基于选择性集成策略的嵌入式网络流特征选择

潘吴斌[1),3)] 程 光[1),3)] 郭晓军[1),2),3)] 王 艳[1),3)]

[1)](东南大学计算机科学与工程学院 南京 210096)

[2)](西藏民族学院信息工程学院 陕西 咸阳 712082)

[3)](计算机网络和信息集成教育部重点实验室(东南大学) 南京 210096)

摘 要 机器学习在网络流量分类中存在特征选择度量指标单一、类别不平衡和概念漂移等问题,使得模型复杂度提高、泛化能力下降.该文提出基于选择性集成策略的嵌入式特征选择方法,根据选择性集成策略选取部分特征选择器集成,再改进序列前向搜索和封装器组合方法二次搜索最优特征子集.实验结果表明该算法在保证分类效果的同时有效降低了特征子集复杂度,从而达到了分类效果、效率和稳定性的最优平衡.

关键词 选择性集成;特征选择;嵌入式;稳定性

中图法分类号 TP393 **DOI 号** 10.3724/SP.J.1016.2014.02128

An Embedded Feature Selection Using Selective Ensemble for Network Traffic

PAN Wu-Bin[1),3)] CHENG Guang[1),3)] GUO Xiao-Jun[1),2),3)] WANG Yan[1),3)]

[1)](*School of Computer Science and Engineering, Southeast University, Nanjing* 210096)

[2)](*School of Information Engineering, Tibet Nationalities Institute, Xianyang, Shaanxi* 712082)

[3)](*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing* 210096)

Abstract The problems of feature selection metrics single, class imbalance and concept drift exist in machine learning for network traffic classification, leading the model complexity increased, the generalization ability decreased. Therefore, an embedded feature selection method based on selective ensemble is proposed, according to the selective ensemble strategy to ensemble part of feature selectors, and then through the combination method of improved sequence forward search and wrapper secondary search optimal feature subset. Experimental results show that the proposed algorithm can reduce the complexity of feature subset effectively while ensuring the classification performance, so as to achieve the optimal balance of the classification performance, efficiency and stability.

Keywords selective ensemble; feature selection; embedded; stability

1 引 言

流量分类技术在网络测量与安全领域应用广泛,一方面,根据应用实时性要求优化网络通信资源;另一方面,实时流量分类提前识别并阻止异常流量.当前,基于统计特征的流量分类方法是最常用的,采集流量的外部特征属性通过机器学习方法进行分类[1-2],尽管该方法可以克服基于端口和深度包检测方法的不足,但特征属性中包含的冗余和不相

收稿日期:2014-04-28;最终修改稿收到日期:2014-07-12.本课题得到江苏省未来网络前瞻性研究项目(BY2013095-5-03)、江苏省科技支撑计划(工业)项目(BE2011173)及江苏省"六大人才高峰"高层次人才项目(2011-DZ024)资助.**潘吴斌**,男,1987 年生,博士研究生,主要研究方向为网络安全、网络测量及流量分类.E-mail:wbpan@njnet.edu.cn.**程 光**,男,1973 年生,博士,教授,博士生导师,主要研究领域为网络安全、网络测量与行为学及未来网络安全.**郭晓军**,男,1983 年生,博士研究生,主要研究方向为网络安全、网络测量及网络管理.**王 艳**,女,1991 年生,硕士研究生,主要研究方向为网络安全、网络测量及流量分类.

关特征会增加模型复杂度、降低模型可信度，导致分类效果和效率同时下降. 然而，特征选择方法可以有效地消除冗余和不相关特征，选取最优特征子集. 当前，借助特征选择方法还存在一定的局限性：(1) 概念漂移使得特征选择结果很难保持稳定，特征属性及其数目随之改变；(2) 不同的特征选择方法缺少统一的评价指标. 当前特征子集的好坏主要由分类性能来评价，而各个特征子集的分类性能不稳定，有时会出现极个别分类性能较低的现象；(3) 有些机器学习算法获得的分类准确率也不稳定，这与机器学习算法数据预处理过程有很大关系，比如 C4.5 决策树会预先离散化数据.

本文受选择性集成思想[3]和 Embedded 方法[4]启发，提出基于选择性集成策略的嵌入式特征选择方法，采用选择性集成方法选取部分特征选择器集成，再通过改进序列前向搜索和封装器组合方法进一步搜索特征子集，该方法可以获得特征较少且稳定的最优特征子集. 另外，采用离散化方法分割连续型数据与类分布一致，简化数据，减少噪声数据，提高机器学习算法分类效果和效率.

本文研究贡献主要在于以下几点：

(1) 本文提出基于选择性集成策略的嵌入式特征选择方法. 一方面，单个特征选择方法很难获得稳定的最优特征子集，选择性集成策略综合多个特征选择器的优点，获得比集成全部更好的性能. 另一方面，采用穷举式搜索耗费时间长，将集成后的特征采用启发式搜索进行二次特征选择，快速剔除不相关特征. 最后，将准确率最高的特征子集作为全局特征子集，有效提高特征子集的稳定性. 该方法有效消除不相关和冗余特征，获得稳定的最优特征子集，有利于提高模型泛化能力、分类效果和效率.

(2) 本文引入稳定性度量评估特征子集. 当前特征选择方法采用单一度量标准(如相关性、一致性、分类精度)选取特征，且没有度量指标评估特征子集，文中将特征加权出现频率作为稳定性度量标准，从整体上评估特征子集，有效应对分类过程中概念漂移问题.

(3) 本文采用最小描述长度原理的离散化方法解决机器学习方法分类准确率不稳定性的问题. 有些机器学习方法分类连续型数据的效果和效率低下，离散化方法将连续属性转化为有限的区间，达到分类准确率与离散区间的最优平衡. 根据信息论原理将连续属性分割成多个离散区间，以最小描述长度为控制离散化算法的结束准则. 该方法有利于简化数据，减少数据中的噪声，提高机器学习方法的分类准确率.

