustr$ 则为邻域半径调节参数。

定义 5：类簇 T 表示为 $\{t_1, t_2, \ldots, t_i\}$ 所构成的集合，其中 t_i 表示聚类划分所产生的第 i 个类簇，令 $C = \{c_1, c_2, \ldots, c_n\}$ 为网络中流量的应用类型，流量分类 中存在映射 $T \Leftrightarrow C$，也即：$\{t_1, t_2, \ldots, t_i\} \Leftrightarrow \{c_1, c_2, \ldots, c_n\}$

算法描述如下：

1）若网络应用程序的类型已被标记的数目记做 s，如果 $s > k$，则表示设定的聚类数目 k 小于网络流量分类数据集中实际存在的类别数量，该算法产生的结果会有极大的误差，不能作为分类识别结果，算法结束，否则，算法继续执行；

2）命名聚类中心点集合 B 并初始化，设 $B = \{\ \}$；

3）对于每个已标识的网络应用类别，根据式（1）计算所有的该类型网络应用数据的密度 $dens(d_i)$，并根据计算，选择其中每个类别密度最大的样本点加入到初始聚类中心集合 $B = \{B_1, B_2, \ldots, Bs\}$；

4）从原数据集 D 中删去这些数据点，根据式（2）和式（3）计算其邻域中的数据对象，并从原数据集合中删除；

5）分别计算余下的每个未标记数据点的密度 $dens(d_i)$，从中选择密度最大的 1 个，并加入到集合 B 中，计算这一中心点的邻域中的数据对象并从原数据集合中删除；

$$B = B \cup \{d_j\}$$
$$D = D\{d_j\} \quad (4)$$

6）迭代执行算法第 4 步，直到集合 B 中初始簇中心点的数目达到 k，即 $|B| = k$；

7）输出中心点集合；

改进后的 k-means 方法在获取 k 个初始的聚类中心后，对数据集进行聚类挖掘分析，然后通过对已标记的数据对象进行最大似然估计，分析聚类识别结果并与相应具体的网络协议和网络应用类型进行匹配，从而获取网络流量分类识别结果。该算法通过改进初始聚类中心的选取，优化了聚类划分结果，同时也加快了算法的收敛速度。

因此，在当初始聚类中心集合 M 被获取后，按照 k-means 流程，运算数据集中每个对象和所选取的 k 个初始聚类中心的距离，得到其最小值并被分配到所代表的簇中，直到算法的准则函数收敛，过程结束。利用改进的 k-means 聚类的半监督学习算法，通过减小初始聚类中心的随机性，来尽可能找到真正中心的簇并将其作为初始聚类中心，以尽可能获取 k-means 中的全局最优解，优化分类识别结果。

聚类结果的类簇和网络应用之间的对应匹配：

改进的 k-means 聚类算法能产生 k 个簇聚类的结果，其目的是为了匹配网络流量识别特定类型。因此，首先需要 k 个对应的网络应用程序的类型，据此来标记聚类划分中的数据所对的特定网络应用类型，我们使用最大似然估计法来创建 1 个基

于概率的对应匹配函数。其中设 $C = \{c_1, c_2, \ldots, c_n\}$ 是一组网络流量所对应的应用类型，采用最大似然估计建立网络流量和对应的应用类型的函数关系，其概率公式标记为：

$$P(C = c_i \mid t_i) = \frac{l_{ji}}{l_j} \tag{5}$$

式中：t_i 表示聚类划分所产生的第 i 个类簇，l_{ji} 表示第 i 个类簇中已标记为应用类型对象 c_j 的数目，l_j 表示第 i 个类簇中的数据总数，概率 P 表示将类簇 t_j 对应为 c_j 应用类别的概率。

4 实验与分析

4.1 实验环境及数据集

实验中的硬件配置主要采用 Intel (R) Xeon (R) 4.00 GHz ×4 的中央处理器以及 8 GB 物理内存，操作系统平台采用 Linux 2.6.18 内核，实验工具利用 L7-filter，MATLAB 和 Weka。

实验利用作者在 UCI 期间采集的数据集，该数据采集于 ICS 网络实验室的节点路由器，该节点有大约 100 名老师和学生，为全报文采集，采用 L7-filter 对报文进行标识，主要分为以下 7 种类型，其详细信息如表 1 所示。

表 1 ICS SET 数据集

Table 1 Data sets of ICS SET

Type of flow	Application names	Number of flow	Percent
WWW	HTTP	38 160	56.43%
P2P	Bittorrent，eDonkey，Gnutella	20 476	30.28%
Mail	Smtp，Pop3，Imap	865	1.28%
Service	Dns，Ntp	61	0.09%
Multimedia	Real，RTSP	4 362	6.45%
BULK	FTP	2 448	3.62%
DB	Oracle，DB2，Sqlnet	1 251	1.85%

