

基于 M3 和 POSS 特征的网络流量分类研究

何继玲, 于威威

(上海海事大学, 上海 201306)

摘要:网络流量分类是网络研究和流量工程的重要基础,网络流量分类大致分为基于端口号、有效负载、主机行为和机器学习等四种分类方法。目前基于机器学习的方法成为了研究热点。在机器学习过程中,特征选择可以实现数据维度约简,从而提高学习模型的泛化能力。针对大规模的流量数据以及网络流量中存在的类别不平衡问题,将最小最大集成策略(min-max module, M3)和多目标演化子集选择算法(Pareto optimization for subset selection, POSS)应用到网络流量分类的特征选择过程中。同时将该方法与其他特征选择方法以及经典的处理类别不平衡问题的方法进行对比。实验结果表明, M3 策略在大部分情况下能获得较好的性能,并能有效处理网络流量中类别不平衡的问题,在流量分类应用中具有一定的实用性。

关键词:网络流量分类;类别不平衡;多目标演化子集选择算法;最小最大模块化

中图分类号: TP301

文献标识码: A

文章编号: 1673-629X(2018)01-0083-06

doi: 10.3969/j.issn.1673-629X.2018.01.018

Research on Network Traffic Classification Based on Min-Max Module and POSS Feature

HE Ji-ling, YU Wei-wei

(Shanghai Maritime University, Shanghai 201306, China)

Abstract: Network traffic classification is an important foundation of network research and traffic engineering. Network traffic classification can be divided into four classification methods like basis on port number, payload, host behavior or machine learning. At present, the machine learning method has become a research hotspot. In the process of machine learning, feature selection can reduce the dimensionality of data and improve the generalization of learning model. In view of the class imbalance of existing large-scale network traffic flow data, min-max module (M3) and Pareto optimization for subset selection (POSS) are applied to feature selection of network traffic classification. It is compared with other feature selection methods and classic methods of dealing with the problem of class imbalance. The experiment shows that the M3 strategy can obtain better performance in most cases and can effectively deal with the problem of class imbalance in network traffic, which has showed its effectiveness in traffic classification.

Key words: network traffic classification; class imbalance; POSS; min-max module

0 引言

随着信息科学技术的不断进步,网络在信息交换中发挥着举足轻重的作用,同时也使网络数据和流量数量呈爆炸性增长。网络流量分类常用于信息识别检测系统中,以自动识别检测各类流量所包含的信息。因此,对网络流量进行正确分类成为了一个热门的研究领域。目前主流的网络流量分类有四种方法^[1],分别是基于端口号的分类、基于主机行为的分类、基于有效负载的分类以及基于机器学习的分类。

文中使用机器学习中的集成特征选择结合支持向量机算法对网络流量进行分类。集成特征选择的过程

包括数据集的划分、特征子集的产生以及特征子集的集成。基于以上步骤,比较了三种集成策略以及三种特征选择方法。除此之外,网络流量具有类别不平衡的特点,为了解决该问题,在特征选择阶段使用基于最小最大模块化的集成特征选择。

1 网络流量分类

1.1 网络流量

网络流量是将网络数据按照[源 IP 地址、源端口号、目的 IP 地址、目的端口号、IP 协议]五元组的格式提取的^[2]。通过五元组,可以将网络流量映射到传输

收稿日期: 2016-12-28

修回日期: 2017-05-03

网络出版时间: 2017-10-19

基金项目: 上海市自然科学基金(14ZR1419300)

作者简介: 何继玲(1991-),女,硕士,研究方向为图像处理与模式识别;于威威,副教授,研究方向为图像处理、模式识别、数据挖掘等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20171019.1559.012.html>

层的 TCP 流和 UDP 流。网络流量分类的目的是将各网络流量对应到应用层中相应的应用类型。

网络流量具有数据量庞大、结构复杂、类别多样和类别不平衡等特点。在目前互联网高速发展的时代,流量数据以惊人的速度增长,导致流量分类的实时性和准确性受到了极大挑战。同时,网络流量类型通常以 HTTP、P2P 等为主,从而导致严重的类别不平衡现象以及新型流量预测等问题。为了解决上述问题,Liu Z 等^[3]提出了 COFS 算法用于解决不平衡问题,Zhang J 等^[4]提出了一种对未知流量预测的方法。