本文第 2 节综述网络流量分类中特征选择研究的现状；第 3 节描述基于选择性集成策略的嵌入式特征选择方法；第 4 节引入两项评价指标，包括性能(平均准确率、查准率、查全率和综合评价 *F-Measure*)和稳定性；第 5 节给出实验数据集、简要说明实验环境和流程；第 6 节从准确性和效率性进行实验分析，并提出离散化方法解决机器学习分类准确率不稳定性问题；第 7 节总结全文并展望未来的工作.

2　相关研究

目前，基于统计特征的机器学习流量分类方法研究广泛[5-7]，但很少有研究人员关注分类过程中存在的类别不平衡性和概念漂移问题[8]. 一般情况下，类别不平衡性主要通过重取样、特征选择和改进分类算法等方法来解决，但当特征维数较高时，重取样和改进算法作用不明显[9]. 所谓概念漂移就是类别分布随时间发生变化，使得分类模型很难保持较高的分类准确率. 这些影响使得基于统计特征的机器学习方法很难获得较高的分类效果和效率. 特征选择方法可以很好的解决维数灾难问题，但现有的单个特征选择方法[10]未考虑特征属性与应用间的内在关联，仅从单一度量指标评估特征，泛化能力和稳定性不高，存在一定的局限性.

Li 等人[11]考虑到不同时间域和空间域对流量分类效果的影响，采用 FCBF 和对称不确定性度量选择特征子集，由于单个 FCBF 特征选择方法对于多个数据集很难保持较高的分类性能，没有很好的解决概念漂移问题. 张宏莉等人[12]提出了基于 Bagging 集成学习的分类方法，虽然 Bagging 集成学习可以提高总体分类精度，但会使实例较少的类别分类准确性降低，类不平衡问题仍然存在. de Souza 等人[13]提出一种动态 Adaboost 集成学习算法，并对 IP 分组来简化分类模型，但分类模型中的 IP 群组不能用于其他网络环境，概念漂移问题仍然存在. Zhang 等人[8]提出了一种采用加权对称不确定性和 ROC 曲线下面积度量的混合特征选择算法，不需要改变类别分布就能提高少数类的查全率和查准率、以及分类的字节准确率，有效解决类别不平衡性，但没有解决动态数据流引起的概念漂移. Fahad 等人[14]提出一种多种特征选择方法集成的混合式特征选择方法，该方法有利于简化分类模型，减少模型建立和分类时间，但是该方法耗费时间长，且没有考

虑类不平衡性和概念漂移问题.

与前述工作相比,本文针对流量分类中类不平衡和概念漂移问题,提出基于选择性集成策略的嵌入式特征选择方法(embedded Feature Selection using Selective ENsemble,FSEN).FSEN 采用选择性集成思想结合部分特征选择器的优点,通过启发式搜索快速剔除不相关和冗余特征,再采用离散化方法将连续型数据优化为离散型数据,简化数据,减少数据噪声,有效提高分类效率和效果.另外,由于特征子集缺乏统一的评价标准,以及特征子集质量对分类的重要性,本文引入一种稳定性度量评价概念漂移引起的特征子集动态变化现象.

3 基于选择性集成的嵌入式特征选择

FSEN 算法主要包括两部分:第 1 部分,将多个特征选择器选取的特征子集根据评价指标进行排序,再根据选择性集成策略选择部分特征选择器,从已有的特征选择器中将作用不大和性能不好的特征选择器剔除,将保留的特征选择器集成;第 2 部分,采用朴素贝叶斯算法评估序列前向搜索产生的特征子集,以分类准确率下降为结束准则,再比较多个数据集的最优特征子集选出全局特征子集,提高特征子集稳定性,FSEN 流程如图 1 所示.该方法不仅可以有效剔除不相关和冗余特征,还能提高特征选择的稳定性,使分类准确率与稳定性达到最优平衡.

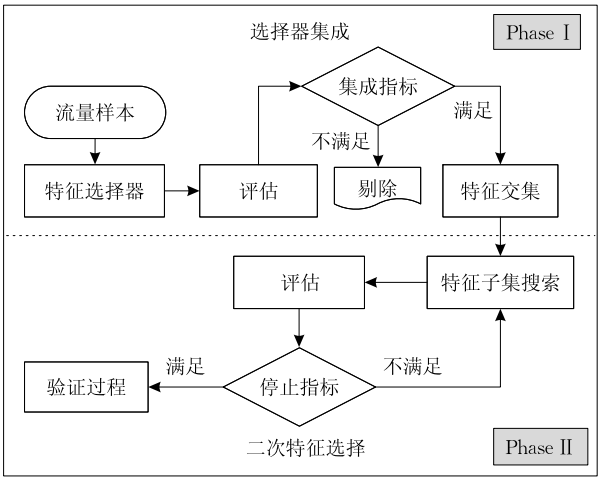


图 1 FSEN 流程图

3.1 选择器集成

选择器集成过程中采用选择性集成(selective ensemble)策略,简单地说,选择性集成就是对同一问题的多种方法进行适当的选择,将所选择的结果进行结合获得比集成全部方法更好的效果[3].特征选择过程中的选择性集成就是从一组特征选择器中选择部分集成,假定在 m 个特征属性上的期望输出 $D=[d_1,d_2,\cdots,d_m]$,其中 d_j 表示第 j 个属性的期望输出,$d_j\in\{-1,1\}(j=1,2,\cdots,m)$.令 $\boldsymbol{f}_i$ 表示第 i 个特征选择器的实际输出,$\boldsymbol{f}_i=[f_{i1},f_{i2},\cdots,f_{im}]^{\mathrm{T}}$,其中 f_{ij} 表示第 i 个特征选择器在第 j 个属性上的实际输出,$f_{ij}\in\{-1,1\}(i=1,2,\cdots,N;j=1,2,\cdots,m)$.当第 i 个特征选择器在第 j 个属性上的实际输出正确时,$f_{ij}d_j=1$,否则 $f_{ij}d_j=-1$.这样,第 i 个特征选择器在这 m 个属性上的泛化误差为

$$E_i=\frac{1}{m}\sum_{j=1}^{m}Error(f_{ij}d_j) \tag{1}$$

$Error(x)$ 定义为

$$Error(x)=\begin{cases}1, & \text{if } x=-1\\ 0.5, & \text{if } x=0\\ 0, & \text{if } x=1\end{cases} \tag{2}$$

