基于流记录偏好度的多分类器融合流量识别模型

董仕[1,2,3]，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和信息集成教育部重点实验室,江苏 南京 211189;
3.周口师范学院计算机科学与技术学院,河南 周口 466001)

摘 要：通过将证据理论引入到流量分类的决策模块中，提出了偏好性和时效性权值，并通过实测数据对多分类器识别模型进行验证，其结果表明该模型较好的克服了单分类器的片面性，通过对多个证据的融合来优化识别的结果。

关键词：多分类器融合；证据理论；偏好性；机器学习

中图分类号：TP393.08 **文献标识码**：A **文章编号**：1000-436X(2013)

Traffic classification model based on fusion of multiple classifiers with flow preference

DONG Shi[1, 2, 3], DING Wei[1, 2]

(1.School of Computer Science and Engineering，Southeast University, Nanjing 211189 China;
2. Key Laboratory of Computer Network and Information Integration
(Southeast University), Ministry of Education, Nanjing 211189, China;
3. School of Computer Science and Technology，Zhoukou Normal University，Zhoukou 466001, China)

Abstract: Introduced the concept of multi-classifier fusion which can improve the classification accuracy and overcome the disadvantage of single classifier. This paper introduced DS theory into decision module of traffic classification and proposed preference and timeliness. After analyzing multi-classifier model by simulation, the results show the new classifier model can overcome one sidedness of single classifier, depend on multiple evidences to optimize the traffic results.

Key words: Multi-classifier; DS theory; Preference; Machine learning

1 引言

随着网络带宽不断增长，各种新的网络应用不断产生，网络流量识别作为网络管理中一个研究热点方向逐渐受到国内外研究人员的关注。在机器学习的流量识别方法中为了获取更加精准的识别精度，更高的识别效率，就需要对分类的数据进行流量的特征选择，把对分类精度影响很大的特征属性通过量化以及有效评估的方式选择出来。目前常见的单分类器的流量识别方法很多，例如：BAYES 神经网络，SVM，C4.5 决策树等方法[1~9]，但是对于不同的样本各算法又存在不同的适应度，因而为了解决单分类器的这种片面性问题，本文提出一种基于多分类器融合的流量识别模型。一方面可以克服单分类器所存在的适应度的缺陷，另一方面也可以提高分类识别的精度。目前，很多研究都是基于单分类器的，且其分类性能的提升已经到达一定的瓶颈，而对多分类器的流量识别的研究比较稀少。不同的单分类器在处理数据噪声问题时有不同的效率，因此当面临不同的网络流数据的时候可能造成因对噪声数据的弱处理而带来单分类器本身分类效果的下降。并且对于在线的流量识别以及抽样对数据分类影响也是目前需要解决的问题，因此本文旨在提出一种基于多分类器融合的流量分类模型，用于解决由于数据噪声和抽样对分类器识别结果的影响，并依据此流量模型对 CERNET 网络中的

收稿日期：2012-07-13；**修回日期**：2013-05-13；
基金项目：国家重点基础研究发展计划(“973”计划)基金资助项目；国家科技支撑计划基金资助项目
Foundation Items: The National Basic Research Program of China(973 Program) (2009CB320505); The National Science and Technology Plan Projects (2008BAH37B04)

数据进行分析。实验结果表明：采用本模型能有效的降低抽样所带来的识别结果抖动的现象，并能有效的解决由数据噪声所带来的识别结果不稳定的问题，且与单分类器相比有较高的识别精度和较低的分类错误率。

本文内容的组织结构如下：第 2 节讲述了相关的研究，并对相关内容进行了进一步分析和讨论。第 3 节提出了多分类器融合的流量识别模型。第 4 节进一步对模型中核心概念和算法进行描述。第 5 至 6 节对实验结果进行评估分析，得出相关的结论并提出展望。

2 相关工作和存在问题

机器学习识别的目标是通过对样本数据的学习来构建学习分类器，然后通过所构建的分类器对测试样本进行分类。机器学习的方法引入到网络流量识别领域是为了解决深度报文检测方法涉及隐私的问题。目前提出的应用识别模型[10~14]可将 Internet 流量划分至 10 个应用类别，准确率达到近 95%甚至更高。但是这些算法仍然存在着一些不足：

第一，所用测度复杂。现有的高精度协议识别方法均使用了 248 个以上的流行为测度，从简单的传输层端口号到复杂的报文首部傅立叶变换，从而导致方法对报文采集及测度计算系统的要求很高，难以达到易用的效果。

第二，识别方法复杂。文献[11]使用了核密度估计(KE, kernel estimation)的方法对 Bayes 方法的变量正态分布假设进行修正。但是，KE 方法时间复杂度很高，且随训练样本流数量线性增长，因此存在着严重的效率-精度的矛盾。文献[13]使用了 Bayesian 神经网络试图进一步提高应用协议识别的精度。但是进一步上升的时空复杂度使其应用识别过程只能离线进行。并且文献[13]中所提算法在训练阶段的时间开销也很大，不可能应用到目前 10Gbps 以上的网络主干信道中。文献[14]中使用了高效的 C4.5 判别树分类方法，使得单流处理时间达到约 553.73 μs。但当前 CERNET 江苏省边界 10Gbps 信道的每小时并发流数约为 60M，即单流处理时间必须小于 60 μs，这样的实时识别高要求和目前的研究成果仍有一个数量级的差距。

第三，对流数较少的应用类别，现有方法的识别准确率很低，从而使得应用类别间的平均识别准确率均低于 69%。更重要的是，现有研究中各应用的识别准确率与该应用在训练样本中所占的比例密切相关。文献[13]中表明，增加某应用类别的训练样本数可以提高该应用的识别精度，但会降低其他应用类型甚至整体流量的识别准确率。但是从理论上说，只要协议训练样本的数量足以刻画该协议的分布情况，所选用的协议识别方法就应该可以准确地标识该协议，而与其在总训练样本中所占的流比重应完全无关；增加某个协议的训练样例应可以细化其行为特征并提高该协议的识别准确率，但不应降低其他协议的和总体流量的识别准确率。

