NBOS_S 流量的标识分类

卓文辉[1,2]，丁伟[1, 2]

（1.东南大学 计算机科学与工程学院，江苏 南京，211189；2. 东南大学计算机网络与信息集成教育部重点实验室，江苏 南京，211189）

摘　要：提出了针对 NBOS_S 流量进行标识分类的方法—基于流属性测度进行 DBSCAN 聚类。采用 DPI、端口、行为特征多融合的方法识别了 NBOS_S 监控环境下 96%的流量，结合本地抓包流量，获取到标准应用流量。将 SU 概念引入到流测度选择中，避免冗余无关测度对聚类造成影响。将核函数引入到 DBSCAN 聚类算法中，消除了聚类对象分布不均对聚类结果造成的不良影响。通过对聚类流样本基于不同的测度组合进行两次 DBSCAN 聚类，优化了聚类的结果，得到了最终的分类目标方案。

关键词：流量分类；DBSCAN；SU；核函数

Traffic Classification for NBOS_S

ZHUO Wenhui[1, 2], DING Wei[1, 2]

(1. School of Computer Science and Engineering, Southeast University, Nanjing, 211189; 2. Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing, 211189)

Abstract: Considered the request of NBOS_S, this paper proposes a method for traffic classification based on flow features using DBSCAN technique. Mixed the method of port-based、payload-based and host behavior classification, we identify approximately 96% traffic monitored by NBOS_S, and then we get standard application traffic by combining it with the local catch traffic. In order to achieve high-density clustering results, Symmetrical Uncertainty notion is applied to obtain the best combination of flow features, furthermore, kernel function is introduced into the DBSCAN clustering algorithm to eliminate the negative effect caused by samples' uneven distribution. For optimizing the clustering results, we use DBSCAN technique twice based on different feature combinations and get the final target classification scheme.

Key words: Traffic identification; DBSCAN; SU; Kernel function

NBOS 是 CERNET 华东北地区网络中心在国家支撑计划的支持下开发的一个用于监控和管理 JSERNET 网络服务质量和网络安全状态的新型网络管理系统，它的一个实用版本 NBOS_S 是在 211 工程三期ＣＥＲＮＥＴ建设项目专题“高性能网络管理与安全保障”中支持面向实时性开发的一个版本。

该系统中的流量识别模块是对其监控环境下的流量进行应用识别并予以展示。由于系统的数据源为 NetFlow 流记录，无应用层负载，所以采用机器学习方法基于流属性对流量进行识别。种类繁多的网络应用以及应用流量间的相似性，使得依赖于机器学习的分类方法在具体应用识别能力上有限。当前流量识别模块选取了 www、P2P、邮件等 9 大类作为应用分类目标，提取 NBOS_S 数据库中的识别结果进行分析，发现大部分情况下，www、p2p 两种流量的比例高达 98%，交互、语音、其他三类所占比例为 0。该分类结果提供给我们有用的信息量不大，为给出一个合理可行且尽可能详细的分类结果展示方案，首先我们要对 NBOS_S 监控环境下的流量分布情况有足够的认识，其次我们要了解流属性特征对不同应用的区分能力。

DPI 是目前公认的识别网络流量最准的方法，但它的识别范围受到了特征库的约束，本文在该方法基础上，结合端口、连接行为特征两种判别方法，对 NBOS_S 监控环境某一时间段下的流量进行了识别，得到了该网络环境下某时段 96%流量的分布情况，同时也获得了部分应用的标准流量，称作 DT。

基金项目：国家 973 基础研究项目（2008BAH37B04）

作者简介：卓文辉，（1989-），女，硕士研究生，E-mail: whzhuo@njnet.edu.cn；丁伟，（1963-），女，教授，E-mail: wding@njnet.edu.cn.

为了弥补 DT 集中应用种类的不足，本文选取一些热点应用进行本地抓包，与 DT 构成总的标准流量，称作 NT。针对 NT，使用基于信息熵理论的 SU 方法，对 NetFlow 流记录可计算属性的重要度进行了排名，选取前 8 的流属性作为区分不同应用流量的测度。

为了获得最终的分类目标，我们采用改进的 DBSCAN 方法基于挑选出的 8 个流测度进行聚类，根据应用流量的聚合结果得出分类方案。

本文内容的组织结构如下：第 1 节对本文中用到的算法、概念进行了介绍，包括对称不确定性 SU 的定义以及 DBSCAN 算法。第 2 节通过引入核函数对 DBSCAN 算法进行了改进。第 3 节详细的介绍了分类目标获取的整个实验过程，包括标准流量的获取，流测度的选择以及聚类过程的实现。第 4 节对实验的结果进行了分析，确定了分类目标。第 5 节对本文进行了总结并提出展望。