和向量 $\boldsymbol{Sum}_j$ 代表所有个体特征选择器在第 j 个属性上的实际输出的和,即

$$\boldsymbol{Sum}_j=\sum_{i=1}^{N}f_{ij} \tag{3}$$

则集成在第 j 个属性上的输出为

$$\hat{f}_j=Sgn(\boldsymbol{Sum}_j) \tag{4}$$

$Sgn(x)$ 定义为

$$Sgn(x)=\begin{cases}1, & \text{if } x>0\\ 0, & \text{if } x=0\\ -1, & \text{if } x<0\end{cases} \tag{5}$$

因此,集成的泛化误差为

$$\hat{E}=\frac{1}{m}\sum_{j=1}^{m}Error(\hat{f}_jd_j) \tag{6}$$

假设集成中剔除第 k 个特征选择器,则新集成在第 j 个示例上的输出为

$$\hat{f}'_j=Sgn(\boldsymbol{Sum}_j-f_{kj}) \tag{7}$$

新集成的泛化误差为

$$\hat{E}'=\frac{1}{m}\sum_{j=1}^{m}Error(\hat{f}'_jd_j) \tag{8}$$

从式(6)和式(8)可知,如果 $\hat{E}$ 不小于 $\hat{E}'$,说明剔除后的集成比原来的集成更好,即

$$\sum_{j=1}^{m}\{Error(Sgn(\boldsymbol{Sum}_j)d_j)-Error(Sgn(\boldsymbol{Sum}_j-f_{kj})d_j)\}\geq 0 \tag{9}$$

当 $|\boldsymbol{Sum}_j|>1$ 时,剔除掉第 k 个特征选择器不影响 d_j,又由于函数 $Error(x)$ 和 $Sgn(x)$ 的性质:

$$Error(Sgn(x)-Error(Sgn(x-y))=-\frac{1}{2}Sgn(x+y) \tag{10}$$

可得

$$\sum_{\substack{j=1\\ j\in\{j\,\|\,\boldsymbol{Sum}_j|\leq 1\}}}^{m} Sgn((\boldsymbol{Sum}_j+f_{kj})d_j)\leq 0 \tag{11}$$

由于 $f_{kj}d_j=-1$，式(11)满足结果.

理论分析表明集成部分特征选择器优于集成所有特征选择器. 特征选择过程中通过对特征选择器排序来选择性集成部分特征选择器，首先，根据准确率评估准则对特征选择器排序，然后，根据指定的特征选择器个数部分选取.

3.2　启发式搜索

FSEN 算法获取的最优特征子集机械式组合容易引起冗余，无法获得较优的特征子集. 假设特征集中有 n 个特征，那么存在 2^n-1 个非空特征子集，搜索策略就是从 2^n-1 个候选特征子集中寻找最优特征子集. 因此，本文改进序列前向搜索算法进一步精选特征子集，每次从未选入的特征中选择一个特征，使它与已选入的特征组合在一起时判据值 J 最大，直到判据值 J 降低为结束准则.

设特征集 $F=\{f_1,f_2,\cdots,f_n\}$，初始时，特征子集 $F_0=\varnothing$，已选入了 k 个特征的特征子集记为 F_k，把未选入的 $n-k$ 个特征 $F_j(j=1,2,\cdots,n-k)$ 逐个与已选入的特征 F_k 组合计算判据值 J，若 $J(F_k+x_1)\geq J(F_k+x_2)\geq\cdots\geq J(F_k+x_{n-k})$，则 x_1 选入，下一步的特征组合为 $F_{k+1}=F_k+f_1$，该过程一直进行到最大判据 J 值降低为止，从而避免搜索整个特征空间，该算法时间复杂度 $\leq n(n-1)/2$，搜索过程如表 1 所示.

表 1　SFS 搜索过程

迭代次数	当前特征子集	评估值	最优特征子集
1	f_1	30	f_3
	f_2	20	
	f_3	35	
	f_4	25	
2	f_1f_3	40	f_2f_3
	f_2f_3	50	
	f_3f_4	45	
3	$f_1f_2f_3$	40	$Stop(f_2f_3)$
	$f_2f_3f_4$	45	

3.3　FSEN 算法

算法 1 的伪代码描述了基于选择性集成策略的嵌入式特征选择方法的具体执行过程. 行 1～4 采用 5 种特征选择器提取特征子集，包括相关性、信息增益、统计、一致性度量，每种算法是各个度量指标的代表性算法，包括 FCBF[15]，InfoGain，GainRatio，Chi-square，CBC. GainRatio 作为一种补偿措施来解决 InfoGain 偏向选择取值多的属性的不足，但它也有可能导致过分补偿，因此两种算法可以互为补充. 行 5～8 根据选择性集成策略选取部分特征选择器，行 5 评估每个特征子集，行 6 选择评估指标最高的 3 个特征子集对应的选择器，行 7 合并特征子集的特征，行 8 返回相关性较高的特征子集，但其中有冗余特征还会降低分类性能，如何消除这些冗余特征是整个特征选择过程的关键. 行 10～16 采用启发式搜索策略从行 8 返回的特征中选择最优特征子集，直至加入特征后分类准确率下降. 行 11 采用序列前向搜索方法产生特征子集，行 12 根据朴素贝叶斯算法评估每轮特征子集 S 的分类准确率，行 13 选出最高的分类准确率，行 17 根据分类准确率找出不同数据集的全局特征子集. 剔除了不相关和冗余特征的全局特征子集有利于简化分类模型，提高分类准确率和稳定性.

算法 1.　FSEN 算法伪代码.