由于 P2P 网络的发展,传统的基于端口号的分类方法只能用于正确地区分拥有常用端口号的流量。而出于安全性的考虑,大部分流量都会对传输层加密,使得基于有效负载或主机行为的分类同样受到限制。为了解决上述方法存在的问题,基于机器学习的方法成为了目前的研究热点。

1.2 基于机器学习的网络流量分类方法

基于统计理论和机器学习的方法使用的流量特征主要是流量数据的统计量,比如包的长度、包到达时间等。该方法首先将网络数据进行统计理论分析,筛选出原始的统计特征作为初始特征集;再进行特征选择筛选出重要特征;最后使用分类或聚类等方法得到最终预测结果。即根据流集合、属性集和类别集,将所有流量映射到对应的类别中。

Moore A 等^[5]提取出 249 维原始特征,使用机器学习的方法对流量进行分类,文中使用朴素贝叶斯作为分类算法;随后 Jeffery Erman 等^[6]使用聚类方法(K-Means, DBSCAN 和 EM)对网络流量进行分类;Wang Y 等^[7]使用 AdaBoost 集成分类方法提高了分类准确率。除了分类算法,研究者在特征选择算法中也进行了改进。Zhou W 等^[8-9]使用 FCBF 特征选择算法结合神经网络对流量进行分类;Kaur J 等^[10]使用了 CFS 和 CON 特征选择算法结合 C4.5、多重感知机、贝叶斯网络等分类算法进行分类;而文中则使用网络流量分类方法对流量进行分类。

2 设计方案

为了得到更好的特征子集并解决类别不平衡问题以提高分类性能,在机器学习过程中的特征选择阶段采用基于最小最大模块化集成特征选择对网络流量进行分类。特征选择算法分别使用 Fisher、Relief^[11]和 POSS^[12],分类算法选择支持向量机。

2.1 集成特征选择

集成特征选择包括三个阶段^[13]:第一,数据集的划分,即由原始数据集通过一定的方法产生多个数据子集;第二,特征选择,即对每一个数据子集使用特征

选择算法,从而产生多个特征选择结果;第三,结果的集成,即对所有数据子集产生的特征子集使用某种方法进行集成,得到最终的特征子集。

2.1.1 集成策略

文中对比了最小最大策略、投票法^[14]、K-中心点法^[15]三种集成策略,本节主要介绍投票法和 K-中心点法,最小最大策略将在下文详细介绍。

投票法是将特征选择器的结果先转化为特征子集,然后统计每个特征被选中的频率,将出现频率最高的 M 个特征作为最终的集成输出:

$$s_p = (s_{p,1}, s_{p,2}, \dots, s_{p,d}) \quad (1)$$

$$s_{p,i} \in \{0, 1\}, i = 1, 2, \dots, d \quad (2)$$

$$S = \sum_{p=1}^B s_p \quad (3)$$

其中, s_p 第 p 个基特征选择器上的特征子集输出; S 为每个特征被选择的频率,最终输出频率最高的前 M 个特征作为特征子集。

K 中心点聚类集成主要是利用聚类的思想,从多个基特征选择器中选择出具有代表性的结果作为最终输出。文中 K 取 1,即选择出一个中心点作为最终输出:

$$W_{\text{mediod}} = 1_Mediod(\{w_1, w_2, \dots, w_B\}) \quad (4)$$

2.1.2 特征选择方法

使用 POSS 特征选择与其他两种常用的特征选择算法(Relief、Fisher)进行比较。

Fisher 特征选择算法应用广泛。它将每个特征的费希尔值作为权重,主要思想是对每一个特征在不同样本上的均值相差越大,方差之和越小,则被赋予的权重越大。

Relief 特征选择算法是根据样本在不同特征上的假设间隔为它们赋予不同的权重。其假设间隔是指保持样本分布不变的情况下决策面所能移动的最大距离。对于每个特征而言,样本的假设间隔越大,则该特征被赋予的权重越大。

POSS 特征选择算法是基于帕雷托优化子集的方法以及改进,从而获得相对最优的特征子集^[12]。大致思想为:假设给定数据集 $V = \{X_1, X_2, \dots, X_n\}$ (n 表示特征数量),定义判别函数 f 以及正整数 k 。为了寻找一个特征子集 $S \subseteq V$,并在约束条件 $|S| \leq k$ 的情况下,判别函数能够取得最优值。其中 $|\cdot|$ 表示数据集的大小。其定义如式(5)所示:

$$\operatorname{argmin}_{S \subseteq V} f(S), \text{ s. t. } |S| \leq k \quad (5)$$