1 相关技术

1.1 SU(Symmetrical Uncertainty)[1]介绍

信息学理论中的条件熵和互信息的计算，对于变量的初始分布没有任何要求，故可适应与任何分布下的变量相关关系分析。

熵是随机变量的不确定性，即包含信息量多少的度量，令 X 为一随机变量，其熵 $H(X)$定义为：

$$H(X)=-\sum_{i=1}^{|X|}p(x_i)\log_2(p(x_i)) \qquad (1)$$

其中 $P(x_i)=P(X=x_i)$。$H(X)$越大，即 X 的不确定性越大，所携带的自信息量越大。在另一随机变量 Y 的值确定的情况下，变量 X 的条件熵 $H(X|Y)$定义为：

$$H(X|Y)=-\sum_{j=1}^{|Y|}p(y_i)\sum_{i}^{|X|}p(x_i|y_j)\log_2(p(x_i|y_j)) \qquad (2)$$

式中 $P(x_i|y_j)$代表在随机变量 Y 的取值为 y_j 的情况下，随机变量 X 取值为 x_i 的概率。由于 $H(X)$表示知道 Y 取值之前的关于 X 的不确定性，而 $H(X|Y)$表示观察到随机变量 Y 的取值后，仍保留的关于变量 X 的不确定性，则差值 $H(X)-H(X|Y)$必然表示了由随机变量 Y 所提供的关于 X 的信息量，即在信息论中被称为 X 和 Y 之间的互信息量，表示为 $I(X;Y)$。

若变量 X 与变量 Y 不相关，则 $I(X;Y)=0$；否则 $I(X;Y)>0$，且 $I(X;Y)$值越大，表明 X 与 Y 的相关性越强。因此，可以用互信息量 $I(X;Y)$来定量衡量两测度之间的相关关系。但是，由于 $I(X;Y)$的结果受变量取值和单位的影响，故进一步对其进行均一化，得到以下对称不确定性 SU(Symmetrical Uncertainty)的定义：

$$SU(X;Y)=SU(Y;X)=2\times\left[\frac{I(X;Y)}{H(X)+H(Y)}\right] \qquad (3)$$

SU取值范围在 [0, 1]之间，且为互信息量$I(X;Y)$的单调增函数，数值越大表示两变量间相关程度越强，反之则越弱；取 0 表示两变量相互独立，取 1 表示两变量间存在严格的函数关系。由于 SU 具有较高的准确性和通用性，因此本文将其引入到网络流测度相关性分析领域中，作为定量衡量两流测度、流测度与流类别之间相关关系的标准。

1.2 DBSCAN 算法介绍

由于网络流量数据具有维数高、不稳定、非线性、复杂性等特点，基于欧式距离的聚类方法只能形成球形的分割面，而基于密度的聚类算法可以形成任意边界形状的簇，DNSCAN 算法是基于密度算法的代表。

1.2.1 DBSCAN 基本思想

DBSCAN 算法的聚类过程也叫密度扩展，它需要输入 2 个参数：邻域半径 Eps 和密度阈值 MinPts。一个数据点的 Eps 邻域半径范围内的数据点数目超过了 MinPts 即为高密度点，也叫核心对象，而 DBSCAN 的主要思想就是将足够高密度的数据点聚类成一个簇，整个过程由迭代的邻域搜索来完成，如果 q 是核心对象，则将所有从 q 密度可达的对象标记为当前类，并从它们开始进一步的邻域搜索，如果 q 是一个边界对象，那么 q 被标记为噪声，然后从下一个对象继续进行处理。一次反复进行下去，直到没有新的对象被加入到当前的聚类中，然后再选择一个新的种子对象开始扩展，得到下一个类，直到所有的对象都被标记为某个聚类或者噪声为止。由此可以看出，DBSCAN 算法对噪声数据是不敏感的，通过邻域搜索结合密度可达思想，DBSCAN 算法能发现空间中任意形状和大小的簇，而且聚类结果不受输入顺序影响，聚类速度快，时