在数据测度属性选择上，需要提取聚类学习算法分类识别所需的数据属性特征行为集，由于空间维度上的测度属性的计算时空复杂度较高，且在多用户多任务环境下，同一主机在同一时间片内往往存在多个网络应用活动，协议行为也难以确定，主要选择一些时间维度上的测度属性集，本文选择的具体特征属性信息如表 2 所示。

表 2 数据集测度属性

Table 2 Measure attribute of data sets

Abbreviations of the names	Feature description
total packets	Total number of packets in biodirection
total bytes	Total number of bytes in biodirection
duration	Duration of the flow
Mean pktiat	Mean packet inter-arrived time
Mean pktl	Mean packet length in biodirection

4.2 实验结果分析

采用均匀抽样的方式，抽取数据子集中每种网络应用各 800 条（因为 mail 类型为 865 条），设定每种数据流中已标记的数据数目分别为 50 100 条，组成各类样本数量均等的训练数据集，其中，由于 service 类型的网络流数目较少，本实验不包括这种类型的流量样本。

在设定不同参数 k 的情况下，分别运行 k-means 算法和改进初始中心点选择的 k-means 算法，选择标记各流量类型为 50 条的测试数据集，参数 k 的值从 30 开始以 30 为步长间隔逐渐递增到 150 的情况下，对比方法的总体准确性，其实验结果如图 1 所示。从图中可以看出，对聚类初始中心点的选择做出优化后，改进后的 k-means 聚类算法的网络流量分类识别结果较直接使用 k-means 算法的准确率更高，并且随着参数 k 的增大，2 种算法的识别准确率都有所增加，在具体问题中，总是希望假设的类簇数目 k 等于或者大于真实的类簇数目，这样，

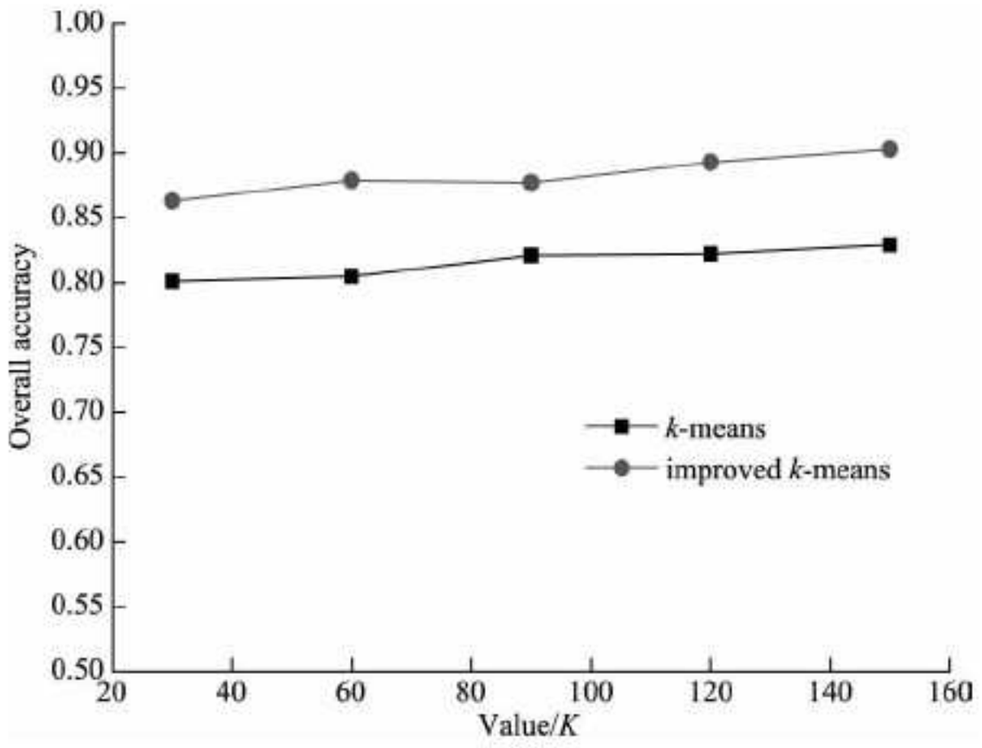

图 1 总体准确率比较

Fig. 1 The comparison of overall accuracy

算法才能取得更为准确的聚类划分结果，因此，k 值的选择对 k-means 算法的聚类结果是一个很重要的影响因素。

随后，测试数据集标记数据的大小对流量分类识别结果的影响，利用上述的各类型流量中标记流数量分别为 50，100 的测试数据集，参数 k 的值同样从 30 开始以 30 为步长间隔逐渐变化到 150，其识别准确率如图 2 所示，从图中可以看出，在参数 k 的值相同的情况下，已标记数据流的数目越大，其分类识别结果的总体准确率越高，当已标记数据流在整体数据集合中所占的比例相同时，参数 k 取值越大，算法结果的总体准确率也逐渐增大。