该式可看作两个优化目标:优化判别函数 f ,即 $\operatorname{argmin}_{S \subseteq V} f(S)$;同时保证 $|S|$ 最小,即 $\min_{S \subseteq V} \max\{|S| - k, 0\}$ 。但一般情况下,想要取得更好的判别函数值,即 $f(S)$ 需要更大的样本数量,所以可以说这两个优化目

标在一定程度上是矛盾的。然而,POSS 方法可以有效地实现上述两个优化目标。其具体步骤描述如下:

在该算法中,使用 n 维二元向量 $s = \{0,1\}^n$ 来表示某一个特征是否被选中,称 s 为特征选择的一种解。即当 $s_i = 1$,表示第 i 个特征将会被选到特征子集 S 中;当 $s_i = 0$,表示第 i 个特征不出现在 S 中。接着,为所有可能的解 s 定义两个属性 o_1, o_2 。前者表示评价标准值,后者表示稀疏度,具体如下所示:

$$s. o_1 = \begin{cases} +\infty, & s = \{0\} \text{ or } |s| \geq 2k \\ f(s), & \text{otherwise} \end{cases} \quad (6)$$

$$s. o_2 = |s| \quad (7)$$

当 o_1 的值趋向于 $+\infty$ 时,说明该方案效果较差,可予以舍弃。然后引入一个隔离函数 $I: \{0,1\}^n \rightarrow R^{[16]}$ 。该函数确定是否允许两个解比较:只有当它们具有相同的隔离函数值时,它们才是可比较的。若隔离函数值相等,对于方案 s' 和方案 s ,若 $s'. o_1 \leq s. o_1$,则说明 s' 弱支配 s 。若 $s'. o_2 \leq s. o_2$ 且 $s'. o_1 < s. o_1$,则 s' 支配 s 。但是如果 s 既不支配 s' , s' 也不支配 s ,则说明它们之间不可比较。具体算法如下:

算法 1:POSS 方法。

输入: $V = \{X_1, X_2, \dots, X_n\}$, 判别函数 f 和正整数 k

参数:迭代次数 T 和隔离函数 $I: \{0,1\}^n \rightarrow R$

算法过程:

$s = \{0\}^n$ 和 $P = \{s\}$

$t = 0$

While $t < T$ do

均匀随机地从 P 中抽取 s

对 s 中的每一位以 $1/n$ 的概率翻转(若当前为 0,则变成 1;反之也成立),产生 s'

if $\forall z \notin P$, 若 $I(z) = I(s')$ 并且 $((z. o_1 < s'. o_1 \wedge z. o_2 < s'. o_2) \text{ or } (z. o_1 \leq s'. o_1 \wedge z. o_2 \leq s'. o_2))$

则 $Q = \{z \in P \mid I(z) = I(s') \wedge s'. o_1 \leq z. o_1 \wedge s'. o_2 \leq z. o_2\}$ $P = (P \setminus Q) \cup \{s'\}$

end if

$t = t + 1$

endwhile

输出: $\underset{s \in P, |s| \leq k}{\operatorname{argmin}} f(s), s. t. |S| \leq k$

2.2 基于最小最大规则(M3)的集成特征选择方法

为了提高分类算法对大规模数据的处理能力, Lu 等提出了最小最大模块化的集成^[17],并结合各个分类器用以解决文本分类、专利分类、人脸识别等领域的问题。将该方法与特征选择结合应用到网络流量分类中,既能得到更好的特征子集又能解决类别不平衡问题,从而达到更好的分类效果。

根据集成特征选择的步骤,基于 M3 的集成特征

选择同样分为三个步骤:首先,将原始数据根据它们的类别信息划分成多个相对较小的平衡数据子集;然后,在每个数据子集上进行特征选择,得到不同的特征选择结果;最后,对多个特征选择结果采用最小最大规则进行集成。