间复杂度为$O(n\log n)$。

1.2.2 DBSCAN 算法不足

DBSCAN 算法主要存在以下不足，1）对输入参数敏感，确定参数 Eps,MinPts 困难，若选取不当，将造成聚类质量下降。2）不能有效地对密度差异较大的数据集进行聚类。在 DBSCAN 算法中，变量 Eps,MinPts 全局唯一，对于数据分布的密集程度有较大差异的簇，在进行密度扩展时容易打破簇的固有结构，造成错误的合并或分割现象。本文将在第 2 节中详细介绍针对 DBSCAN 算法做出的改进。

2 DBSCAN 聚类算法的改进

针对 DBSCAN 算法存在的问题，虽然已经有了些改进方法[2,3,4]，但它们多数从分割数据集，设置多组邻域、密度阈值参数方面考虑，均需人工观察判断数据集中出现的密度层次，增加了用户的额外负担，且与数据分布相关，不具有适用性，并没有很好解决存在问题。本文换一个角度出发，仍采取全局统一的邻域、密度参数，但是通过引入核函数，对聚类的流量样本进行非线性映射，使得样本集分布密度尽可能地均匀，有用的特征被较好地分辨、提取并放大，不同类别间距增大，从而达到更好的聚类效果。

2.1 高斯核函数

假设 $X=\{x_k\in R^N, k=1,2,\cdots l\}$ 是一个非空的输入空间的样本集，被某种非线性映射 Θ 映射到某一特征空间 H 得到 $\Theta(x_1),\Theta(x_2),\cdots,\Theta(x_l)$。如果函数 $K: X\times X\to R$ 满足：

$$K(x_i,x_j)=(\Theta(x_i)\cdot\Theta(x_j)),\forall x_i,x_j\in X \quad (4)$$

则称 K 为核函数。任何一个函数只要满足 Mercer 条件，就可以作为核函数。在无监督学习模型中，核函数一般是凭经验选取，在一般情况下，首先考虑的是高斯核函数（见公式 2），因为高斯核函数对应的特征空间是无穷维的，有限的样本在该特征空间肯定是线性可分的。

$$K(x_i,x_j)=\exp(-\beta\|x_i-x_j\|^2),\beta>0 \quad (5)$$

2.2 核空间距离

在输入空间样本 X 被 Θ 映射到特征空间 H 后，特征空间的 Euclidean 距离可以表示为：

$$\begin{aligned}D_{H(x_i,x_j)}&=\sqrt{\|x_i-x_j\|^2}\\&=\sqrt{\Theta(x_i)\cdot\Theta(x_j)-2\Theta(x_i)\Theta(x_j)+\Theta(x_j)\Theta(x)}\end{aligned} \quad (6)$$

将公式（4）代入公式（6），即可得到公式（7）

$$D_{H(x_i,x_j)}=\sqrt{K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)} \quad (7)$$

再由公式（5），便可以方便的计算出特征空间中点 x_i 和点 x_j 的距离。因此，我们便将公式（7）作为 DBSCAN 聚类算法中，样本距离的度量函数。

2.3 改进算法描述

改进后的 DBSCAN 算法步骤如下：

1）将待聚类的对象构造成向量矩阵 X。假设样本中有 n 个对象，每个对象有 m 个测度，则 X 可表示成：

$$X=[x_1,x_2,\cdots,x_n]^T, x_i=[x_{i1},x_{i2},\cdots,x_{im}]$$

2）利用公式（5）对初始数据矩阵 X 中的 n 个对象两两进行计算，得到 $n*n$ 维的对称核矩阵 K，K 中的每一个元素

$$k_{ij}=K(x_i,x_j)$$

3）任选一未处理对象 p，利用公式（7）和核矩阵 K 计算出对象 p 与 X 中其他对象的距离，统计距离小于参数 Eps 的数目 N。

4）若 N 小于参数 MinPts，则将点 p 标记为噪声对象，寻求下一个处理点；若 N 大于参数 MinPts，则该点 P 与其 Eps 范围内所有点形成一个簇 C，将点 p 标记为已处理，然后递归，以相同的方法处理簇内所有未被处理的点，从而对簇进行扩展。

5）重复步骤 3~4，直到所有对象都归入了某个簇或标记为噪声。

3NBOS_S 标识改进方法

为给出一个合理可行且尽可能详细的分类结果

展示方案，我们的工作分为三部分，1）得到网络中各应用、协议的标准流量；2）利用 SU 算法挑选出合适的流属性测度。3）利用改进的 DBSCAN 方法，基于挑选出的流测度对标准应用流量进行聚类，根据聚类结果，得到分类目标。