输入：

数据集 *Data*：Traffic data sets

特征选择器 *T*：Five Feature selectors

特征子集 *Subset*：Feature subset

特征集合 *F*：Features in subsets

输出：

全局特征子集 *Global feature subset*

```
1.  for data in Data do //Part 1 Selectors Ensemble
2.    for t in T do
3.      Subset[optimal]:=FindOptimalSubset(data,t)
4.    end for
5.    β:=Evaluate(Subset[optimal])
6.    β[top]:=FindTopThree(β)
7.    F∩=F{β[top]}
8.    return F //Part 1 finished
9.    repeat // Part 2 Secondary Feature Selection
10.     for f in F do
11.       Subset:=GenerateSubset(F)
12.       θ:=Evaluate(Subset)
13.       Subset[best]:=FindBestSubset(Max(θ))
14.       F−=f
15.     end for
16.   until F∈∅ ‖ θ[iteration+1]<θ[iteration]
17.   Subset[global]=FindOptimal(Subset[best])
18. end for
19. return Subset[global] //Part 2 finished
```

4 评价标准

4.1 性能指标

准确率常用于评价识别新流量的能力.假设 N 为流量样本数,m 为应用类型数.n_{ij} 表示实际类型为 i 的应用被标记为类型 j 的样本数.真正 TP 代表实际类型为 i 的样本中被正确标记的样本数,$TP_i=n_{ii}$.假负 FN 代表实际类型为 i 的样本中被误标识为其他类型的样本数,$FN_i=\sum_{j\neq i}n_{ij}$.假正 FP 代表实际类型为非 i 的样本中被误标识为类型 i 的样本数,$FP_i=\sum_{j\neq i}n_{ji}$.根据这些概念,给出衡量分类模型整体准确率(*Overall Accuracy*)、查准率(*precision*)、查全率(*recall*)和综合评价(*F-Measure*)的形式化描述.

$$\text{整体准确率 } OA=\frac{\sum_{i=1}^{m}TP_i}{\sum_{i=1}^{m}(TP_i+FN_i)} \tag{12}$$

$$\text{查准率 } precision=\frac{TP_i}{TP_i+FP_i} \tag{13}$$

$$\text{查全率 } recall=\frac{TP_i}{TP_i+FN_i} \tag{14}$$

$$F\text{-}Measure=\frac{2\times presion\times recall}{presion+recall} \tag{15}$$

查准率和查全率体现了识别方法在每个单独协议类别上的识别效果,整体准确率体现了识别方法的总体准确率,*F-Measure* 是查准率和查全率的综合评价指标.一个好的方法不仅要求具有较高的总体准确率,还应该在各个类别上具有较高的查准率、查全率和 *F-Measure*,特别当样本类别分布不均匀时,查准率、查全率和 *F-Measure* 可以准确获知每个类别的分类情况.

4.2 稳定性

概念漂移是实际分类过程中最常见的问题,类别分布随时间发生改变,特征选择方法很难选取稳定的特征子集来保持较高的分类精度.文献[16]采用汉明距离(Hamming Distance)和 Tanimoto 系数作为稳定性度量,但只适用于固定大小的特征子集.另外,由于不同特征选择方法度量标准不统一,无法比较.本文提出的稳定性度量可以统计不同大小的特征子集,也可以比较不同特征选择方法获得的稳定性.另外,针对出现频率高的特征对稳定性贡献大,采用加权方式突出其稳定性作用.因此,有必要评估特征子集的稳定性,特征选择的稳定性主要研究当样本类别分布发生变化时,特征选择算法的鲁棒性.特征选择方法不仅要获得很高的分类准确率,可靠的稳定性也必不可少.

令特征子集 $S=\{S_1,S_2,\cdots,S_n\}$,集合 $X=\{f\mid f\in S,F_f>0\}=\bigcup_{i=1}^{n}S_i$,$X\neq\varnothing$ 包含 S 的所有特征,特征 f 出现次数为 F_f,总出现次数 $N=\sum_{y\in X}F_f=\sum_{i=1}^{n}|S_i|$,特征一致性为

$$C(f_i)=\frac{F_{f_i}-F_{\min}}{F_{\max}-F_{\min}} \tag{16}$$

最小出现次数 $F_{\min}=1$,最大出现次数 $F_{\max}=n$.$C(f_i)=0$ 表示 f_i 出现次数为 1;$C(f_i)=1$ 表示 f_i 出现次数为 n,平均一致性为

$$C(S)=\frac{1}{|X|}\sum_{f\in X}C(f_i)=\frac{1}{|X|}\sum_{f\in X}\frac{F_{f_i}-F_{\min}}{F_{\max}-F_{\min}} \tag{17}$$

带权重的一致性为

$$CW(S)=\sum_{f\in X}w_f\frac{F_{f_i}-F_{\min}}{F_{\max}-F_{\min}} \tag{18}$$

即

$$CW(S)=\sum_{f\in X}\frac{F_{f_i}}{N}\cdot\frac{F_{f_i}-F_{\min}}{F_{\max}-F_{\min}} \tag{19}$$

如果 $CW(S)=0$,当且仅当 $N=|X|$,每个特征只出现一次;如果 $CW(S)=1$,当且仅当 $N=n|X|$,每个特征只出现一次;如果 $N>|X|$,肯定有特征出现超过一次,$CW(S)>0$.

5 实 验

5.1 实验数据集

目前还没有一个权威的数据集来评测流量分类性能,采用不同数据集进行分类研究,可以有效验证该算法的有效性,本文采用 CERNET 华东(北)地区网络中心采集的 CERNET 数据集(CNT)和 Moore_Set[15](MS)两组数据集.

CNT 数据集是采用 tcpdump 抓取华东(北)网络中心 16 个 C 类地址 2014 年 4 月 2 日 13:00 约 60 分的双向全报文数据,大小为 30 GB,构成 5 个数据集,采用改进 OpenDPI 获取五元组标准集[17],再利用 tcptrace 获取双向流的 110 种统计特征,CNT 数据集共包含 352 884 个完整的双向流网络流样本,被分为 7 类,流样本类别具体分布如表 2 所示.

MS 数据集是在同一结点处随机抽样产生,数据集只选用语义完整的 TCP 双向流作为网络流样

本，每条流包含 249 项特征属性，共包含 9 种类型 376832 个网络流样本，被分为 5 个数据集，每类网络流的数量和所占的比例见表 3.