2.2.1 任务分解

在任务分解阶段,对于一个 K 类的分解问题,首先采用“一对一”的策略将其分解为 $K(K-1)$ 个二分类问题^[1]。假设 K 类的训练数据集表示为:

$$X = \{(x_l^i, y^i)\}_{l=1}^{L_i}, i = 1, 2, \dots, K \quad (8)$$

其中, L_i 为第 i 类样本的个数; x_l^i 为第 i 类样本的第 l 个样本; y^i 为第 i 类样本的标签

通过“一对一”的策略,第 i 类样本和第 j 类样本的训练数据集可以表示为:

$$X_{i,j} = \{ \{(x_l^i, y^i)\}_{l=1}^{L_i} \cup \{(x_l^j, y^j)\}_{l=1}^{L_j} \}, i = 1, 2, \dots, k, i \neq j \quad (9)$$

如果两类问题的规模较大或者是不平衡性,可以进一步将它们划分成规模更小的较为平衡的子问题^[18]。由于要求每一个数据子集中样本个数几乎一致,使得每个组合的二类问题中的类别都是平衡的,故 M3 可以有效地解决类别不平衡的问题,在网络流量分类中具有一定的实用性。

2.2.2 集成结果

根据上述划分策略,得到 $N^+ \times N^-$ 个不同类别之间相对平衡、规模较小的样本子集,然后在每个样本子集上使用某种特征选择算法,从而得到 $N^+ \times N^-$ 个特征选择结果。对这些结果使用 MIN-MAX 规则。

MIN 规则:对拥有相同正类训练样本集和不同负类训练样本集的分类结果取最小值;

MAX 规则:对拥有相同负类训练样本集和不同正类训练样本集的分类结果取最大值。

具体过程如下:

$$W^{i,j} = F(T^{i,j}), i = 1, 2, \dots, N^+, j = 1, 2, \dots, N^- \quad (10)$$

其中, $F(\cdot)$ 表示特征选择器; $W^{i,j}$ 表示在样本子集 $T^{i,j}$ 上进行特征选择的结果。

$$W^i = \operatorname{Min}(W^{i,1}, W^{i,2}, \dots, W^{i,N^-}), i = 1, 2, \dots, N^+ \quad (11)$$

$$W = \operatorname{Max}(W^1, \dots, W^{N^+}) \quad (12)$$

其中, W^i 表示对包含相同正类样本、不同负类样本的数据子集的特征选择结果采用 MIN 规则,MIN 规则是取每个特征在不同特征选择结果中的最小权重作为输出; W 表示对上一步输出的 N^+ 个特征选择结果使用 MAX 规则,MAX 规则是取每个特征在不同特征选择结果中的最大权重作为输出。

最小最大策略的集成特征选择流程如图 1 所示。

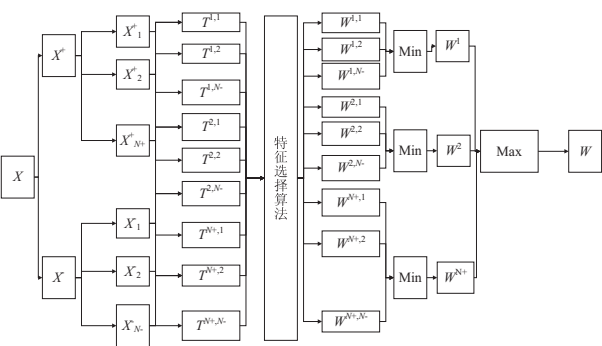


图 1 最小最大策略的集成特征选择模型

该算法描述如下：

算法 2：基于最小最大规则的特征选择集成算法。

输入：训练集 x ，正类样本划分个数 N^+ ，数据划分

方法 P ，特征选择算法 F

输出：特征权重 W

(1) 数据划分阶段。

统计 x 中正类和负类样本个数 l^+, l^-

计算负类样本划分的块数

$$N^- = \lfloor l^- / l_{\text{arg}}^+, l_{\text{arg}}^+ = l^+ / N^+ \rfloor$$

将 X 按正类、负类分为 x^+, x^-

$$\{X_1^+, X_2^+, \dots, X_{N^+}^+ = p(X^+)\}$$

$$\{X_1^-, X_2^-, \dots, X_{N^-}^- = p(X^-)\}$$

for $i = 1:N^+$

for $j = 1:N^-$

$$T^{i,j} = \{X_i^+\} \cup \{X_j^-\}$$

end for

end for

(2) 最小最大集成阶段。

for $i = 1:N^+$

for $j = 1:N^-$

$$W^{i,j} = F(T^{i,j})$$

end for

$$W^i = \text{Min}(W^{i,1}, W^{i,2}, \dots, W^{i,N^-})$$

end for

$$W = \text{Max}(W^1, W^2, \dots, W^{N^+})$$

3 实验

3.1 数据集

通过实验验证基于最小最大模块化集成特征选择筛选出的特征在网络流量分类中的分类性能，及其处理类别不平衡的能力。主要对比了三种集成策略、三种特征选择算法以及两种处理类别不平衡的方法。