第四，目前绝大多数研究只针对 TCP 流量甚至完整 TCP 流量进行应用协议识别，然对现有 Trace 和当前 CERNET 网络环境的监测可知，目前 CERNET 主干网和 Internet 上所承载的 UDP 流量已经超过 50%，而完整 TCP 流仅占网络总流数的约 10%。因此，仅识别完整 TCP 流具有较大的片面性，且方法不具有实用意义，有必要将应用流量识别对象扩展至使用 TCP 和 UDP 协议传输的所有流量。文献[14]对 UDP 上的应用协议流量识别进行了一些讨论，但其所标识的对象也仅为完整 UDP 流，不仅流量极少，且应用协议情况简单。

因此为了解决上述的问题，本文提出了一种基于多分类器融合的流量识别模型。该模型有以下特点：1)与常见的单分类器算法相比，基于多分类器模型的识别方法增加了决策力度；2)引入偏好度和时效度等概念能充分发挥各分类器的优势，从而获取更准确和高效的识别结果；3)采用证据理论对分类结果进行融合能更好的提高分类的准确性。

3 多分类器流量识别模型

模型的相关概念和流程如下。

定义 1 流量统计特征：它是由带有流超时的流记录特征组成，所谓的流特征是通过对 TRACE 数据组流计算所得。具体的表现形式为

$$T_{nm}=\begin{pmatrix} T_{11} & \mathrm{K} & T_{1m} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ T_{n1} & \mathrm{L} & T_{nm} \end{pmatrix}$$

其中，$n>0$,$m>0$。

目前，多分类器类型大致分为 2 类。

定义 2 同类分类器：具有相同的分类器所组成的混合分类器，$C=\{c1,c2,c3\mathrm{L}\ ,cn\}$。

定义 3 多样性分类器：由不同类型分类器所组成的混合分类器，$C=\{a1,b2,c3\mathrm{L}\ ,fn\}$。本文为

了体现不同分类器间对分类结果的影响，以多样性分类器为研究对象。

决策类型：是由均值决策，贝叶斯决策和投票决策等共同组成。

定义 4　贝叶斯决策：根据贝叶斯理论，在 T_{nm} 已知的情况下，判别 T_n 属于的类别。假设 x 为待识别的对象，而识别类别为 $C_1, C_2, \mathrm{L}\ C_M$，则 C_M 要满足如下条件

$$IFP(C_m \mid t_1, t_2, \mathrm{L}\ , t_n) = \max P(C_j \mid t_1, t_2, \mathrm{L}\ , t_n)\ THEN\ x \in C_m$$

必须计算输入在不同假设下的概率，然后选取最大可能概率的类别作为该输入所属类别。其中

$$P(C_j \mid t_1, t_2, \mathrm{L}\ , t_n) = \frac{P(t_1, t_2, \mathrm{L}\ , t_n \mid C_j) P(C_j)}{P(t_1, t_2, \mathrm{L}\ , t_n)}$$

$$P(t_1, t_2, \mathrm{L}\ , t_n) = \sum\ P(t_1, t_2, \mathrm{L}\ , t_n \mid C_j) P(C_j)$$

定义 5　均值决策：使用各基分类器输出均值。即为：$P(C_m \mid t) = \frac{1}{n} \sum_{k=1}^{k} P_k(C_m \mid t)$

$$H(x) = \underset{j=1}{Max} \left(\sum_{i=1}^{n} P_{ij}(C_m \mid t) \right)^{l}$$

对每个类，各基分类器都会给出属于该类的概率，并求归属每类概率的均值，取均值最大者的类标为最终的分类结果。

定义 6　投票决策：根据每个分类器的权重来决定投票所占的比例。

$$H(x) = \underset{j=1}{\overset{m}{Max}} \sum_{i=1}^{T} (w_i * c_{i,j})$$

上式中的 m 代表类别数，T 代表分类器的个数，w_i 代表分类器的权重。$c_{i,j}$ 表示第 i 个分类器在第 j 个类别中的分类结果。如果 x 被第 i 个分类器所识别为 j 类，则 $c_{i,j}$=1 否则 $c_{i,j}$=0

D-S 证据理论：

D-S 证据理论是一种不确定性的推理与决策过程，在决策层融合算法中得到广泛的应用。[15]

D-S 证据理论具有较强的理论基础，针对实际的应用则要求证据之间严格的独立，证据的合并规则也被证明为 P 完全难解问题[17]。冲突证据可能引发悖论。在本文所描述的应用场景中不同的流量类别之间是相互独立的，满足证据理论的先决条件。

证据理论中需要提到下面相关概念：

假定存在互不相容的命题集合，为一有限集合，记作 $\Theta = \{A_1, A_2, \mathrm{L}\ , A_n\}$，$\Theta$ 为一待辩识框架。其中 A_i 代表命题“待识别流量属于第 i 类”。

概率分配函数：若存在从 Θ 的幂集到[0,1]区间的映射 $m: P(\Theta) \to [0,1]$，且满足

$$\begin{cases} m(\phi) = 0 \\ \sum_{A \in P(\Theta)} m(A) = 1 \end{cases}$$

那么称 $m(A)$ 为概率分配函数。

信任函数：

$$BEL(A) = \sum_{B \in A} m(B), \forall A \subseteq \Theta$$

似然函数：

$$PL(A) = 1 - BEL(\overline{A}) = \sum_{B \cap A \neq \phi} m(B)$$

其中，*BEL* 为下限函数，表示命题成立的最小不确定性函数，*PL* 函数为上限函数或不否定函数。表示非假的信任程度的不确定性度量。利用 *BEL* 和 *PL* 的定义可以生成称为证据间隔的量。