表 2　CNT 数据集统计信息

类别	数目	百分比/%	类别	数目	百分比/%
HTTP	274958	77.92	QQ	7312	2.07
Flash	12110	3.43	Bittorent	2796	0.79
SSL	9559	2.71	Nomatch	44350	12.57
ICMP	2937	0.83	Total	352884	100.00

表 3　MS 数据集统计信息

类别	数目	百分比/%	类别	数目	百分比/%
WWW	328092	87.07	P2P	2094	0.56
MAIL	28567	7.58	Data-base	2648	0.70
FTP-control	3054	0.81	FTP-data	5797	1.54
FTP-pasv	2688	0.71	Services	2099	0.56
Attack	1793	0.48	Total	376832	100.00

5.2　实验流程

图 2 描述了特征选择算法的具体流程[18]. 首先，评价函数对产生的特征子集进行评价，直至评价结果满足结束准则为止，否则继续评价下一组特征子集. 然后，根据选取的特征子集建立分类模型，根据模型识别新的应用，再计算分类准确率，最后，对同一算法产生的多个特征子集计算稳定性. 机器学习算法分类精度采用十折交叉验证进行评估，十折交叉验证是将数据集分成 10 份，轮流将其中 9 份作为训练数据，1 份作为测试数据，10 次结果的正确率的平均值作为对算法准确率的估计.

本文基于 Weka-3.7.10[19] 二次开发，在 eclipse 上调用 Weka 的 API 完成特征选择和流量分类任务. 所用实验平台运行 Windows 7 操作系统，CPU 为

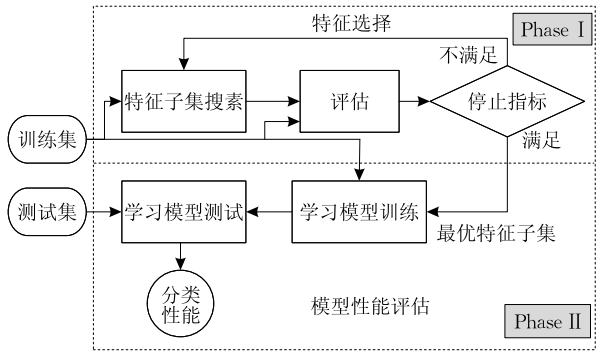


图 2　特征选择评估过程

Intel Core i5-3210，2.5GHz，内存为 DDR3 1600MHz 4GB，Java 开发平台 eclipse-4.2.2.

6　实验分析

6.1　准确性

当前网络上运行着大量的应用，同时新应用不断出现，每个应用都有独特的流统计特征，流统计特征随着时间推移发生概念漂移，使得分类器很难保持较高的分类准确率. 概念漂移导致类分布不断变化，使得特征选择方法难以获得稳定的特征子集，因此，有必要选择稳定的特征子集，使其能在很长一段时间维持稳定的分类准确率. 将 FSEN 算法在 CNT 数据集上进行特性选择，采用朴素贝叶斯算法获得分类准确度，并与 5 种常用特征选择算法（FCBF[15]，InfoGain，GainRatio，Chi-square，CBC）进行对比，其中 FCBF 在文献[15]中被采用，分类准确率如表 4 所示. 为了进一步验证算法的可行性，再统计 MS 数据集分类准确率，如表 4 所示.

表 4　分类准确率

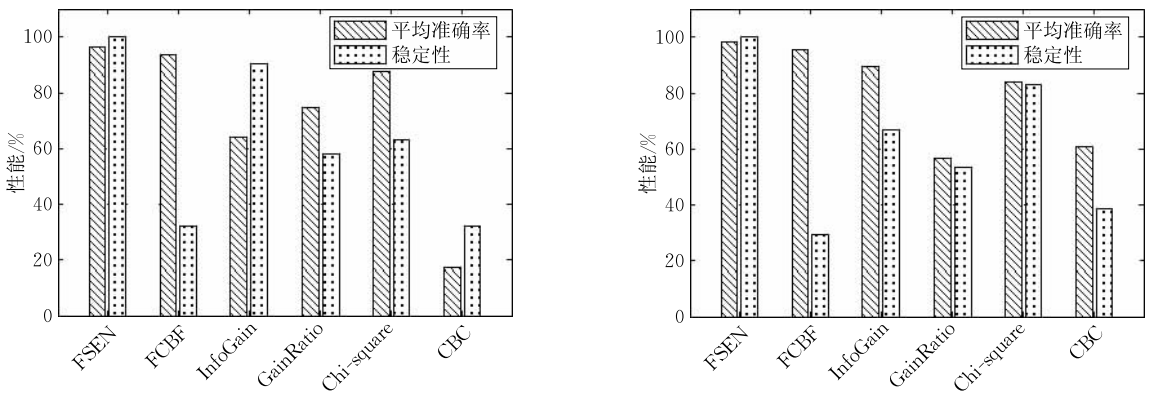
方法	数据集									
	CNT1	CNT2	CNT3	CNT4	CNT5	MS1	MS2	MS3	MS4	MS5
FSEN	96.22	96.54	96.42	95.64	96.49	97.46	98.86	98.43	97.42	97.98
FCBF	92.21	95.01	92.51	92.43	94.52	96.98	94.60	93.65	95.48	94.91
InfoGain	44.33	77.53	76.14	94.57	28.36	90.86	89.04	96.50	85.23	83.91
GainRatio	92.68	94.09	90.86	92.79	4.37	10.37	10.24	84.13	89.68	88.23
Chi-square	59.55	95.59	90.02	95.49	95.61	96.26	95.05	96.96	91.99	39.99
CBC	35.50	13.02	13.54	5.53	18.48	21.58	93.58	76.93	31.51	79.43
Original	12.03	11.14	19.93	11.62	18.12	57.89	60.70	84.45	74.51	79.29

注：Original 代表未进行特征选择的特征全集.