3.1 应用标准流量的获取

3.1.1 流量数据源

应用流量的来源有三个：1）IPTAS：江苏省网边界到 CERNET 国家主干路由之间时长 1 小时带报文负载的 IPTrace 数据。2）本地端系统对热点应用的抓包数据（.pcap 文件）。3）IBR（互联网背景辐射）流量：IBR 是客观存在于互联网中的一种无功流量[]。蠕虫或黑客的扫描、拒绝服务攻击的反向散射、网络设备的错误配置等均可能会导致 IBR 产生。该流量数据由本实验室 IBR 相关研究工作人员提供。表 1 给出了这三类流量的概要信息：

表 1 概要信息

DATA	流数	报文数	字节数
IPTrace	5.67E+7	4.86E+8	51G
本地抓包	9.13E+5	2.4E+7	17G
IBR 流量	2.48E+7	2.87E+7	2G

3.1.2DPI 方法获取应用流量

nDPI 是在 OpenDPI 上扩充改进的 DPI 库，能识别多达 170 种应用，本文基于该开源软件，在融合端口识别、连接行为特征[5]的基础上，对 IPtrace 中的流量进行了识别，得到了该数据源中 96%流量的应用分布情况，如表 2 所示的标准流量，我们称之 DT。

表 2 DT 流量分布情况

应用类别	所含协议	字节比重
Mail	pop,smtp,imap,etc	0.196
域名解析	dns,mdns,whois-das,etc	11.25
DataBase	postgresql,mssql,mysql,etc	0.498
P2P	eDonkey,BT,pplive,etc	12.67
Interactive	telnet,rdp,ssh,vnc,smb,etc	0.46
VoIP	skype,teamspeak,viber,etc	4.52
WWW	http,http_proxy,etc	63.41
BULK	ftp	0.054
Service	ntp,netbios,nfs,upnp,etc	0.65
其他	Remotescan,games,vpn	6.286

3.1.3 本地抓包获取应用流量

由于 DT 数据集为某一时段的流量分布情况，为了弥补 DT 集中应用种类的不足，本文选取一些热点应用，在本地利用 Wireshark 软件进行抓包，热点应用选取参考了 CNNIC 发布的《第 32 次中国互联网网络发展状况统计报告》[6]、Alex 中国[7]、中关村在线软件下载排行榜[8]、百度搜索风云榜[9]发布的统计数据。主要集中在视频电影网站、P2P、即时通信、游戏 4 类。详见表 3：

表 3 本地抓包应用

应用类别	应用名称
视频网站	优酷、CNTV、酷 6、搜狐、爱奇艺、乐视、腾讯视频、土豆、新浪、PPS 网络电视、风行
P2P	uTorrent、Bitcomet、eDonkey、PPTV、PPS、迅雷、暴风影音
即时通信	QQ、阿里旺旺、YY 语音
游戏	wow、crossfire、dnf、kartrider、梦幻西游、腾讯游戏系列(飞车、三国、音速、炫舞、华夏)

综上可知，DT 流量、本地应用抓包流量以及 IBR 流量，共同构成了本文的应用标准流量，称作 NT。

3.2 基于 SU 算法的流测度选择

NBOS_S 的数据源为 NetFlow 流记录，基于 NetFlow 的固有字段，可计算的流属性有源/宿端口、报文数、持续时间、平均报文到达间隔等共计 20 个。好的流测度组合应包含足够的类别信息且空间维数尽可能的低，因此测度选择的目的主要有两点，一是去除对类别属性无关的特征，二是去除冗余的特征。

本文采用 1.1 节中介绍的 SU 算法来进行流测度选择，步骤如下：

1）对 NT 流量按照 5 元组进行组流，计算出每

条流 $m(m=20)$个测度的值，随机选取各种应用的流共计 $n(n=4041)$条，构成测度集。测度集可以看作是一个 $n\times m$ 维的矩阵 X，矩阵中 $n\times 1$ 维向量 X_j 表示第 j 个测度 n 条流的所有取值，元素 x_{ij} 表示第 i 条流第 j 个测度的取值。n 条流相对应的应用类型可描述为一个 $n\times 1$ 的矩阵 C，矩阵中元素 c_i 表示第 i 条流所属的应用类别。如此，扩展矩阵 M=[X;C]便构成了我们的属性选择的样本集。

