基于RVM的网络流量分类研究

柏　骏[1]，夏靖波[1]，鹿传国[1]，李明辉[2]，任高明[1]

(1. 空军工程大学信息与导航学院　西安　710077；2. 空军后勤部　北京 东城区　100720)

【摘要】将相关向量机(RVM)分类模型应用于网络流量分类问题中。首先对实验数据进行了标准化处理，然后将RVM与其他机器学习算法进行了性能比较，最后在RVM分类结果预测概率中引入置疑区间概念，研究了置疑区间范围及其对分类准确性的影响，并基于此提出了一种新的混合流量分类方法。实验结果表明：1) RVM在准确性等3方面性能指标上优于SVM，且在小样本情况下仍具有较高的分类准确率；2) 置疑区间[0.1,0.9]内的分类预测准确率较低，而置疑区间之外的分类预测准确率在98%以上。

关　键　词　置疑区间;　机器学习;　相关向量机;　流量分类

中图分类号　TP393　　文献标志码　A　　doi:10.3969/j.issn.1001-0548.2014.02.016

Network Traffic Classification Based on Relevant Vector Machine

BAI Jun[1], XIA Jing-bo[1], LU Chuan-guo[1], LI Ming-hui[2], and REN Gao-ming[1]

(1. Information and Navigation College, Air Force Engineering University　Xi'an　710077;

2. Logistics Department of Air Force　Dongcheng Beijing　100720)

Abstract　Relevant vector machine (RVM) is applied in network traffic classification. Firstly, experiment data is standardized, and then RVM is compared with other machine learning tools. Lastly, doubting interval is introduced to analyze predicted probability of classification, based on which a new hybrid traffic classification approach is proposed. Experiment studies illustrate that: 1) RVM excels the support vector machine (SVM) in three performances, and moreover, its classification accuracy is rather high in the situation of small sample circumstances; 2) probabilistic classification in doubting interval has a rather low classification accuracy while an accuracy above 98% outside doubting interval.

Key words　douting interval;　machine learning;　relevant vector machine;　traffic classification

通过准确、实时地分类网络流量，网络运营商能够及时获知当前占据主要带宽的流量应用类型，依此调整网络运行策略以支持其不同的网络运行目标，从而达到优化网络性能、提升服务质量的目的。

目前主要的流量识别、分类技术有4种，分别是端口识别技术[1]、深层数据包检测技术(DPI)[2]、基于网络行为的流量分类技术[3]和基于机器学习的流量分类技术[4]。其中基于机器学习的流量分类方法不受动态端口、载荷特征甚至网络地址转换(NAT)的影响，其效果、灵活性以及可扩展性等，较之前所述各种方法都有所突破，因而受到学者广泛关注，成为学术界的研究热点[5-7]。但其需预先训练分类模型且测试时间相对较长，难以在高速网络中胜任实时分类的任务。

RVM[8]是一种新的机器学习方法，具有与SVM相同的决策形式，通过引入稀疏贝叶斯学习理论，不仅使其具备了SVM避免过学习的优点，且极大地减少了核函数的计算量，还克服了SVM存在的稀疏性不强、计算量大、核函数必须满足Mercer条件以及需人为凭经验确定参数等缺点。因此，与SVM相比，RVM更稀疏，测试时间更短，更适用于在线分类。另外，RVM利用概率模型来解释数据中的噪声，使得RVM具备了预测结果概率的能力，通过概率预测可进一步加深对分类结果的理解。目前RVM已被广泛应用于故障预报[9]、高光谱影像分类[10]等领域中，取得了良好的分类效果。

本文将RVM应用于网络流量分类中，在对样本数据标准化处理的基础上，利用RVM训练流量样本并建立分类模型，以此模型对测试样本进行分类，并输出分类概率。

收稿日期：2012－11－28；修回日期：2013－05－20

基金项目：陕西省科技计划自然基金重点项目(2012JZ8005)；全军军事学研究生课题(2010XXXX-488)

作者简介：柏骏(1985－)，男，博士生，主要从事网络管理与信息安全方面的研究.