定义 7　证据间隔

$$EI = [BEL(B), PL(B)]$$

因此可以输出一系列的证据间隔的集合：

$$\begin{bmatrix} BEL(A_1) & PL(A_1) \\ BEL(A_2) & PL(A_2) \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ BEL(A_K) & PL(A_K) \end{bmatrix}$$

定义 8　*Dempster* 规则形式化定义

设 m_1 和 m_2 为 2^{Θ} 上相互独立的基本概率分配，则这 2 个证据的基本概率分配定义为

$$m(A) = \frac{\sum_{A_1 \cap A_2 = A} m_1(A_1) m_2(A_2)}{1 - K}, A \neq \phi$$

其中，K 为归一化因子，$K = \sum_{A_1 \cap A_2 = A} m_1(A_1) m_2(A_2)$。

可将 2 个证据的组合进行推广。推导出 n 个证据进行组合的 *Dempster* 的一般式。

$$m(A) = \frac{\sum_{A_1 \cap A_2 \cap \ldots \cap A_n = A} m_1(A_1) m_2(A_2) \mathrm{L}\ m_n(A_n)}{1 - K_n}, A \neq \phi,$$

K_n 为归一化因子。$K_n = \sum_{A_1 \cap A_2 \cap \ldots \cap A_n = A} m_1(A_1) m_2(A_2) \mathrm{L}\ m_n(A_n)$

采用证据距离方法：

在有 n 个证据的辨识框架中，假定 m_1，m_2 为概

率分配函数，则 m_1 和 m_2 之间的距离表示为 $d(m_1,m_2)=\sqrt{\frac{1}{2}(\|\vec{m}_1\|^2+\|\vec{m}_2\|^2-2<\vec{m}_1,\vec{m}_2>)}$ 其中 $|\vec{m}_1|^2$ 表示 $|\vec{m}_1|^2=<\vec{m}_1,\vec{m}_1>$ 为两向量的内积。从 n 个分类器中得到 n 个证据，每个证据为一 m 维的向量。使用如下矩阵表示

$$DM=\begin{bmatrix}0 & d_{12} & \cdots & d_{1m}\\ d_{21} & 0 & \cdots & d_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ d_{n1} & d_{n2} & \cdots & 0\end{bmatrix} \tag{1}$$

根据专家经验和多次实验结果得出阈值 ε

则根据下式中 d_{ij} 与 ε 关系来确定对证据的相似性的判定。

$$de_{ij}=\begin{cases}1, de_{ij}\le\varepsilon\\ 0, de_{ij}>\varepsilon\end{cases} \tag{2}$$

从式(1)经过式(2) 演变为证据相似性矩阵

$$DE=\begin{bmatrix}1 & de_{12} & \cdots & de_{1m}\\ de_{21} & 1 & \cdots & de_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ de_{n1} & de_{n2} & \cdots & 1\end{bmatrix} \tag{3}$$

在式(3)中若 de_{ij} 为 0，且 $0<i<1,0<j<1$ 则表示第 i 个分类器和第 j 个分类器的证据有冲突，即为具有不相似性。会降低融合决策精度，若 de_{ij} 为 1，且 $0<i<1,0<j<1$，则表示第 i 个分类器和第 j 个分类器的证据是相似的，相互支持，这样会增加融合决策精度。考虑多分类器的偏好性:即为不同的分类器对同种流量的产生不同的识别精度。因为考虑到各基分类器的差异性以及对于分类样本的偏好性。这样就需要考虑到在决策过程中加入权重的概念，而这个权重的概念是根据不同分类器的偏好性所确定的。例如分类器 A 对 t_n 类具有较高的偏好性，那么就赋予比较高的权重。例如下面章节中将详细讲述分类器对于流量分类的偏好性研究。

融合流量识别模型的逻辑结构（如图 1 所示），

此流量识别模型大致分为三层：由输入层，分类层，决策层组成。

1) 输入层：主要把流量的统计特征作为输入层的输入数据。主要由在第 5 节中的测度属性所组成。

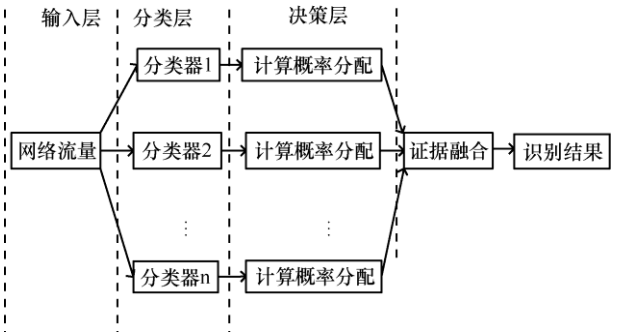


图 1　多分类器融合流量识别模型

2) 分类层：这一层是分类融合模型的核心层，主要由各基分类器构成。

3) 决策层：决策层主要依据评估机制来完成最后的抉择，例如投票机制等。

流量分类器

定义 9　偏好度定义：

为了定量分析流量分类器对各分类的偏好性，现引入偏好度 H

$$H=\frac{\mathrm{Pr}ecision}{\mathrm{Re}call}\times 100\%$$

根据公式 8、9 可以推出：

$$H=\frac{TP+FN}{TP+FP}\times 100\%=1+\frac{FN-FP}{TP+FP}\times 100\% \tag{4}$$

其中，Precision 和 Recall 是机器学习算法评估测度，详细定义在第 5 节中。

命题 1　流量分类器对每个类标的偏好度随着 FN 误报率增大而增大，随 FP 的增大而减小。

证明　对于任意多个分类器 C，设共有 M 个分类器，且每个分类器经过训练而得到的 $FN=\{FN_1,FN_2,\ldots,FN_M\}$，其中最大的 FN 误报率记作 $F_{\max}=\max\{FN_k,0<k\le M\}$，最小的 FP 误报率记作 $F_{\min}=\min\{FP_k,0<k\le M\}$。依据式(4)中所定义的偏好度可知，当 FN 和 FP 相对独立时，取 $F_{\max}$ 和 $F_{\min}$ 使得偏好度 H 值最大化。对于任意的 $0<k\le M$，都有 $FN_k\le F_{\max}$；$FP_k\ge F_{\min}$。当且仅当 $FN_k=F_{\max}$ 且 $FP_k=F_{\min}$ 时，H 取得最大值。证毕。