实验数据集^[5]包含 20 000 条样本，在原始数据集内进行训练集和测试集以及验证集的划分。Moore A 等在文献[5]中提取了 249 个初始特征，这些特征大致可分为单向流特征和双向流特征，单向流为客户端

或服务端单方发送的数据包的统计，双向流则为两个方向均发送的数据包的统计。

根据需要，文中选取 29 个关键特征，其数据集中共有六种流量类型，每类流量的个数，所占比例，在实验中的类别标签以及使用 M3 时每个类别分块的个数如表 1 所示。

表 1 流量类型及描述

类型名称	样本个数	所占比例	类型标签	M3 中分块个数
DATABASE	337	0.016 85	1	2
SERVICE	489	0.024 45	2	2
P2P	802	0.040 1	3	4
FTP	2 635	0.131 75	4	16
MAIL	6 037	0.301 85	5	36
HTTP	9 700	0.485	6	60

为了去除各个属性定义域的区别对分类效果的影响，采用 0-1 归一化的方法将数据缩放到 0 到 1 的闭区间内；原始数据集中包含部分缺失值，采用拉格朗日插值法填补缺失值。

采用的分类器算法是支持向量机。实验采用 10 折交叉验证法计算分类准确率。在支持向量机中，采用高斯核函数，其 sigmoid 值设置为 2，损失函数 C 设置为 32，采用 SMO 算法来计算其参数。

对于类别不平衡的问题，常用的评价准则有 ROC 曲线、F-Measure^[19]、G-Mean^[20]等，文中使用 F-Measure 度量方法^[21]。该方法能考虑到每个类别上的分类准确率，并将每个类别的准确率同等对待，解决了忽略不平衡数据集少样本重要性的缺陷。F-Measure 的计算方法为：

$$\text{F-Measure} = 2 / (\frac{1}{\text{precision}} + \frac{1}{\text{recall}}) \tag{13}$$

$$\text{precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{14}$$

$$\text{recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{15}$$

其中，recall 表示正类样本的分类准确率；precision 表示负类样本的分类准确率；TP 表示被正确分类的正类样本数量；FN 表示被错误分类的正类样本数量；TN 表示被正确分类的负类样本数量；FP 表示被错误分类的负类样本数量。

3.2 实验结果分析

在预处理之后，先对原始数据集使用超平面划分正负类样本。对于多分类问题，采用“一对一”的策略将其转换成二分类问题进行划分。后续步骤仅考虑二分类情况。

对每一块数据子集使用同样的特征选择算法。分别采用了 POSS、Fisher 和 Relief 算法，得到特征排序；根据特征排序，对每一个特征子集分别使用投票法、 K

-中心点法以及最小最大策略三种方法各自的特征子集。将权重排名前 6,8,10,12,14,16,18,20 个特征的数据子集进行训练与测试;将得到的特征子集使用支持向量机预测得到分类结果,计算其 F-Measure 值。