1 相关向量机

RVM采用了与SVM相同的决策形式：

$$t_n = y(\boldsymbol{x}_n;\boldsymbol{\omega}) + \varepsilon_n = \sum_{n=1}^{N} \omega_n K(x,\boldsymbol{x}_n) + \omega_0 + \varepsilon_n \quad (1)$$

式中，t_n 为类别标号；$\boldsymbol{x}_n$ 为输入样本向量；$K(x,x_n)$ 是选用的核函数；$\{\omega_n\}_{n=0}^{N}$ 代表不同的权重；ε_n 为噪声，假设其服从均值为零，方差为 τ^{-1} 的高斯分布。

同时引入了贝叶斯概率模型来解释噪声对预测结果的影响。贝叶斯概率为：

$$p(t_*\,|\,t) = \int p(t_*\,|\,w,\sigma^2)\,p(w,\sigma^2\,|\,t)\mathrm{d}w\mathrm{d}\sigma^2 \quad (2)$$

根据贝叶斯准则进行推理，有：

$$p(t_*\,|\,t) = \int p(t_*\,|\,w,\sigma^2)\,p(w\,|\,t,\alpha,\sigma^2)\,p(\alpha,\sigma^2\,|\,t)\mathrm{d}w\mathrm{d}\alpha\mathrm{d}\sigma^2 \approx \int p(t_*\,|\,w,\partial_{\mathrm{MP}},\sigma_{\mathrm{MP}}^2)\,p(w\,|\,t,\partial_{\mathrm{MP}},\sigma_{\mathrm{MP}}^2)\mathrm{d}w \quad (3)$$

用拉普拉斯方法将 $p(w\,|\,t,\partial)$ 近似为高斯分布，有：

$$\partial_i^{\mathrm{new}} = \frac{\gamma_i}{\mu_i^2},\ (\sigma^2)^{\mathrm{new}} = \frac{\|t-\Phi\mu\|^2}{N-\sum_{i=0}^{N}\gamma_i},\quad \gamma_i = 1-\partial_i\sum_{i,i}$$

通过不断地迭代，以逼近 ∂_{MP} 及 σ_{MP}^2。其中大部分 ∂_i 趋近无限大，对应的 ω_i 则为零；少量的 ω_i 稳定趋于有限值，对应的 $\boldsymbol{x}_i$ 即为相关向量(relevant vectors)。

应用回归方法中解决分类问题的Logistic模型：

$$p(t_n = 1\,|\,\boldsymbol{\omega}^{\mathrm{T}}\phi(x_n)) = \frac{1}{1+\exp(-\boldsymbol{\omega}^{\mathrm{T}}\phi(x_n))} \quad (4)$$

有结果预测概率为：

$$p(t\,|\,w) = \prod_{i=1}^{N}\sigma[y(\boldsymbol{x}_n;\boldsymbol{\omega})]^{t_i}\{1-\sigma[y(\boldsymbol{x}_n;\boldsymbol{\omega})]\}^{1-t_i} \quad (5)$$

RVM判别准则为：如果 $p_i = \dfrac{1}{1+\exp(-y)} < 0.5$，则 $t_i = 0$；如果 $p_i = \dfrac{1}{1+\exp(-y)} \geqslant 0.5$，则 $t_i = 1$。

2 实验环境

2.1 实验数据

为了便于对比、分析，本文采用了两个实验数据集：文献[1]中所用的实验数据集Moore_Set和采自某大学校园网的网络出口数据集IN_Set。

文献[11]对于 Moore_Set 进行了详细说明，在此不再赘述。

由于Moore_Set以完整的TCP流作为流量样本，忽略了目前大量存在于网络中的UDP数据包，尤其是占用大量带宽的P2P应用，其UDP数据包甚至占到P2P数据包的60%左右。为此，本文按照五元组，以间隔T(T=120 s)为流结束标志，采集TCP/UDP双向网络流作为流量样本，并按照NetMate[12]中定义的44种网络流量属性特征提取流量特征，产生数据集IN_Set。