2）计算向量 X_j 与向量 C 之间 SU 系数，若 SU 系数小于某阈值 $\delta_1(\delta_1=0.45)$，则认为测度 j 不能提供对分类有用的信息，属于无效测度，不选，即 $X=X-\{X_j\}$；反之不作处理。

3）对 X 中剩余的测度，计算向量两两测度向量间的 SU 系数，若 SU 系数大于某阈值 $\delta_2(\delta_2=0.85)$，则认为测度 j 与测度 l 相互冗余，删除两者中与类别向量 C 的 SU 系数较小的测度；反之不作处理。

表格 4 给出了该方法选择出来的测度结果：

表 4 测度描述

测度	描述	重要度
低位端口	NetFlow 字段	1
传输层协议	NetFlow 字段	0.621
双向字节数	前/后向字节数之和	0.582
平均报文长度	双向字节数/双向报文数	0.563
BPS	字节数/持续时间	0.538
平均到达间隔	持续时间/报文数	0.538
双向报文数	前/后向报文数之和	0.521
双向报文数比	前/后向报文数之比	0.519

注：重要度为 $SU(X_i,C)$与 $MAX(SU(X_i,C))$的比值

3.3 改进 DBSCAN 基于流测度聚类

我们利用第二节中介绍的方法步骤进行聚类，算法描述中的 n 个对象对应着选取的 4041 条流，m 个属性测度对应着 3.2 节选出的 8 个属性测度，在实现 DBSCAN 聚类过程中，有两个需要手动输入的参数 Eps 和 MinPts，为了确定适合本样本的参数组合，我们根据评价聚类质量的两个原则—紧密度及分离度，即同簇距离尽可能近，簇与簇之间的距离尽可能地远，给出度量聚类分析质量的指标，轮廓系数 sc(sihouette coefficient):

$$sc=\frac{b-a}{\max(a,b)} \tag{8}$$

其中 a 表示某一样本点与其同簇中其他所有点之间的平均距离。b 表示某一样本点与其相邻簇中所有点之间的平均距离。

由定义可知，该系数结合了凝聚度和分离度，因此，可以以此来判断聚类的优良性。轮廓系数的取值在区间[-1,1]间，值越大表示聚类效果越好。依据这个原理，我们可以尝试选取多组聚类参数（Eps，MinPts）的值，反复计算在每组参数条件下的轮廓系数，得到合适的聚类参数值。

为了对聚类实验结果进行更好的分析，我们再给出两个聚类分析度量参数，完整性系数 c 和一致性因子 h：

完整性系数 c（completeness）

$$c=1-\frac{H(K|C)}{H(K)} \tag{9}$$

其中，H(K)表示簇分布的信息熵（公式 10），$H(K|C)$表示在给定协议类型为 C 的流样本中，簇分布的信息熵（公式 11）。

$$H(K)=-\sum_{k=1}^{|K|}\frac{n_k}{n}\cdot\log\left(\frac{n_k}{n}\right) \tag{10}$$

$$H(K|C)=-\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{n_{k,c}}{n_c}\cdot\log\left(\frac{n_{k,c}}{n_c}\right) \tag{11}$$

由定义易知，该系数表示的是同一协议的流样本被聚类到同一簇的程度，取值区间为[0,1]，数值越大，完整性越好，当协议类型为 C 的流均被聚类到一个簇中，则 H(K|C)的值为 0，完整系数为 1。我们可以将该测度来评判分类目标的可行性。

一致性因子 h(Homogeneity)

$$c=1-\frac{H(C|K)}{H(C)} \tag{12}$$

其中 H(C)表示应用分布的信息熵，H(C|K)表示给定簇下，应用类型分布的信息熵。该系数表示同一簇中，应用类型的一致程度，取值区间为[0,1], 数值越大，一致性越高，当簇中全为一种应用类型对象时，该系数值为 1.