3.2.1 特征选择算法的结果对比

该实验首先使用超平面划分 (HP) 和三种集成方法。比较结果如图 2~4 所示。

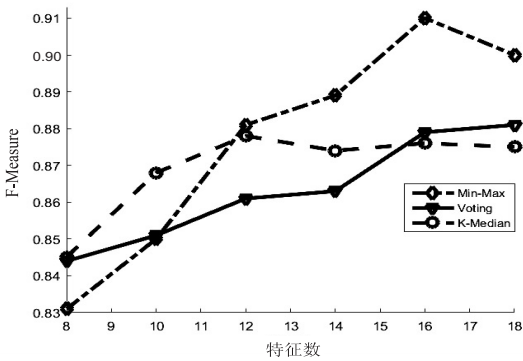


图 2 POSS 在不同的集成策略算法的结果比较

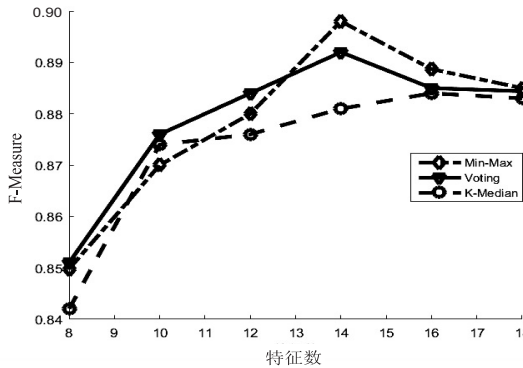


图 3 Fisher 在不同的集成策略算法结果的比较

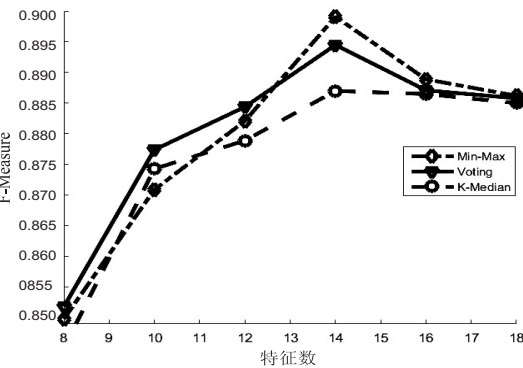


图 4 Relief 在不同的集成策略算法的结果比较

可以看出在集成策略方面,随着特征数量的变化,三种方法均有较好的表现。从整体来看,基于 M3 的集成策略效果优于其他两种方法。并且在选择 15 到 18 维特征时分类效果较好,在特征数到 20 维时,结果趋于稳定或下滑。

如图 5 所示,在大部分情况下,当选取 18 个特征时,三种特征选择算法的性能可以达到最好。从不同

特征选择对比的情况来看,POSS 特征选择算法在大多数情况下的性能都优于其他算法。

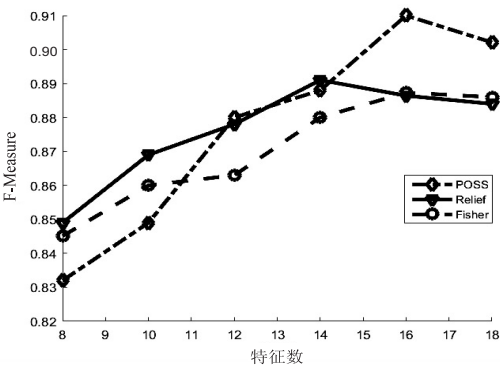


图 5 三种不同特征选择算法的结果比较

3.2.2 不同的采样策略比较

为了比较处理数据不平衡的能力,使用经典的处理不平衡数据的方法-1:1 随机欠采样和 1:1.5 随机欠采样策略^[22]选取数据子集,并使用 POSS 算法和支持向量机算法与基于 M3 的策略进行比较。

图 6 比较了三种策略处理不平衡数据的能力,在特征选择阶段均选择权重排名前 16 的特征作为训练样本。结果表明,1:1.5 的采样虽然能保持比 1:1 更好的完整性,但效果略次于 1:1,而基于 M3 可同时处理大规模数据和不平衡数据,既能保持完整性,也能保持相对较高的准确性,在处理不平衡数据时效果最佳。

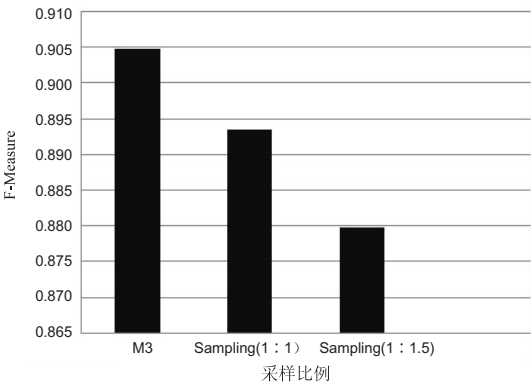


图 6 不同采样方法的比较

4 结束语

文中将集成方法与 POSS 特征选择算法相结合的集成特征选择方法应用到网络流量分类领域,比较了三种不同集成方法以及三种不同的特征选择算法的分类效果,同时比较了 M3 处理类别不平衡的能力,并且将用于解决二分类的 M3 方法拓展到了解决多分类问题,从而对流量进行分类。实验结果表明,基于 M3 的划分和 POSS 方法在大部分情况下优于其他划分与集成的策略和方法,在处理不平衡数据时同样优于其他两种方法。在流量分类领域中能取得良好的效果,能