IN_Set数据集包含两个数据集，第1个数据集捕获于2012年5月22日14:30～17:30，记为IN_Set$_1$；第2个数据集捕获于2012年9月7日14:30～17:30，记为IN_Set$_2$。两个数据集都用“端口号+DPI”与人工处理相结合的方法对数据集进行应用识别，对于无法识别的流量标记为未知，实验数据中不包括未知流量。数据集的统计信息如表2、表3所示。

表2 IN_Set$_1$数据集的统计信息

流量样本	流量数量	百分比/(%)
WWW	69 101	48.02
FTP	9 176	6.38
MAIL	7 429	5.16
DataBase	5 648	3.93
P2P	40 063	27.84
GAME	938	0.65
SERVER	241	0.17
Unknown	11 307	7.85
Total	143 903	100.00

表3 IN_Set$_2$数据集的统计信息

流量样本	流量数量	百分比/(%)
WWW	72 906	49.08
FTP	8 253	5.56
MAIL	7 134	4.80
DataBase	6 479	4.36
P2P	39 457	26.57
GAME	1 058	0.72
SERVER	461	0.31
Unknown	12 769	8.60
Total	148 517	100.00

2.2 实验工具与工作平台

本文使用RVM_Matlab工具箱[13]、LIBSVM[14]和数据挖掘软件Weka-3.6.4[15]为主要实验工具，对实验结果进行比较分析；实验平台为一台普通PC机，其CPU为Intel Pentium(R) Dual-Core 2.60 GHz，内存为DDR-667 2 GB；运行Windows XP操作系统。

2.3 准确性评价指标

对于某一类样本，其类召回率记为 $\mathrm{recall}_i = \dfrac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}$；对于所有样本，整体准确率为：

$$\mathrm{OverallAccuracy} = \frac{\sum_{i=1}^{m}\mathrm{TP}_i}{\sum_{i=1}^{m}(\mathrm{TP}_i+\mathrm{FN}_i)} \quad (6)$$

整体准确率应用最广，为相关领域研究人员广

泛采纳，它反映了分类模型正确预测样本数在预测总数中的比例。本文主要采用整体准确率作为评价工具。

3　实验结果与分析

3.1　数据预处理

流量特征的选取直接决定了流量分类的准确性与速度。Moore_Set包含的249项网络流属性，IN_Set包含了44项网络流量属性，这些数据中存在众多的冗余属性和无关属性，其存在不仅会降低分类模型的准确率，而且会极大地加重分类模型的计算负载。

利用CFS(correlation-based feature selection)方法对数据集的流量特征进行筛选，选取与分类结果相关性最强且互相关最弱的几项流量特征。但是通过前期实验发现，若直接使用原始数据进行训练、分类，其分类效率低下，准确性不高且相关向量较多。究其原因，主要有两点。首先，在流量特征的统计过程中不同特征采用不同测量单位，各变量大小在数值上差异很大，直接使用原始数据进行分类可能导致信息丢失并引起数值计算的不稳定；其次，大部分流量样本数据过度集中而少量的野值数据偏离度较大，导致RVM难以产生有效相关向量以区分不同类型数据。因此，需对流量特征向量做预处理。

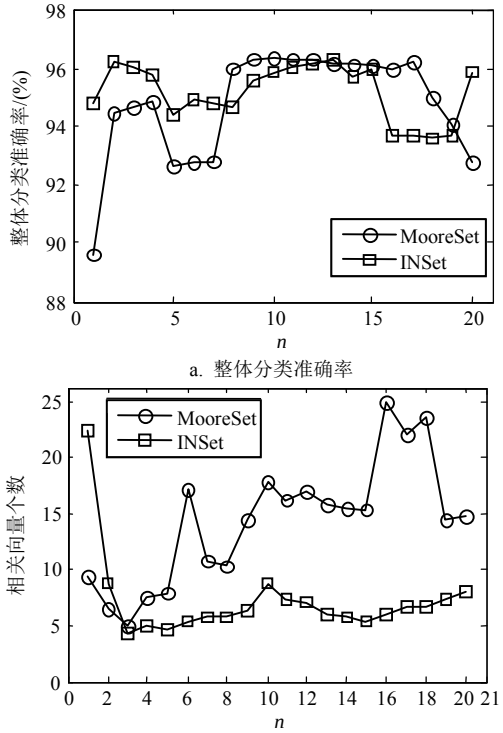
首先考虑采用线性归一化方法对其进行标准化处理，但实验结果发现该方法并没有有效解决少量数据偏离度过大的问题；然而，利用对数函数进行标准化，数据中的零元素以及极小元素会使数据之间的偏离度更大。因此，出于减小数据偏离度、尽量合理分布样本数据的考虑，本文采用如下方法：