由表 4 可以看出 FSEN 特征选择后的分类准确率均高于其他特征选择算法，而且准确率比较稳定；FCBF 获得较稳定的分类准确率，但总体分类准确率低于 FSEN，Original 分类准确率不高，因为特征全集中存在不相关和冗余特征，与 Original 比较可以看出，有些数据集特征选择后的分类准确率反而变低，特征选择后的分类模型反而变差了，说明有些特征选择方法鲁棒性差. 另外，从表 4 中可以看出，FSEN 和 FCBF 算法分类准确率相对稳定，其余算法随着时间推移准确率有较大变化，很不稳定，说明传统的特征选择方法无法获取稳定的特征子集应对概念漂移问题. 而 FSEN 分类准确率稳定在 95% 以

上,因为 FSEN 算法综合多个特征选择方法的优点,剔除概念漂移产生的局部最优特征,获得稳定的特征子集.为了进一步验证 FSEN 算法应对概念漂移的有效性,采用平均准确率和稳定性从整体上评价特征选择方法,如图 3 所示.

图 3 可以看出 FSEN 算法的平均分类准确率和稳定性均高于其他算法,其他算法存在分类准确率不稳定现象,FCBF 虽然获得了较高的分类准确率,但稳定性相对较低;而 Chi-square 和 InfoGain 算法的分类准确率和稳定度均高于 60%.从稳定性来看,FSEN 算法明显高于其他算法,因为 FSEN 算法选取稳定的特征子集作为不同数据集的全局特征子集.综合来看,FSEN 算法具有较好且稳定的特征选择能力,达到分类准确率和稳定性的最优平衡.一个好的识别方法不仅要有较高的识别准确率,还应该在每个待识别的应用上具有较高的查准率、查全率和 *F-Measure*.各个应用的样本分布不均匀,对每个应用的查准率、查全率和 *F-Measure* 特别重要.分类准确率只能综合评价整个数据集的识别精度,查准率、查全率和 *F-Measure* 可以有效评价各类的分类情况,各个特征选择算法的准确率、查准率、查全率和特征数目如表 5 所示.



(a) 数据集CNT (b) 数据集MS

图 3 平均准确率和稳定性

表 5 查准率和查全率

方法	准确率		查准率		查全率		特征数	
	CNT	MS	CNT	MS	CNT	MS	CNT	MS
FSEN	96.26	98.03	96.86	98.10	95.88	97.94	2	3
FCBF	93.34	95.12	95.98	93.34	93.32	84.32	5	7
Chi-square	64.19	89.11	96.96	97.02	64.22	89.10	5	10
GainRatio	74.96	56.53	95.80	69.20	74.98	56.50	5	10
InfoGain	87.25	84.05	96.04	97.02	87.24	84.08	5	10
CBC	17.21	60.61	96.56	84.84	17.20	60.60	9	8

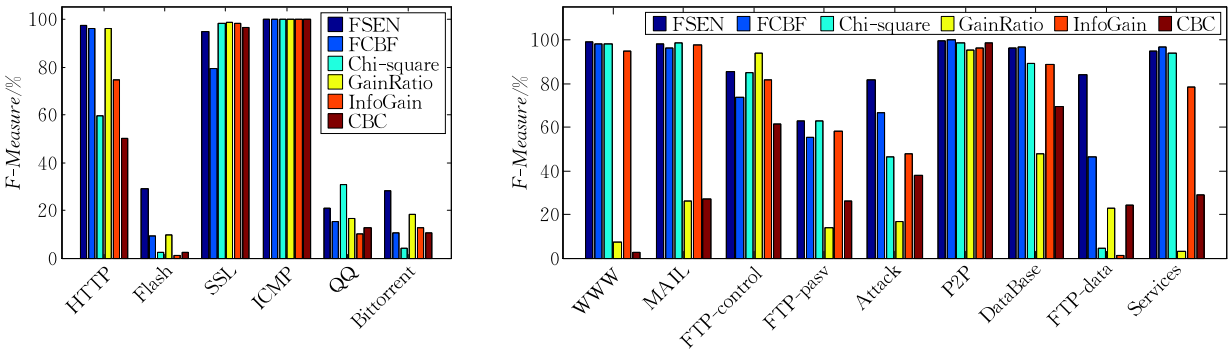
表 5 可以发现 FSEN 算法的查准率和查全率较为接近,表明分类性能稳定.然而,FCBF 具有较高的查准率及较低的查全率,表明其中有应用类型分类精度不高,需要对该类型进行增量学习.从特征数目来看,FSEN 算法选取的特征子集数目最小,因为 FSEN 算法借助集成学习的优势,保留了不同数据集的本质特征,对部分相关的特征予以剔除,具体特征描述如表 6 所示.CNT 数据集的特征子集是 Server Port(端口号)和 Missed data(丢失字节数),虽然基于端口号的分类方法由于动态端口号而失效,但端口信息仍然是重要的特征;Missed data 表示实际收到的字节数与期望收到的字节数的差值;MS 数据集的特征子集是 Server Port, Ave_seg_size(平均包大小)和 Init_win_bytes(初始窗口字节

表 6 FSEN 算法产生的特征子集

数据集	简称	特征描述
CNT	Server Port	Port Number at server
	Missed data	Difference between ttl stream length and unique bytes sent
MS	Server Port	Port Number at server
	Ave_seg_size	Average segment size:data bytes divided by #packets(server to client)
	Init_win_bytes	Total number of bytes sent in initial window(client to server & server to client)

数)，Server Port 是重要特征，Ave_seg_size 表示平均包大小，不同的应用包大小差别很大，Init_win_bytes 表示初始窗口发送的字节数，两组数据集包含的特征区别性都很强，而且特征之间也不存在冗余性. 说明 FSEN 算法选取的特征作为全局特征子集，不受概念漂移作用的影响.

图 4 描述了数据集 CNT1 和 MS1 各个类别的综合评价 *F-Measure*，从 CNT1 图中可以看出，FSEN 算法除了类 SSL 和 QQ 的 *F-Measure* 低于 Chi-square 算法以外，其余类的 *F-Measure* 都高于其他算法，特别是类 Flash 和 Bittorrent 的 *F-Measure* 明显高于其他算法. 另外，从 MS1 图中可以看出，FSEN 算法的 *F-Measure* 在各个类别上都超过 60%，除了类 DataBase 和 Services 的 *F-Measure* 略微低于 FCBF 算法以外，其余类别的 *F-Measure* 都是最高的. 因为 FSEN 算法集成多个特征选择方法的优点，同时兼顾了多种度量标准，而不是从单一的度量考虑. 综合来看，FSEN 算法的 *F-Measure* 明显优于其他算法，在各个类别上都获得较高的分类性能，有效处理了类不平衡问题.