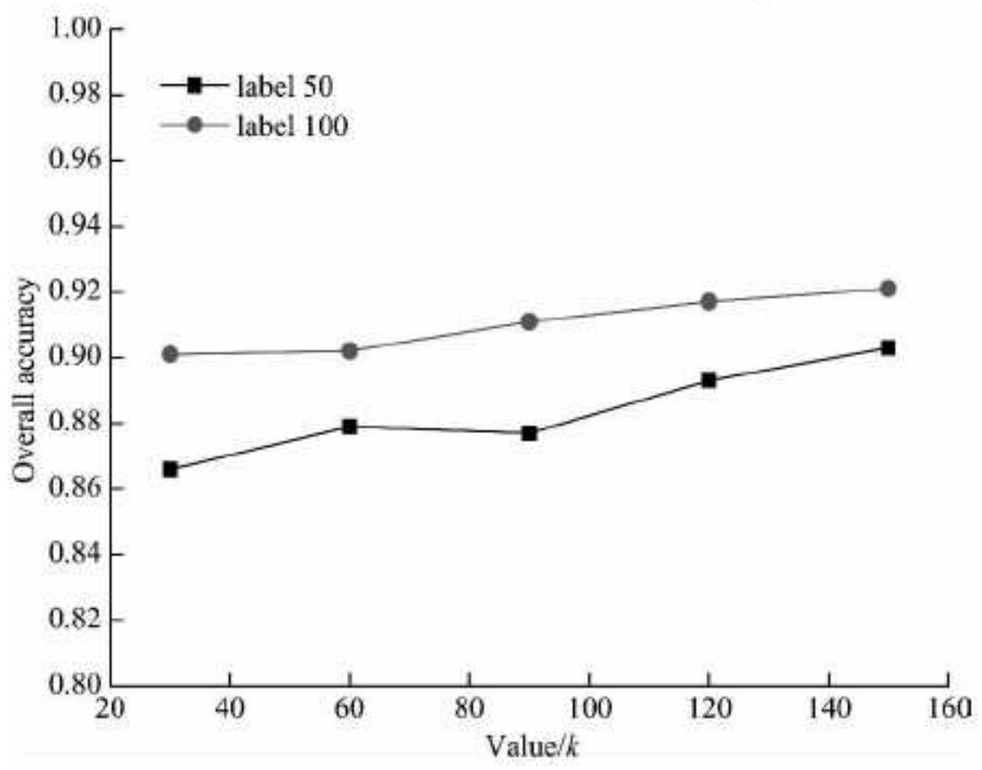

图 2　不同标记数据总体准确率比较

Fig. 2　The comparison of overall accuracy rate of different labeled data

在改进的 k-means 算法中，所选择的中心点除了考虑在距离维度上的分布，也衡量中心点周围的密度分布，这样以尽可能满足在同一流形上的数据样本点具有较高的相似性的特征，相比于高斯函数，还可避免其对尺度敏感的缺点，从而使得算法所选择的初始簇中心有可能更为接近实际的簇中心，而最大似然估计计算也能获得更优的分类识别结果，从总体上来说，改进后的基于半监督学习的网络流量分类识别方法在准确性上表现得更为稳定，算法平均准确率在 90% 以上，可以有效地解决相关网络流量分类识别问题。此外，k 值的增大有利于获取更优的聚类划分，但也使得相应聚类算法的运行时间变长，时空开销加大。

后续的研究工作中，在基于密度因子的初始聚类中心选择上，由于密度参数和邻域半径调节系数 $adjustr$ 的取值因实验数据集的不同而有所不同，并会影响到初始中心点的选择和最终的聚类结果，一般根据经验，密度参数 $\delta > 1$，邻域半径调节系数 $0 < adjustr < 1$。此外，参数 k 的值选择比较敏感，其预先设定的初始值对聚类的数目及识别的结果有着直接的影响，当需要对未知网络流量或者更新的网络应用产生的流量进行分类识别分析时，准确的 k 值设定还存在一定的困难，有待进一步研究。