$$y=\left(n\frac{x-x_{\min}}{x_{\max}-x_{\min}}\right)^{\frac{1}{n}} \tag{7}$$

对数据进行标准化处理，其中，n为标准化系数。

取$n=1,2,\cdots,20$，按上述方法分别对Moore_Set和IN_Set数据集进行标准化处理，分类准确性和相关向量数量如图1所示。在Moore_Set数据集中，实验中以$entry_{01}$为训练数据，其余9组为测试数据；在IN_Set数据集中，分别将IN_Set_1、IN_Set_2均分为4组，分别记为IN_Set_{1a}、IN_Set_{1b}、IN_Set_{2a}和IN_Set_{2b}，a组作为训练集，b组作为相应测试集。对训练集随机抽样，建立分类模型，两组实验结果都取均值。

当n=1时，即是线性归一化，其分类效果不是很理想，整体分类准确率相对较低，相关向量个数也较多。从图1a中可以看出，两个数据集在准确率上的走势相似，在[8,15]区间内准确率相对稳定，达到96%左右，而在图1b中的[8,15]区间内，相关向量数量也相对较少。这说明实验所采用的数据标准化方法是有效的。因此，在数据标准化过程中，n在[8,15]区间内即可达到令人满意的分类准确率，同时，相关向量数量亦可接受。



a. 整体分类准确率

b. 相关向量个数

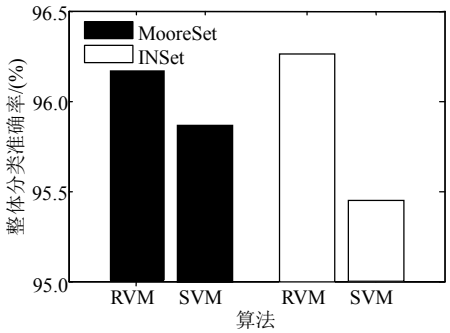
图1　n在[1,20]区间上对整体分类准确率与相关向量个数的影响

3.2　分类性能

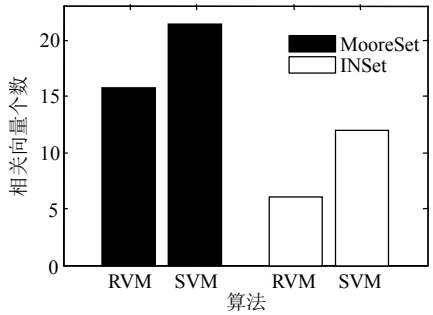
3.2.1　与SVM的性能比较

为了进一步分析RVM与SVM的性能，本文选取n=13对上述两组数据进行标准化处理，分别利用RVM和SVM进行网络流量分类。上述实验重复10次，结果如图2所示。从图2a可以看出，RVM在准确性上稍优于SVM算法，主要是因为SVM需对惩罚因子做出设置以平衡经验风险和置信区间，但试验结果对该数据十分敏感，易引起过学习等问题，从而影响了分类的准确性；RVM相关向量随样本数量的增加而增长的速度较SVM要慢很多，因此在图2b中RVM的相关向量少于SVM的支撑向量；RVM通过不断迭代以获得最优解，然而随着样本数量的增加，其迭代次数显著增加，使得训练时间也迅速增加，这在图2c中的IN_Set数据中可以明显地看出，所以RVM的训练时间长于SVM；RVM与SVM相比更稀

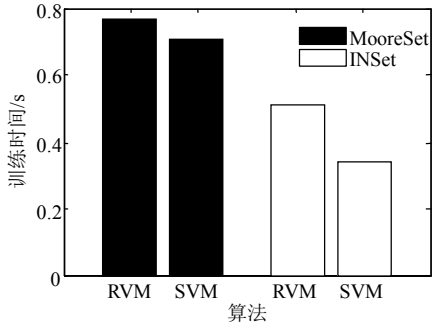
疏，且极大地减少了核函数的计算量，所以测试时间更短，从实验数据上看，前者仅为后者的1%，这个特性使得RVM更适用于在线流量识别、分类。