H 是由 $\{h_1,h_2,\ldots,h_n\}$ 构成。通过大量的实验和相关文献[18]表明查准率和查全率之间存在相反的依赖关系，如果提高输出查准率，就会降低查全率。反之亦然。

定义 10　时效度定义。

为了考虑到能用于高速在线的流量识别，因此在模型中引入时效度 T

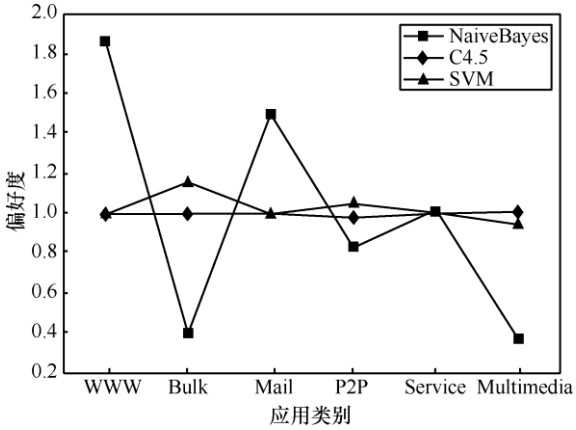


图 2　不同分类器对不同类别的偏好性

$$T = Training\_time + Classification\_time$$

T 是由 $\{t_1, t_2, \mathrm{L}\ , t_n\}$ 构成。不同的分类器对同一个样本有不同的时间效率，当我们在设计融合的分类器的时候需要考虑这一因素。从图 3 中可以看到不同数据集的分类器的时效性的对比。

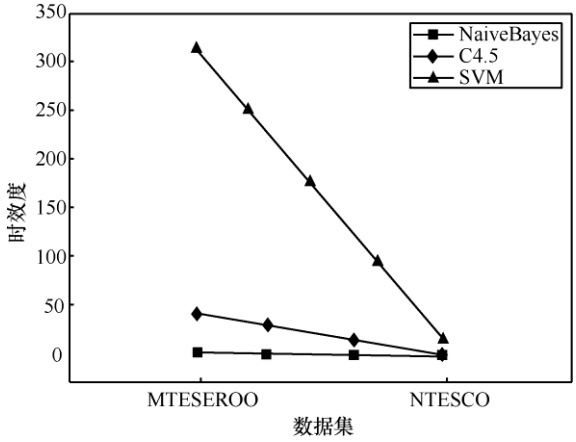


图 3　不同分类器作用于不同数据集的时效度对比

定义 11　动态权值矩阵

$$w = \begin{pmatrix} w_{11} & \mathrm{K} & w_{1m} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ w_{n1} & \mathrm{L} & w_{nm} \end{pmatrix} \tag{5}$$

一共有 n 个基分类器，训练的类别为 m 个，那么 w 即为 n 个基分类器在 m 个类别上的权值矩阵。其中 $w_{nm} = \dfrac{H}{T}$。并对权值进行归一化处理，采用公式 5 获取归一化后的权值 $w_{nm(new)}$。

$$w_{nm(new)} = \frac{w_{nm} - w_{\min}}{w_{\max} - w_{\min}} \tag{6}$$

考虑到本文所给出的流量分类器偏好性以及时效度因素,引入基于动态权值矩阵的 D-S 融合方法，因此 D-S 证据理论规则演变为

$$m(A) = \frac{\sum_{A_1 \cap A_2 \ldots \cap A_m = A} w_1 m_1(A_1).w_2 m_2(A_2) \ldots w_m m_m(A_m)}{1 - K_m}, A \neq \phi \tag{7}$$

其中

$$K_m = \sum_{A_1 \cap A_2 \ldots \cap A_m = A} w_1 m_1(A_1).w_2 m_2(A_2) \ldots w_m m_m(A_m)$$

$$w_1 = w_{11(new)} {}^{*} w_{21(new)} {}^{*}, \ldots {}^{*} w_{m1(new)}$$

$$w_2 = w_{12(new)} {}^{*} w_{22(new)} {}^{*}, \ldots {}^{*} w_{m2(new)}$$

$$w_m = w_{1m(new)} {}^{*} w_{2m(new)} {}^{*}, \ldots {}^{*} w_{mm(new)}$$

根据本文所设置的场景，对于输入的流记录 F，若有 $A_i = F \in i(i = 1, 2, \mathrm{L}\ m)$，这样就构成 m 个相互独立的命题，这样就符合了证据理论所需要的假设条件，所有命题构成的待辨识别窗口为 $\Theta = \{A_1, A_2, \mathrm{L}\ , A_m\}$。$n$ 个分类器 $C_1\ C_2 \ldots C_n$，可以得到 n 个证据 $C_k(F) = c_j$ 其中 $j = 1, 2, \mathrm{L}\ , n$，其表达式表示流记录 F 经过分类器分类后的类别为 $c_j \in \{1, 2, \mathrm{L}\ , m\}$。分类器 C_k，设其查准率为 $\varepsilon_r^k = \varphi_k(A_j^k)$，误判率为 $\varepsilon_s^k = \varphi_k(\neg A_j^k)$。因此存在 $\varphi_k(\Theta) = 1 - \varepsilon_r^k - \varepsilon_s^k$，对于所有的证据 $C_k(F) = c_j$ 就存在 n 个 φ_k，决策层的关键是通过证据理论获取 $\varphi = \varphi_1 \oplus \varphi_2 \oplus \cdots \oplus \varphi_n$，计算 n 个证据对于 $\forall i \in \{1, 2, \mathrm{L}\ , m\}$ 的信任度 $bel(\{Ai\})$。