5　结　论

结合现有的有监督机器学习算法和无监督机器学习算法，提出一种基于半监督学习的流量分类识别算法，改进了基于 k-means 的网络流量分类算法，通过密度敏感的相似性函数来满足聚类过程中数据的全局一致性要求，并以此获取算法聚类初始簇中心，以获得更好的聚类划分结果并利用已知标记信息完成相应类簇的对应匹配过程，最终的实验结果表明该半监督分类识别算法可以取得较好的分类识别效果，能够有效满足流量分类识别的应用需求，在后续工作中，算法仍需要对参数的初始值设定和实时性改进等方面进行进一步的研究。

参考文献

[1] NGUYEN T T, ARMITAGE. A survey of techniques for internet traffic classification using machine learning [J]. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56-76.

[2] FRALEIGH C, MOON S, LYLES B, et al. Packet-level traffic measurements from the spirint IP backbone [J]. IEEE Network, 2003, 17(6): 6-16.

[3] PAPAGIANNAKI K, TAFT N, ZHANG Z, et al. Long-term forecasting of internet backbone traffic: observations and initial models [C]. Proceedings of INFOCOM, London, UK, 2003: 753-764.

[4] TRIKHA P, S Vijendra. Fast density based clustering algorithm [J]. International Journal of Machine Learning and Computing, 2013, 3(1): 10-12.

[5] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifier [J]. Machine Learning, 1997, 29(2): 131-163.

[6] QUINLAN J R. C4.5: programs for machine learning [M]. California: Morgan Kaufmann, 1993.

[7] 顾成杰，张顺颐，基于改进 SVM 的网络流量分类方

法研究[J]. 仪器仪表学报, 2011, 32(7): 1507-1513.

[8] LIU B, HSU W, MA Y. Integrating classification and association rule mining[C]. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, USA, AAAI press, 1998.

[9] NG T R, HAN J W. Efficient and effective clustering method for spatial data mining[C]. Proceedings of the 20th International Conference on Very Data Bases, Santiago, Chile, 1994: 144-155.

[10] WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained K-means clustering with background knowledge[C]. Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann Publishers, Williamstown, 2001: 577-584.

[11] 杨会锋, 曹洁, 帅立国. 基于改进 K-均值聚类算法的背景建模方法[J]. 电子测量与仪器学报, 2010, 24(12): 1114-1118.

[12] ESTER M, KRIEGEL H, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, USA, 1996: 226-231.

[13] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[C]. Proceedings of the 30th IEEE Conference on Local Computer Networks. (LCN 30), Sydney, Australia, 2005: 220-227.

[14] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.

[15] YU S X, SHI J. Segmentation given partial grouping constraints[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2004, 26(2): 173-183.

[16] HERTZ T, SHENTAL N, Bar-Hillel A, et al. Enhancing image and video retrieval: learning via equivalence constraint[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Madison, 2003: 668-674.

[17] BERKHIN P, BECHER J. LEARNING Simple relations: theory and applications[C]. Proceedings of the 2nd SIAM ICDM, Arlington, 2002: 333-349.

[18] ERMAN J, MAHATI A, ARLITT M, et al. Semi-supervised network traffic classification[C]. Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, USA, 2007: 369-371.

[19] PARK K, WILLINGER W. Self-similarity network traffic and performance evaluation[M]. New York: John Wiley& Sons, 2000.

[20] PAPAGIANNAKI K, TATF N, ZHANG Z, et al. Long-term forecasting of internet backbone traffic: observations and initial models[C]. Proceedings of INFOCOM, London, UK, 2003: 753-764.

作者简介

周文刚, 1981 年出生, 电子科技大学博士, 2009 - 2011 年美国加州大学欧文分校, 主要研究方向为网络管理、机器学习等。

E-mail: wengangz@uci.edu

Zhou Wengang was born in 1981, Ph. D. candidate in University of Electronic Science and Technology of China, visited the University of California, Irvine CA, USA, from November 2009 to December 2011. His research interests are in the fields of machine learning and network measurement.

陈雷霆, 1966 年出生, 电子科技大学教授, 主要研究方向为网络计算、图像处理、虚拟现实技术等。

E-mail: richardchen@uestc.edu.cn

Chen Leiting was born in 1966, professor, received Ph. D. in computer science from University of Electronic Science and Technology of China. His research interests are in the areas of computer graphics and virtual reality technology.

董仕, 1980 年出生, 东南大学博士, 主要研究方向为网络安全、网络管理。

E-mail: shdong@njnet.edu.cn

Dong Shi was born in 1980, and now he is Ph. D. candidate of Southeast University, China. His major research fields include network security and network management.