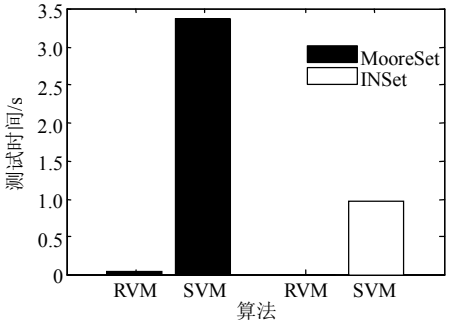
a. 整体分类准确率



b. 相关向量



c. 训练时间



d. 测试时间

图2 RVM与SVM的性能比较

另外，在实验误差方面，RVM分类结果误差非常小，而SVM的误差则相对较大，这说明利用RVM进行分类，其结果更加稳定。造成上述现象的原因在于：SVM是根据结构风险最小化原则寻找最优超平面以确定支撑向量，由于在本文实验中训练样本是通过随机抽样获取，因此每次实验的边缘数据分布不一致使最优超平面存在较大差异，由此引起较大实验误差；而RVM是通过先验概率分布来确定相关向量，因此在保证足够训练样本的情况下，随机抽样对其分类结果的影响很小。

3.2.2 与其他算法的比较

为了比较、分析基于不同机器学习算法流量分类方法的准确性，本文以IN_Set数据集为实验对象。将 IN_Set_{1a} 和 IN_Set_{2a} 作为训练集，IN_Set_{1b} 和 IN_Set_{2b} 作为测试集；使用CFS方法选取流量特征，并按照上述标准化方法将数据集标准化；对训练集 IN_Set_{1a} 分别按照1%、10%、20%和50%抽样，利用RVM、SVM、C4.5和NaiveBayes算法对抽样数据进行训练，建立流量分类模型，并测试分类结果。上述实验重复10次，分类准确性取多次实验均值，实验结果如图3所示。

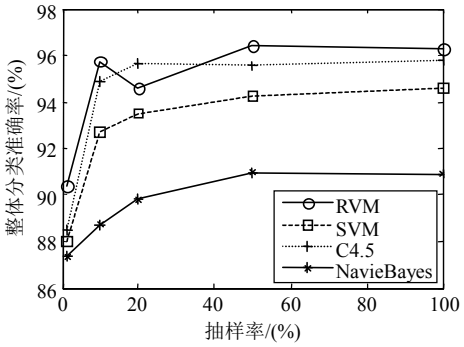


图3 分层抽样下的整体分类准确率

从图3可以发现，在分层抽样下使用RVM对流量进行分类取得了较好的整体分类准确率，分类效果优于SVM和NaiveBayes算法，略优于C4.5算法。在小样本情况下，RVM表现出良好的分类性能，分类准确率达到90%左右；当抽样率为10%时，RVM、SVM和C4.5的分类准确率均有较大提升，此后随着样本的增加，SVM和C4.5分类准确率稳步上升并逐渐趋于平缓。然而随着训练样本的增加，RVM的分类准确率出现了抖动，且有明显下降。这主要是因为利用CFS算法选取流量特征是从全局角度出发，根据全部训练数据的相关性进行筛选，会导致选取的流量特征在抽样数据中达不到最优分类性能，从而引起分类结果的显著变化。而NaiveBayes分类方法是基于条件独立假设和高斯分布假设，在实际网络中难以满足上述假设，因此导致分类结果不理想。

3.3 置疑区间

对于二元分类，RVM不仅可以获得二值输出，

还能得到分类结果的预测概率。预测概率越趋近于0或1，其预测的准确性也就越高。而对于某一区间的预测值，其预测准确性有明显下降，分类结果存在较大的不确定性，可称之为置疑区间。下面给出RVM置疑区间的定义。