最终的分类结果可以根据下式来确定：

$$E(x) = \begin{cases} j, & bel(\{Aj\}) = \max bel(\{Ai\}) \geqslant a \\ M + 1, & \text{其他} \end{cases}$$

其中，$0 < a \leqslant 1$

4　分类器融合识别算法

融合分类器在决策层引入改进的证据理论融合策略，在分类层所采用的基分类器都是可扩展的，设各基分类器为 $C_1, C_2, C_3, \mathrm{L}\ , C_n$，采用并行的策略，首先考虑依据流量分类器的偏好性进行样本的偏好性训练，例 C_1，C_2，C_3 分别对 WWW，P2P，Bulk 有较高的偏好度。并求出每个分类器的时效度 T，得到动态权值矩阵 w，从 n 个不同的分类器中获取证据，并计算证据之间的距离，得到相容性矩阵，若不相容则摈弃其中一条证据，若相容则利用

基于动态权值融合公式进行融合。

预处理模块：

输入：拥有特定流测度的流记录

输出：每个分类器证据和偏好度

过程：对标准数据集进行训练得到不同分类器对不同类别的偏好性以及时间度，计算 $w_{nm}=\frac{H}{T}$，获取动态权值矩阵，并对权值进行归一化处理。从不同分类器中获得证据

算法 1

Input: flow record $FR=\{f_1, f_2, \mathrm{L}, f_n\}$,Classifier $C=\{c_1, c_2, \mathrm{L}, c_k\}$

Output: flow record with label

1) FOR (flow record FR) DO

2) FOR (each classifier c_k) DO

3) H_i =Get_Preference(training_samples,appi);

4)/*获取偏好性*/

5) T_i = Get_Timedegree(training_samples,appi);

6)/*获取时间度*/

7) W = Get_W(training_samples, c_k , H_i , T_i);

8)/*求出动态权值矩阵*/

9) E =Get_E(training_samples, c_k);

10) ENDFOR ENDFOR

融合模块：

描述如下所示：

输入：拥有特定流测度的流记录

输出：带有标签的流记录

过程：求出证据之间间隔距离，得到相容矩阵，根据相容性来决定是去掉冗余的证据还是采用动态权值融合公式来求解。这样通过证据理论的组合规则来完成证据融合。下面是伪代码:

算法 2

Input: flow record $FR=\{f_1, f_2, ..., f_n\}$,

Classifier $C=\{c_1, c_2, ..., c_k\}$

Output: flow record with label

1) FOR (each flow record f_n) DO

2) FOR (each classifier c_k) DO

3)Cluster (label_result) = {E1, E2, …El}

4)If (*Ei* in set {*E1,E2,...El*})

5)/*获取每个分类器的证据*/

6) IF(de_{ij}≤ε)

7)K++;/* k 表示相似证据个数；

8) FOR (each classifier C_k)

9) Compute value of $E(f_n)$;

10)Else IF(de_{ij}>ε)

11)删除融合分类器；

12) ENDIF ENDFOR ENDFOR ENDFOR

算法复杂度分析:

多分类器融合识别算法的时间复杂度为流数 n 和分类器数目 k 乘积的线性函数，即为 O(n*k)，因此总的算法复杂度为 O(n*k)，由于在进行分类器融合之前，需要计算并存储 k*n 个数值，其中 n 代表类别个数，存储每个分类器所产生的证据需要 K 个数值，因此总的空间复杂度为 O(n*k)+O(k)。

5 实验与结果分析

5.1 测度属性

目前研究所采用的数据大部分是针对全报文采集的数据，这样可以从中获取更多的信息,能更准确的对流量进行识别和分类。但是这些分类识别目前只能通过离线采集数据，然后再通过在线的方式进行识别，这样识别效率比较低。而且采用的属性特征大部分是针对 TCP 协议的，基于 P2P 应用的 UDP 协议流量很大一部分不能有效识别，在本文中提出既针对 TCP 流又适用于 UDP 流的测度，具体见表 1 所示。由于 NETFLOW 流的出现，我们尽量考虑采用 NETFLOW 固有流的属性来实现流量识别，这样可以减少很大的流量负载所带来的压力，又可以提高识别效率，真正实现在线的流量识别。鉴于此，本文考虑采用 NETFLOW 以及扩展的 NETFLOW 流记录作为研究对象。

在具体介绍模型之前，先引入本文中所要使用的带扩展的 NETFLOW 流记录以及识别的应用类型目标的描述：

定义 11 带扩展的 NETFLOW 流记录描述：$X=(x_1, x_2, x_3, ..., x_t)$；

定义 12 应用类型识别目标集合的描述：

$$Y=F(X)=(y_1, y_2, y_3, ..., y_n);$$

通过样本数据可以确定函数的参数，而分类器就是函数 $F(X)$ 本身。

注：表 1 中共列出了 16 种测度，其中有 5 个测度可以在 NETFLOW 中直接得到，而其余的部分需要进行相应的计算。

表 1　　测度描述

测度	测度描述
双向报文数	前向和后向的报文数之和
双向字节数	前向和后向的字节数之和
平均报文长度	双向字节数/双向报文数
持续时间	流结束时间-流开始时间
TOS	NETFLOW 中双向 TOS 之 OR
TCPFLAGS1	某一方向流的 TCPFLAGS；
TCPFLAGS2	另一方向流的 TCPFLAGS；
传输层协议	NETFLOW 直接得到
低位端口	NETFLOW 直接得到
高位端口	NETFLOW 直接得到
PPS	报文数/持续时间
BPS	字节数/持续时间
平均报文到达间隔	持续时间/报文数
双向报文数比	流中双向报文数的比
双向字节数比	流中双向字节数的比
双向报文长度比	流中双向报文长度的比

5.2　机器学习算法评估

目前流量识别算法有效性的评估标准有如下几个概念：

混淆矩阵(confusion matrix) ：它是分类结果的一种输出方式，能够表征预测结果和标准数据之间的对应关系。具体形式如表 2 所示。其中 h_{ij} 代表实际类型为 i 被分类模型判断为 j 的样本数。