有效地对其进行分类。

参考文献:

- [1] NAMDEV N, AGRAWAL S, SILKARI S. Recent advancement in machine learning based internet traffic classification [J]. *Procedia Computer Science*, 2015, 60: 784–791.
- [2] KATRIS C, DASKALAKI S. Comparing forecasting approaches for Internet traffic [J]. *Expert Systems with Applications*, 2015, 42(21): 8172–8183.
- [3] LIU Z, WANG R, TAO M, et al. A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion [J]. *Neuro-computing*, 2015, 168: 365–381.
- [4] ZHANG J, CHEN C, XIANG Y, et al. An effective network traffic classification method with unknown flow detection [J]. *IEEE Transactions on Network and Service Management*, 2013, 10(2): 133–147.
- [5] MOORE A, ZUEV D, CROGAN M. Discriminators for use in flow-based classification [M]. [s. l.]: [s. n.], 2005.
- [6] ERMAN J, ARLITT M, MAHANTI A. Traffic classification using clustering algorithms [C]//*Proceedings of the 2006 SIGCOMM workshop on mining network data*. [s. l.]: ACM, 2006: 281–286.
- [7] WANG Y, XIANG Y, YU S. Internet traffic classification using machine learning: a token-based approach [C]//*IEEE 14th international conference on computational science and engineering*. [s. l.]: IEEE, 2011: 285–289.
- [8] ZHOU W, DONG L, BIC L, et al. Internet traffic classification using feed-forward neural network [C]//*International conference on computational problem-solving*. [s. l.]: IEEE, 2011: 641–646.
- [9] HU L T, ZHANG L J. Real-time internet traffic identification based on decision tree [C]//*World automation congress*. [s. l.]: IEEE, 2012: 1–3.
- [10] KAUR J, AGRAWAL S, SOHI B S. Internet traffic classification for educational institutions using machine learning [J]. *International Journal of Intelligent Systems and Applications*, 2012, 4(8): 37–45.
- [11] ROBNIK-SIKONJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF [J]. *Machine Learning*, 2003, 53(1–2): 23–69.
- [12] YU Y, YAO X, ZHOU Z H. On the approximation ability of evolutionary optimization with application to minimum set cover [C]//*Proceedings of the twenty-third international joint conference on artificial intelligence*. Beijing: AAAI Press, 2012: 3190–3194.
- [13] AWADA W, KHOSHGOFTAAR T M, DITTMAN D, et al. A review of the stability of feature selection techniques for bioinformatics data [C]//*13th international conference on information reuse and integration*. [s. l.]: IEEE, 2012: 356–363.
- [14] LI Y, GAO S Y, CHEN S. Ensemble feature weighting based on local learning and diversity [C]//*Proceedings of the twenty-sixth AAAI conference on artificial intelligence*. Toronto, Canada: AAAI Press, 2012.
- [15] WOZNICA A, NGUYEN P, KALOUSIS A. Model mining for robust feature selection [C]//*Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. [s. l.]: ACM, 2012: 913–921.
- [16] ZHAO Z, GUO S, XU Q, et al. G-means: a clustering algorithm for intrusion detection [C]//*Proceedings of the 15th international conference on advances in neuro-information processing – volume part I*. Auckland, New Zealand: Springer-Verlag, 2009: 563–570.
- [17] LU B L, ITO M. Task decomposition and module combination based on class relations: a modular neural network for pattern classification [J]. *IEEE Transactions on Neural Networks*, 1999, 10(5): 1244–1256.
- [18] CHU X L, MA C, LI J, et al. Large-scale patent classification with min-max modular support vector machines [C]//*International joint conference on neural networks*. [s. l.]: IEEE, 2008: 3973–3980.
- [19] HUANG Y J, POWERS R, MONTELLIONE G T. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics [J]. *Journal of the American Chemical Society*, 2005, 127(6): 1665–1674.
- [20] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection [C]//*Proceedings of the fourteenth international conference on machine learning*. Stanford, USA: ICML, 2000: 179–186.
- [21] LI Y, FENG L L. Integrating feature selection and min-max modular SVM for powerful ensemble [C]//*International joint conference on neural networks*. [s. l.]: IEEE, 2012: 1–8.
- [22] SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J, et al. RUSBoost: a hybrid approach to alleviating class imbalance [J]. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 2010, 40(1): 185–197.