置疑区间：设 $p(x;t)$ 为RVM分类预测概率(其中 x 表示分类样本，t 为分类结果)，$e(p)$ 为该预测概率相对应的分类错误概率，对于定值 α $(0<\alpha<1)$，若存在 $\underline{p}$ 和 $\overline{p}$ ($\underline{p}<\overline{p}$)，使得：

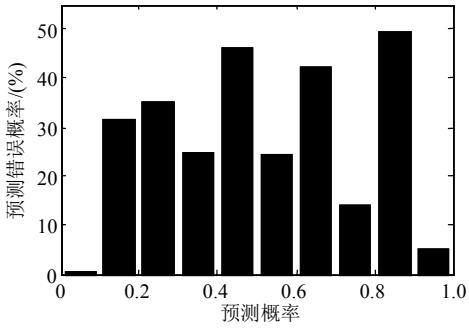
$$e(p)\geqslant 1-\alpha \qquad \underline{p}\leqslant p\geqslant \overline{p} \tag{8}$$

则称区间[$\underline{p}$, $\overline{p}$]是RVM分类预测概率中可疑度为 α 的置疑区间，$\underline{p}$ 和 $\overline{p}$ 分别称为RVM分类预测概率可疑度为 α 的置疑区间的置疑下限和置疑上限。

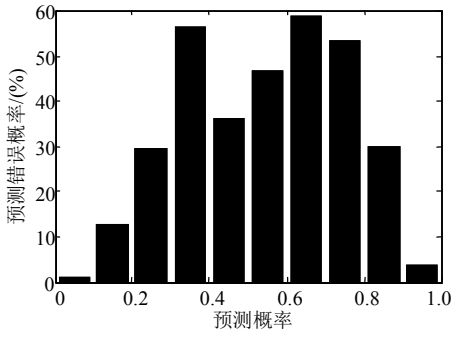
为了研究置疑区间范围及其对分类准确性的影响，本文实验分别以Moore_Set和IN_Set数据集为实验对象。在第一组实验中，将$entry_{01}$作为训练集，其余9个数据集作为测试集；在第二组实验中，以IN_Set_{1a}、IN_Set_{2a}作为训练集，并分别以IN_Set_{1b}和IN_Set_{2b}为相应测试集。实验结果取均值，如图4所示。从图4可以看出，当分类结果预测概率p在[0,0.1]区间时，预测错误概率最低(在Moore_Set中错误率仅为0.3%，在IN_Set中为0.6%)，即分类准确率最高；在[0.9,1]区间内，预测错误率也相对较低(在Moore_Set中错误率为4.85%，在IN_Set中为3.74%)。而在此外的概率区间，预测错误概率明显上升。在Moore_Set数据中，有3个区间段的预测错误率达到了40%以上；在IN_Set数据中，有4个区间段的预测错误率达到了40%以上，其中[0.3,0.4]、[0.6,0.7]和[0.7,0.8]区间的错误率更是在50%以上。

进一步研究发现，上述两组实验预测概率在[0,0.1]和[0.9,1]区间的分类结果占到总样本数的90.47%(Moore_Set)或94.58%(IN_Set)，同时这两个预测概率区间内的分类结果准确率则高达98.79%(Moore_Set)或98.02%(IN_Set)。此外，在[0.1,0.9]区间内的分类准确率仅为65.72%(Moore_Set)或40.12%(IN_Set)，远不能满足分类的准确性需求，与“端口号+DPI”方法90%左右的分类准确率[16]仍相距甚远。因此，本文取 $\alpha=0.9$，根据上述定义，则区间[0.1,0.9]是RVM分类预测概率可疑度为0.9的置疑区间。