表 2　　混淆矩阵

		分类结果			
		Class 1	*Class 2*	.....	*Class n*
正确的分类	*Class 1*	h_{11}	h_{12}	.....	h_{1n}
	Class 2	h_{21}	h_{22}	.....	h_{2n}
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	Class n	h_{n1}	h_{n2}	.....	h_{nn}

真正 TP(true positive)：实际类型为 i 的样本中被分类模型正确预测的样本数 $TP_i = h_{ii}$

假正 FP(false positive)：实际类型为非 i 的样本中被分类模型误判为类型 i 的样本数量 $FP_i = \sum_{j \neq i} h_{ji}$

假负 FN(false negative)：实际类型为 i 的样本中被分类模型误判为其他类型的样本数 $FN_i = \sum_{i \neq j} h_{ij}$

Precision（查准率）

$$\text{Precision}=\frac{TP_i}{TP_i + FP_i} \tag{8}$$

Recall （查全率）

$$\text{Recall}=\frac{TP_i}{TP_i + FN_i} \tag{9}$$

Overall accuracy(整体准确率)

$$\text{Overall accuracy=OA}=\frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \tag{10}$$

5.3　实验数据

实验数据的采集分两类：1)本地端系统抓包数据(.pcap 文件) ；2)IPTAS：江苏省网边界路由器的 IPTrace 数据。

本地抓分组数据

本文主要研究了几种应用：WWW，Bulk，Mail，P2P，Service，Interactive，Multimedia，Voice。对 8 种应用分别抓分组。

IPTrace 数据

trace 数据共 3 组：第 1 组采集于 2010 年 5 月 18 号 00:00 ~ 1:00，第 2 组采集于当天的 1:00 ~ 2:00，第 3 组采集于当天的 19:00 ~ 20:00，江苏省教育网边界到 CERNET 国家主干路由之间，每个报文均为 68 字节，前 8 字节为时间戳，后 60 字节为截取的报文长度；时间戳中 usec 的最后 1bit 为报文方向标志(0、1 分别表示出、进江苏省网)。由于信道吞吐量大，也为了保证 IP 流的完整性，报文抽样采用流抽样，抽样比为 1/4，抽样比的计算条件是所有被使用的 IP 地址的最后 3 位（bit）呈均匀分布。总体 trace 的 3 组数据具体情况如表 3 所示。

表 3 中三段不同的 trace 采集于一天当中的不同时段，体现出了人们行为作息对流量的影响：比如晚上 7-8 点为上网高峰期，流量大，而凌晨 0:00-1:00 相对较少，5-6 点则更少，此时大部分用户正在睡眠中，因此流量极少。利用改进的 L7-filter[16]软件对本地数据和 IPTRACE 数据进行打标签，并形成 NOC_SET 数据集。如表 4 所示。其中流数采用 NETFLOW 组流规范对 Trace 组流得到，超时采用 16s。

5.4　实验结果与分析

本文分别对基于 NOC_SET 数据分别用 4 种分

类器算法进行实验，采用多分类器模型算法（采用 C4.5 决策树、朴素贝叶斯、SVM 三种分类器进行融合）采用经典的机器学习三种单分类器。评估验证采用十折交叉验证对数据进行交叉验证。十折交叉验证法是常用的精度测试方法，它的基本思想是将数据集分成 10 份，轮流将其中的 9 份作为训练数据，1 份作为测试数据，进行实验。每次实验都会得出相应的正确率，10 次结果的正确率的平均值作为对算法精度的估计。通过对本文中数据进行交叉验证并通过分类算法进行评估后，所得结果如下面各图表所示。

表 3　　IPTrace 数据

Trace	Pkts count	Bytes	Flows count
Trace1(0:00-1:00)	3.70E+8	2.52E+10(24.0G)	1.28E+6
Trace2(5:00-6:00)	1.23E+8	8.39E+9(8.0G)	3.74E+5
Trace3(19:00-20:00)	9.78E+8	5.73E+10(53.36G)	2.42E+6

表 4　　NOC_SET

应用协议标号	应用类别	所含协议举例	流数	比重(%)
1	WWW	HTTP,etc	304572	74.01
2	Bulk	FTP	5483	1.33
3	Mail	IMAP, POP3, SMTP	385	0.09
4	P2P	BitTorrent, eDonkey, Gnutella, XunLei,ppl ive	71186	17.3
5	Service	DNS, NTP	3035	0.74
6	Interactive	SSH, CVS, pcAnywher e,etc	60	0.014
7	Multimedia	RTSP, Real	20	0.0049
8	Voice	SIP, Skype	276	0.067
9	Others	games, attacks,etc	26500	6.44

从图 4，5 可以看出，多分类器模型比传统的单分类器算法有更高的查准率和查全率，由于训练样本的比例的不均衡性，因此不同的网络流量应用有不同的识别率，在样本总体一定的情况下，应用类型样本比例较大者，训练越充分，查准率越高。从表 5 的总体准确率上也可以发现其总体识别效果很好。