4 实验结果与分析

在初始聚类过程中，并没有达到很好的效果，

聚类算法将 84%的流聚成了 10 类，16%是噪声数据，对这 10 类进行分析，发现存在以下两个现象：1）应用如 Telnet 等，其聚类后的完整性系数为 1，其对应簇的一致性因子却小于 0.4，这表明虽然某个应用的流量全部聚成了一类，但是该类中还存在其他的应用流量。2）应用如 BitTorrent 等，被归类到了多达 6 个类别。对这两种情况下应用的流测度属性进行对比分析，发现，第一种情况中的大部分应用具有明显的端口特征，但其他属性测度特征不够明显，这样，聚类算法在用公式（7）计算空间距离过程时，其他 7 个流属性测度，会减轻端口属性的差异，使得不能够很好的区分；第二种情况产生的原因是，应用的流量特征不唯一，流与流的空间距离较大，在流总数大的情况下，为考虑聚类的全局效果，DBSCAN 算法的参数 Eps 必定不大，导致找不到该应用流量的聚类点，因而导致该流量与其附近其他协议流量划分成一类的结果。

为了消除和改善这两种情况，本文采用改进的 DBSCAN 算法对不同的流属性测度进行两次聚类，第一次聚类过程只针对端口属性进行，实验过程中聚类参数 Eps=0.03，MinPts =30，轮廓系数 sc=0.854，得到 26 个簇，对完整性系数 c 为 1，一致性因子系数 h 为 1 的点，直接作为可分目标，满足条件的簇共有 21 个，分别是 sip、vnc、telnet、ssh、mysql、mssql、flashget、gre、imap、pop、icmp、smb、netbios、ftp、snmp、rsync 等。

将剩下的流量样本进行二次聚类，聚类过程针对 8 个属性测度，实验过程 Eps=0.4，MinPts=10，轮廓系数 sc=0.703，得到了 10 个簇，簇标号为 0~9。对结果进行统计发现，两次聚类方法能够有效地改善应用流量的聚类结果，图 5 中给出了部分统计结果：

表 5 聚类结果统计

协议	类别（比例）
http	0(0.69) 3(0.24) 5(0.026) 6(0.035)
flash	3(1.00)
viber	7(1.00)
eDonkey	1(1.00)
pplive	1(1.00)
BitTorrent	0(0.46) 1(0.48) 4(0.01) 7(0.05)
IBR	1(0.05) 2(0.91) 3(0.01) 4(0.01) 5(0.01)

注：0(0.69)表示 http 流量中 69%的流聚集到了簇 0 中。

结合两次的聚类结果，我们可以得到如表 6 所示的分类目标：

表 6 分类目标方案

一级分类	二级分类
www	http flash http_proxy
Mail	--
域名解析	--
DataBase	--
P2P	Flashget BT/eDonkey/PPlive Others
Interactive	--
VoIP	skype/viber others
BULK	ftp
Service	--
IBR	--
其他	--

5 结束语

本文利用 DBSCAN 聚类算法对 NBOS_S 监控环境下不同应用的流属性特征进行聚类，根据聚类结果得到了可行的流量标识分类方案。为了能得到高密度的聚类结果，将核函数引入到 DBSCAN 聚类方法中，减轻了流量样本不均匀性对聚类产生的负面影响，同时将信息熵中的 SU 概念引入到流属性测度选择中，选出了 8 个最佳的测度。为了确定 DBSCAN 算法中参数 Eps,Minpkts 的合适取值，使用了轮廓系数的概念。为了能够定量地分析聚类结果的质量，我们使用了完整系数和一致性因子的概念，使用基于不同流测度组合两次聚类的方法，优化了聚类结果。本文的下一步工作为，依据制定的分类目标方案，基于机器学习方法构造分类器，对 NBOS_S 监控环境下的流量进行分类。

参考文献：

[1] W H Press, S A Teukolsky, W T Vetterling, B P Flannery. Numerical Recipes in C [M]. London: Cambridge University Press, 1988.

[2] Jiang H, Li J, Yi S, et al. A new hybrid method based on partitioning-based DBSCAN and ant clustering[J]. Expert Systems with Applications, 2011, 38(8): 9373-9381.

[3] Chowdhury A K M R, Mollah M E, Rahman M A. An efficient method for subjectively choosing parameter ‘k’automatically in VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) algorithm[C]//Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. IEEE, 2010, 1: 38-41.

[4] Yu Y, Zhou A. An Improved Algorithm of DBSCAN[J]. Computer Technology and Development, 2011, 21(2).

[5] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark[C]//ACM SIGCOMM Computer Communication Review. ACM, 2005, 35(4): 229-240.

[6] 中国互联网络信息中心（CNNIC） 第 32 次中国互联网络发展状况统计报告 2013.7.17[Online]. http://www.cnnic.cn/

[7] Alexa 中国[Online.]http://top.chinaz.com/.

[8] 中关村在线软件下载排行榜[Online].http://xiazai.zol.com.cn/download_order/soft_order.html

[9] 百度搜索风云榜[Online].http://top.baidu.com