基于以上分析，为了获得更高的分类准确率，本文提出一种新的混合流量分类方法：利用RVM对网络流量分类，得到分类结果概率；判断预测概率是否在置疑区间[0.1,0.9]内，如果是，则采用“端口号+DPI”相结合的方法对该流重新进行识别，否则完全采纳其分类结果。利用上述方法进行了初步试验，对于Moore_Set，由于无法进行深度检测，仅使用端口号作为辅助手段，分类准确率达到98.12%；而对于IN_Set，使用“端口号+DPI”的方法，分类准确率达到98.53%。关于上述方法的系统研究将是下一步工作的研究重点。



a. Moore_Set



b. IN_Set

图4　分类错误率在预测概率上的分布

4　结 束 语

随着P2P协议的广泛应用，基于机器学习的网络流量方法在识别载荷加密流量上的优点引起学者的广泛重视，成为学术界研究热点。

本文在分析了相关向量机原理的基础上，给出了RVM分类模型，并研究了基于RVM的网络流量分类方法。通过Moore_Set与IN_Set两组数据的对比实验，可以发现RVM在准确率、相关向量数量及测试时间上，都要优于支持向量机方法(SVM)；分层抽样实验表明RVM稳定性较好，在小样本情况下仍具有较高的分类准确率；研究、分析了RVM分类结果的置疑区间范围及其对分类准确性的影响，并提出了一种新的混合流量分类方法，初步实验表明该方法的分类准确率能达到98%以上。

在本文实验中遇到的一些问题还需进一步深入研究，最后提出的混合流量分类方法仍需系统地论

证、分析。与此同时，RVM测试时间快的特点尤其适用于在线流量识别，这也是下一步工作的重点。

参 考 文 献

[1] TOUCH J, LEAR E, MANKIN A, et al. Internet assigned numbers authority (IANA)[EB/OL]. [2010-08-28]. http://www.iana.org/assignments/port-numbers.

[2] SEN S, SPATSCHECK O, WAND D. Accurate, scalable in-network identification of P2P traffic using application signatures[C]//Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters (WWW'04). New York, USA: ACM, 2004: 512-521.

[3] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: Multilevel traffic classification in the dark[C]//ACM SIGCOMM. Phila-delphia, PA, USA: ACM, 2005.

[4] ZUEV D, MOORE A W. Traffic classification using a statistical approach[C]//Proceedings of PAM 2005. Boston, USA: [s.n.], 2005.

[5] DAINOTTI. A, PESCAPE. A, SANSONE C. Issues and future directions in traffic classification[J]. IEEE Network, 2012, 26(1): 35-40.

[6] GU Cheng-jie, ZHANG Shun-yi, SUN Yan-fei. Real-time encrypted traffic identification using machine learning[J]. Journal of Software, 2011, 6(6): 1009-1016.

[7] HURLEY J, GARCI-PALACIOS E, SEZER S. Classifying network protocols: a 'two-way' flow approach[J]. Communications, 2012, 5(1): 19-89.

[8] TIPPING M. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001, 1(1): 211-244.

[9] 胡昌华, 王兆强, 周志杰, 等. 一种RVM模糊模型辨识方法及在故障预报中的应用[J]. 自动化学报, 2011, 37(4): 503-512.
HU Chang-hua, WANG Zhao-qiang, ZHOU Zhi-jie, et al. An RVM fuzzy model identification method and its application to fault prediction[J]. Acta Automatica Sinica, 2011, 37(4): 503-512.

[10] ZHANG Mian-ji, ROBUST Y. Hyperspectral classification using relevance vector machine[J]. Geoscience and Remote Sensing, 2011, 49(6): 2100-2112.

[11] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[C]//Proc of the 2005 ACM SIGMETRICS Int'l Conf on Measurement and Modeling of Computer Systems. Banff: ACM, 2005: 50-60.

[12] ZANDER S. NetMate meter[EB/OL]. [2013-04-08]. http://sourceforge.net/projects/netmate-meter/.

[13] TIPPING M. Sparse Bayesian models (and the RVM) [EB/OL]. [2006-10-12]. http://www.relevancevector.com.

[14] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[EBPOL]. [2007-08-06]. http://www. csie. ntu.edu. tw/～cjlin/libsvm.

[15] DURRANT B, FRANK E, HUNT L, et al. Waikato environment for knowledge analysis (WEKA)[EB/OL]. [2012-10-18]. http://www.cs.waikato.ac.nz/ml/weka/.

[16] CALLADO A, SZAB'O, GERO C K G, et al. A survey on internet traffic identification[J]. IEEE Communications Surveys & Tutorials, 2009, 11(3): 1-16.

编 辑 漆 蓉