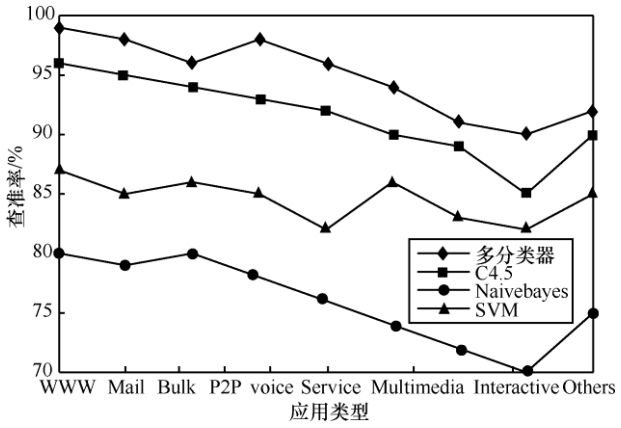


图 4　多分类器与传统机器学习算法查准率

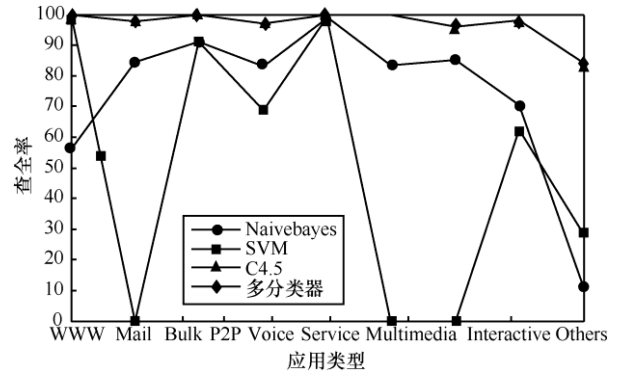


图 5　多分类器与传统机器学习算法查全率

表 5　识别算法对 Overall accuracy 的影响情况

识别算法	Overall accuracy (%)
多分类器算法	97.36
C4.5 算法	96.65
Naivebayes 算法	58.5
SVM	92.5

从表 6 中可以看到随着流数量级的增加时间消耗呈线性增长趋势，且多分类器的时间消耗为最少。结合图 6 中所示的总体正确率，可以看出当训练样本是 10^3 时，多分类器方法训练时间仅需 30 秒就能使识别模型达到 93.8%的准确率，时效度为 35.2。而其他三种方法训练时间都要超过 1 分钟，且时效度较高。而随着训练样本的增加各识别算法的识别正确率趋于平稳。表 7 表明了各算法所产生的内存开销，随着流数量的线性增长，所消耗的内存也在不断的增大，当训练样本为 10^3 时，多分类器所消耗的内存仅为 90MB，而其他三个单分类器算法都要超过 110MB，出现这种情况的原因主要是因为多分类器的训练过程内存开销基本不变且较低，而在分类识别过程中内存开销主要用来存储权

表 6　　不同算法的时间消耗

流数量	多分类器		C4.5		Naïvebayes		SVM	
	时间消耗		时间消耗		时间消耗		时间消耗	
	训练时间 (s)	分类时间 (s)	训练时间 (s)	分类时间 (s)	训练时间 (s)	分类时间 (s)	训练 时间 (s)	分类时间 (s)
10^3	30.0	5.20	84.6	12.58	46.63	13.24	116.6	13.2
10^4	256	7.11	284.7	16.36	275.5	12.44	325.6	19.4
10^5	3503	15.24	3870	23.57	3781	18.96	4294	25.9
10^6	36172	59.2	38476	79.84	37276	67.01	42162	86.7

值和中间计算结果，这部分是随流记录增加而增加的，且所占用的内存较小。

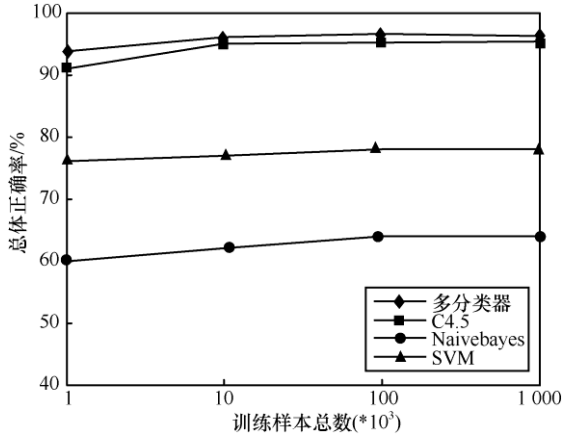


图 6　样本大小对识别正确率的影响

表 7　　不同算法的内存开销

流数量	多分类器	C4.5	Naïvebayes	SVM
	内存开销 (MB)	内存开销 (MB)	内存开销 (MB)	内存开销 (MB)
10^3	90	145	174	116.6
10^4	158	234.7	275.5	224.6
10^5	260	364	388.1	323.4
10^6	298	364.43	392.37	339.62

报文抽样技术目前已经得到广泛的应用，例如 NETFLOW 抽样，它是高速网络流量管理与测量的一项关键技术。因此对于分析报文抽样对分类算法的影响尤为重要，为了分析报文抽样对多分类器融合模型的影响，并与单分类器进行对比，分别通过五种不同的抽样率来观察对总体正确率的影响，结果如图 7 所示。

图 7 中不同抽样比率下采用多分类器模型总体正确率比较平稳，而传统的单分类器算法就出现随抽样比率的增加而出现总体正确率的下降现象。这主要是因为本文采用偏好度来度量分类器的性能，然后根据权值进行优化，这样就弥补了由于样本数据的不平衡性而引起的抽样后的剧烈抖动。

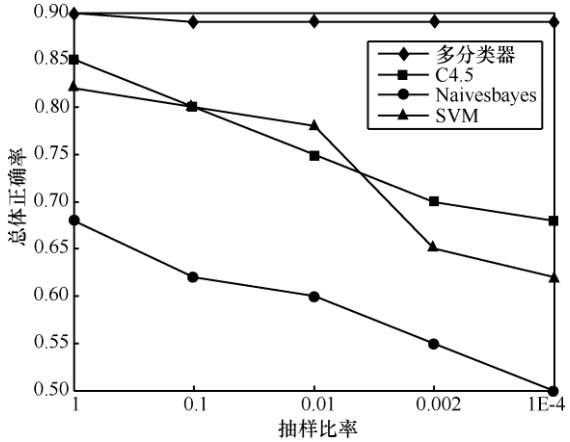


图 7　抽样对识别结果的影响

综上可得，本文基于扩展的 NETFLOW 流记录并将偏好度等定义融入到多分类器模型中，识别结果表明：采用多分类器模型无论是在查准率还是在查全率上都比传统的单分类器算法更高，另外从整体正确率方面也得到了验证。并且从结果上看，采用基于流记录进行流量识别，虽然相关的属性很少，但是识别结果却可以达到和采用全报文的数据集几乎差不多的分类效果，因此这样就为流量的在线分类提供了一种很好的方法。可以在基于 NETFLOW 现有的字段中加入上述的少量的属性特征也可以达到很好的分类效果，而且可以提高在线进行分类的效率。