(a) CNT1　　(b) MS1

图 4　综合评价 *F-Measure*

6.2　效率性指标

选取不同大小的 CNT 数据集(10 000，20 000，40 000，80 000，160 000)执行 FSEN 特征选择，每个过程执行 10 次取平均值，各个特征选择算法的运行时间如图 5 所示. 另外，采用 FSEN 特征选择在 3 个 CNT 数据集上执行特征选择，特征子集的模型建立时间和分类时间分别如图 6、图 7 所示.

从特征选择执行时间可以发现 FCBF、GainRatio、Chi-square 和 InfoGain 所需的时间较少，而 CBC 的执行时间较长，FSEN 执行时间长是因为集成了多个特征选择算法，可以采用并行计算来加快

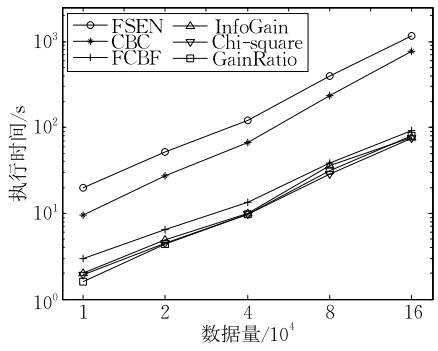


图 5　特征选择执行时间

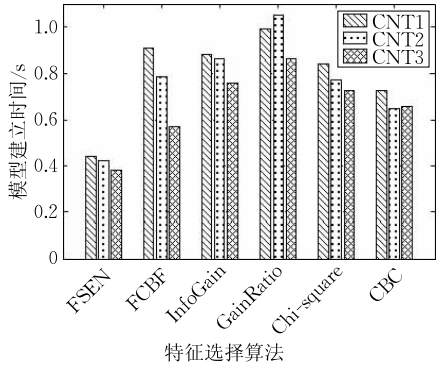


图 6　模型建立时间

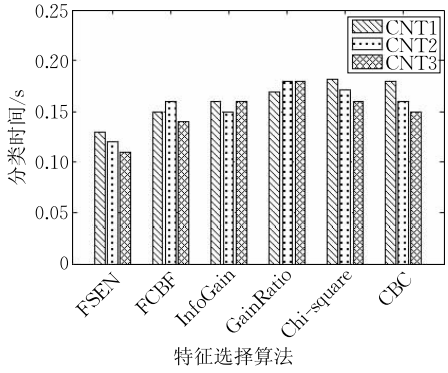


图 7　分类时间

处理速度.从模型建立时间来看,FSEN 低于其他特征选择算法,但相差不大.而从分类时间来看,FSEN 算法明显低于其他算法.综合来看,FSEN 算法不论是模型建立时间还是分类时间都少于其他特征选择算法,主要是因为 FSEN 算法产生的特征数目较少,简化了分类模型,同时提高了分类准确率和稳定性,最终达到分类准确率和稳定性的最优平衡.

6.3 特征离散化

采用 5 种常用的机器学习算法(C4.5,BayesNet,KNN,NB 和 SMO)分类 FSEN 算法获取的特征子集,结果显示 NB 和 SMO 算法的识别精度不太稳定,如图 8 所示.不管采用什么流统计特征,C4.5[20]、BayesNet 和 KNN 分类准确率总高于 NaiveBayes 和 SMO,发现前 3 种算法分类前对数据进行"离散化"预处理[21].机器学习算法主要用于处理离散型数据,虽然可以分类连续性数据,但效果和效率低[22].由于流统计特征中存在连续型数据,导致机器学习分类性能下降.离散化可以简化数据,消除噪声,使得分类器更快、更精确、鲁棒性更好.同时,最小化类和属性之间的相互依赖.因此,采用一种基于最小描述长度原理的启发式离散化算法[23]来解决.离散化方法根据信息论原理将连续属性分割成多个离散区间,以最小描述长度为控制离散化算法的停止指标,在分类错误与离散区间之间找到一个最优平衡.采用离散化方法后获得的分类准确率如图 9 所示.

通过比较不同机器学习算法采用离散化预处理前后的分类性能,对比图 8、图 9 可以发现离散化后的 NB 和 SMO 分类准确率明显提高,另外 3 种算法保持较高的分类准确率.离散化有助于分割连续型数据与类分布一致,简化分类模型,同时消除数据中部分噪声,提高机器学习算法分类效果和稳定性.

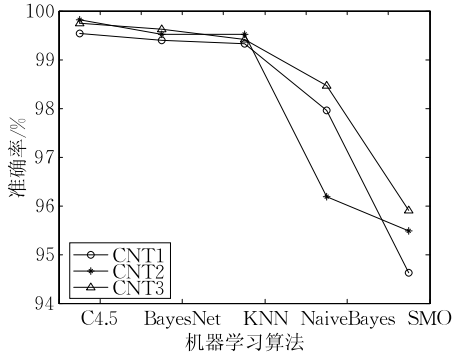


图 8 分类准确率

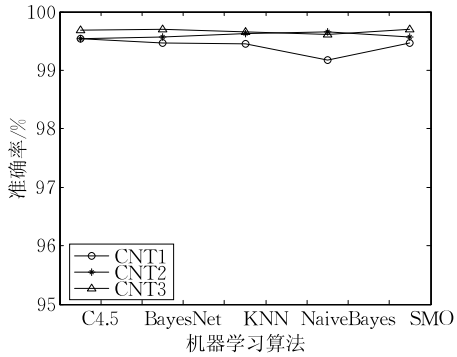


图 9 离散化方法的分类准确率

7 结 论

特征选择从高维数据中选取最优特征子集,有利于提高模型鲁棒性,减少模型建立时间和分类时间,从而提高分类准确率和泛化能力.本文提出了一种基于选择性集成策略的嵌入式特征选择方法,通过准确率、稳定性和时间性能比较不同特征选择算法的性能.实验结果表明该特征选择算法稳定性强,而且特征子集较小,有效简化分类模型,提高分类效果和效率,使分类准确率与稳定性达到最优平衡.另外,采用离散化方法简化数据,有效减少数据噪声,进一步提高机器学习算法分类稳定性.下一步主要研究将结合本文提出的嵌入式特征选择算法,采用集成学习与代价敏感学习方法,通过特征选择和机器学习紧密结合解决流量分类中的类不平衡和概念漂移问题.