6　结束语

本文在江苏省网边界以及本地分别获取数据，并利用 L7-filter 对数据进行打标签，得到基准数据集 NOC_SET。提出了采用多分类器组合的方式来解决目前网络流量识别中存在的问题，并与常用的流量识别机器学习算法进行了对比，结果表明了本文

提出的多分类器模型在流量识别中可以使分类的结果更好，识别率更高，且不受因抽样而带来的结果的抖动和不确定性，具有较高的稳定性和鲁棒性。

本文的创新点在于：1) 基于江苏省网边界数据构建了 NOC_SET 标准数据集；2) 提出基于流的几种测度属性；3)提出用于流量识别的多分类器模型。

基于本文的研究，下一步工作主要是：基于本文的流数据以及所提出的流测度属性，为后续研究提供数据支持，也能够通过改善多分类器模型的识别时效性来更好的满足在线的流量识别。

参考文献：

[1] KARAGIANNIS T, PAPAGIANNAKI K, FALOUTSOS M. BLINC: multilevel traffic classification in the dark[A]. Proc of the ACM SIGCOMM[C]. Philadelphia, 2005. 229-240.

[2] ROUGHAN M, SEN S, SPATSCHECK O, et al. Class-of-service mapping for QOS: a statistical signature-based approach to IP traffic classification[A]. Proc of the ACM SIGCOMM Internet Measurement Conf[C]. Taormina, 2004. 135-148.

[3] MOORE AW, ZUEV D. Internet traffic classification using Bayesian analysis techniques[A]. Proc of the 2005 ACM SIGMETRICS Int'1 Conf on Measurement and Modeling of Computer Systems[C]. Banff, 2005. 50-60.

[4] 李君, 张顺颐, 王浩云等. 基于贝叶斯网络的 Peer to peer 识别方法[J]. 应用科学学报, 2009, 27 (2) : 124-130
LI J, ZHANG S Y, WANG H Y, et al. Peer to peer identification using Bayesian networks [J]. Journal of Applied Sciences, 2009, 27 (2): 124-130.

[5] 徐鹏, 刘琼, 林森. 基于支持向量机的 Internet 流量分类研究[J]. 计算机研究与发展, 2009, 46 (3) : 407-414.
XU P, LIU Q, LIN S. Internet traffic classification based on support vector machines [J]. Journal of Computer Research and Development, 2009, 46 (3): 407-414.

[6] LI Z, YUAN R X, GUAN X H. Accurate Classification of the Internet traffic based on the SVM method [A] Proc. Of IEEE International Conference on Communications (ICC) [C].Glasgow, Scotland, United Kingdom, 2007.1373-1378.

[7] MA Y L, QIAN Z J, SHOU G C. Study on preliminary performance of algorithms for network traffic identification [A] Proc Of 2008 International Conference on Computer Science and Software Engineering[C]. Wuhan, China, 2008. 629-633.

[8] ALSHAMMARI,R;ZINCIR-HEYWOOD A N. Investigating two different approaches for encrypted traffic classification [A] Proc Of Sixth Annual Conference on Privacy,Security and Trust (PST) [C]. Fredericton, NB, Canada, 2008. 156-166.

[9] HIRVONEN M, LAULAJAINEN J P. Two-phased network traffic classification method for quality of service management[A] Proc. Of the 13th IEEE International Symposium on Consumer Electronics (ISCE2009) [C]. Kyoto, Japan, 2009. 962-966.

[10] ZUEV D, ANDREW W M. Traffic classification using a statistical approach [A]. Proc of the 6th annual Passive and Active Measurements Workshop (PAM'05) [C]. Boston, USA, 2005. 321-324.

[11] ANDREW W M, DENIS Z. Internet traffic classification using bayesian analysis techniques [A]. Proc of ACM SIGMETRICS'05[C]. Banff, Canada, 2005. 50-60.

[12] JIANG H B, MOORE A W. GE Z H, et al. Lightweight application classification for network management [A]. Proc of the SIGCOMM Workshop on Internet Network Management'07[C]. Kyoto, Japan, 2007.299 - 304.

[13] AULD T, MOORE A W, GULL S F. Bayesian neural networks for Internet traffic classification [J]. IEEE Transactions on Neural Networks, 2007, 18(1): 223-239.

[14] LI W, CANINI M, MOORE A W, et al. Efficient application identification and the temporal and spatial stability of classification schema [J]. Computer Networks, 2009, 53: 790–809.

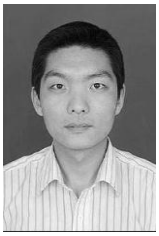
[15] HALL D L. Mathematical techniques in multi sensor Data Fusion.Bosston:Artech Hous.2004:125-13

[16] L7-filter, Application Layer Packet Classifier for Linux.[EB/OL]http://l7-filter.sourceforge.net.2003.

[17] ORPONEN P. Dempster's rule of combination is #P-complete [J].Artificial Intelligence, 1990, 44(1, 2):245-253.

[18] DAVIS J, GOADRICH M. The relationship between Precision-Recall and ROC curves [A]. Proceedings of the 23rd International Conference on Machine learning (ACM, 2006) [C].Pittsburgh, PA, United. 233-240.

作者简介:



董仕（1980-），男，河南周口人，东南大学博士生，主要研究方向为网络测量与网络行为学。

丁伟（1962-），女，江苏南京人，东南大学教授，博士生导师，主要研究方向为网络行为学。