致 谢 审稿人对论文提出了宝贵意见,在此表示感谢!

参 考 文 献

[1] Grimaudo L, Mellia M, Baralis E, et al. Self-learning classifier for Internet traffic//Proceedings of the IEEE INFOCOM 2013. Turin, Italy, 2013: 3381-3386

[2] Ding L, Yu F, Peng S, et al. A classification algorithm for network traffic based on improved Support Vector Machine. Journal of Computers, 2013, 8(4): 1090-1096

[3] Zhou Z H, Wu J, Tang W. Ensembling neural networks: Many could be better than all. Artificial intelligence, 2002, 137(1): 239-263

[4] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23(19): 2507-2517

[5] Jin Y, Duffield N, Erman J, et al. A modular machine learning

system for flow-level traffic classification in large networks. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 4

[6] Xie G, Iliofotou M, Keralapura R, et al. Subflow: Towards practical flow-level traffic classification//Proceedings of the IEEE INFOCOM 2012. Orlando, USA, 2012: 2541-2545

[7] Lee S, Kim H, Barman D, et al. Netramark: A network traffic classification benchmark. ACM SIGCOMM Computer Communication Review, 2011, 41(1): 22-30

[8] Zhang H, Lu G, Qassrawi M T, et al. Feature selection for optimizing traffic classification. Computer Communications, 2012, 35(12): 1457-1471

[9] Chen X, Wasikowski M. Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 124-132

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

[11] Li W, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema. Computer Networks, 2009, 53(6): 790-809

[12] Zhang Hong-Li, Lu Gang. Machine learning algorithms for classifying the imbalanced protocol flows: Evaluation and comparison. Journal of Software, 2012, 23(6): 1500-1516 (in Chinese)
(张宏莉，鲁刚. 分类不平衡协议流的机器学习算法评估与比较. 软件学报，2012, 23(6): 1500-1516)

[13] de Souza E N, Matwin S, Fernandes S. Network traffic classification using AdaBoost dynamic//Proceedings of the IEEE International Conference on Communications Workshops 2013(ICC). Budapest, Hungary, 2013: 1319-1324

[14] Fahad A, Tari Z, Khalil I, et al. Toward an efficient and scalable feature selection approach for internet traffic classification. Computer Networks, 2013, 57(9): 2040-2057

[15] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. ACM SIGMETRICS Performance Evaluation Review, 2005, 33(1): 50-60

[16] Somol P, Novovicova J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 1921-1939

[17] Carela-Español V, Bujlow T, Barlet-Ros P. Is our ground-truth for traffic classification reliable? //Proceedings of the Passive and Active Measurement. Los Angeles, USA, 2014: 98-108

[18] Fung P C G, Morstatter F, Liu H. Feature selection strategy in text classification//Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg, Springer, 2011: 26-37

[19] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18

[20] Xu Peng, Lin Sen. Internet traffic classification using C4.5 decision tree. Journal of Software, 2009, 20(10): 2692-2704 (in Chinese)
(徐鹏，林森. 基于 C4.5 决策树的流量分类方法. 软件学报，2009, 20(10): 2692-2704)

[21] Xie Hong, Cheng Hao-Zhong, Niu Dong-Xiao. Discretization of continuous attributes in rough set theory based on information entropy. Chinese Journal of Computers, 2005, 28(9): 1570-1574(in Chinese)
(谢宏，程浩忠，牛东晓. 基于信息熵的粗糙集连续属性离散化算法. 计算机学报，2005, 28(9): 1570-1574)

[22] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining. Knowledge and Information Systems, 2008, 14(1): 1-37

[23] Garcia S, Luengo J, Sáez J A, et al. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 734-750

PAN Wu-Bin, born in 1987, Ph. D. candidate. His research interests include network security, network measurement and traffic classification.

CHENG Guang, born in 1973, Ph. D., professor, Ph. D. supervisor. His research interests include network security, network measurement and behavior, future Internet security.

GUO Xiao-Jun, born in 1983, Ph. D. candidate. His research interests include network security, network measurement and network management.

WANG Yan, born in 1991, M. S. candidate. Her research interests include network security, network measurement and traffic classification.

Background

Network traffic classification is a basic technology in network management and security. Machine learning algorithms are widely used to recent traffic classification, feature selection are used as a pre-processing step to eliminate irrelevant and redundant features, but existing feature selection techniques metrics single, overlook the intrinsic link of characteristic

properties and application itself, and sensitive to variation in traffic data, leading it difficult to keep a better performance. Meanwhile, class imbalance and concept drift affect feature selection to obtain stable feature subset. Therefore, an embedded feature selection based on selective ensemble is proposed, according to the selective ensemble strategy to ensemble part of feature selectors to get a better perfor-mance than ensemble all. And then through the combination method of improved sequence forward search and wrapper secondary search optimal feature subset. Finally, the subset with highest accuracy is selected as a global feature subset. The proposed algorithm can eliminates irrelevant and redundant features, improves the stability of feature subset effectively.

Moreover, a metric of stability is introduced to measure the feature subset, the metric of weighted occurrence frequency is used to evaluate whole feature subset, it can deal with concept drift effectively. Experimental results show that the proposed algorithm can reduce the complexity of feature subset effectively while ensuring the classification performance, so as to achieve the optimal balance of the classification performance, efficiency and stability.

This work was supported by a grant from the National Basic Research Program (973 Program) of China (2009CB320505), the National Natural Science Foundation of China (60973123), the Science and Technology Support Programs (Industry) of Jiangsu Province (BE2011173), and the Prospective Research Programs Future Internet of Jiangsu Province (BY2013095-5-03) to solve traffic classification and network management in High-Speed network and Future Internet.

The research team has focused on network management and security for years, and more than ten papers in this domain have published in highly-ranked conferences and journals.