高技术通讯 2009 年 第 19 卷 第 10 期:998 ~ 1005

doi:10.3772/j.issn.1002-0470.2009.10.002

一种基于扩充特征集的流分类方法①

戴 磊②*** 云晓春* 肖 军*** 陈 友***

(*中国科学院计算技术研究所 北京 100080)

(**中国科学院研究生院 北京 100049)

摘 要 鉴于当前流分类研究均建立在使用载荷无关的流特征的基础上,而载荷无关的特征一般无法为准确分类提供充足的分类信息的问题,提出了一种基于扩充特征集的流分类方法,该方法首先提取载荷特征扩充流分类特征集合,特征集合扩充后,特征的数目显著增加,呈现出高维特性,进而针对高维特征空间,提出了一种混合型特征选择算法,并基于该算法选取的特征构建流分类器。实验结果表明,相对于使用载荷无关特征集的方法,所提出的方法能够显著改善分类效果,同时能够提升分类速度,更适用于现实网络环境。

关键词 流分类,特征选择,信息增益

0 引 言

流分类是增强网络可控性的关键技术之一,它对服务质量(QoS)、安全监控以及网络计费等很多方面都有重要作用。近年来,随着互联网的快速发展,流分类在学术界和业界也备受关注,形成了一个相对独立的研究领域。利用机器学习技术实现流分类是目前研究的热点与难点,该技术能够在很大程度上弥补传统流分类方法的不足,不仅可以提供较快的分类速度,并且具有处理加密流和未知网络应用的能力。此技术使用一组特征向量描述流,特征向量中一般包含包间隔时间、包大小、端口号等信息,分类器根据特征向量的值判断网络流的类别。当前该领域的相关研究都建立在使用载荷无关的流特征的基础上,但载荷无关特征往往无法提供充足的分类信息[1],这导致了该类技术在实际应用中分类准确率不佳,一般只能达到 70%[2-4],尤其是对于 P2P 类的协议,载荷无关特征的描述能力严重不足,其召回率难以超过 60%[1]。针对该问题,此外,本文提出了一种基于扩充特征集的流分类方法。本文使用的数据采自主干网,而已有的研究使用的数据都来自校园网或实验室的网络,相对而言,此项研究的结果更能反映出流分类方法在复杂网络环境下的分类效果。

1 相关工作

随着互联网复杂性的提高,传统流分类技术显现出许多弊端,难以满足应用需求。近年来研究者的目光日益集中在利用机器学习技术进行流分类上。

在文献[5]中,McGregor 等人采用 EM 算法对流进行聚类研究,使用到的流特征包括包大小、包间隔时间及流持续时间。Roughan 等人也根据包的大小和间隔时间提出了一种基于统计指纹方法的流分类方法[6]。在文献[7]中,作者考虑到实时性问题,根据流的前 4 个包提取统计特征,比较了 3 种聚类算法的效果。在文献[8]中,William 等人关注了计算性能问题,他们使用 NetMate[9] 工具获取 26 个流特征,并在特征选择后比较了 5 种分类算法的性能。以上研究都将单独的网络应用作为一类业务类型,如 SMPT、FTP 等均作为单独一类流,而在文献[1]中,Moore 与 Zuev 将相似的网络应用归为一类流,如 BitTorrent 与 Gnutella 同属于 P2P 类别,DNS 与 NTP 同属于 SERVICES 类别,并利用核估计的贝叶斯分类器(Naïve Bayes kernel estimation, NBK)进行流分类研究。

特征选择是重要的分类预处理步骤,利用特征选择技术降低特征维度,不仅能够大量消减建模与

① 863 计划(2007AA01Z444, 2007AA01Z474, 2007AA010501, 2007AA01Z467)和国家自然科学基金(60703021, 60573134)资助项目。

② 男,1979 年生,博士生;研究方向:流分类,信息安全;联系人,E-mail: dailei@software.ict.ac.cn

(收稿日期:2008-07-07)

分类过程中需要处理的数据量,提高计算性能,而且可以消减冗余和无关特征的干扰。该技术已被应用于流分类研究中。在文献[8]中,William 等人分别使用相关性特征选择(correlation-based feature selection, CFS)与一致性特征选择(consistency-based feature selection, CON)选取分类特征子集,从而大幅提高了流分类的性能。在文献[1]中,作者首先从网络流数据中产生 248 个载荷无关的流特征,然后使用快速相关性特征选择(fast correlation-based filter, FCBF)[10]算法选择优化特征子集,实验结果表明只需要选取不超过 20 个特征就能获得最佳分类效果。特征选择算法一般可以归为两类:过滤型和封装型[11,12]。过滤型特征选择算法利用数据本身的特性作为特征子集的评价指标,而封装型特征选择算法依赖特定分类器的正确率作为特征子集的评价指标。通常过滤型特征选择速度比较快,选择的结果与特定分类算法无关,选择效果相对较差;封装型特征选择速度慢,需要耗费大量的计算资源进行交叉验证,选择结果依赖于采用的分类算法,选择效果一般较好。对于高维特征空间,多采用过滤型特征选择技术,相关的流分类研究也都是基于过滤型特征选择技术。为了结合两类算法的优点,更好地处理高维数据,一些研究者提出了混合型特征选择算法[13,14]。

综上所述,近年来研究者们通过使用经典的数据挖掘与机器学习方法,极大地推动了流分类领域的研究发展。为了进一步提高流分类的准确性,本文提出了一种基于扩充特征集的流分类方法。

2　扩充特征集的构造

由于载荷无关的流特征中包含的分类信息有限,基于载荷无关特征的流分类方法分类准确率难以提升,一种改进的途径是同时利用载荷无关特征和载荷特征构造流分类特征集合,并以此为基础建立流分类器。在文献[15]中,作者以流中的包大小、包间隔时间、包头部信息以及持续时间等特性为基础,提取了 248 种可用于流分类的载荷无关流特征,这是已知对载荷无关流特征最完备的描述,表 1 中将其中的一部分特征列出作为示例。

在此基础上,本文考虑选取流的前 k 个字节构成载荷特征扩充流分类特征集合,其中每一个字节均代表一个特征,其值在 -128 至 127 之间。图 1 是根据流的前 k 个字节作特征,使用 C4.5[16] 分类器进行 5 折交叉验证的实验结果,实验中的数据采自骨干网。不难看出,分类准确率随着 k 值的增加而提升,在 200 个字节时基本达到最佳,为 97.99%,因此,选取流的前 200 个字节作为载荷特征扩充流分类特征集合,这样就形成了拥有 448 个特征的扩充特征集,表 2 中列出其中部分特征作为示例。考虑到流分类的实时性需求,所提取的特征均来自流的前 5 个包,在已有的研究中,只有 Zander 在文献[6,17]注意到了这个问题。

表 1　载荷无关流特征示例

特征名称
流持续时间
TCP 端口
包间隔时间(均值,方差,…)
载荷大小(均值,方差,…)
初始窗口大小
包间隔时间傅立叶变换

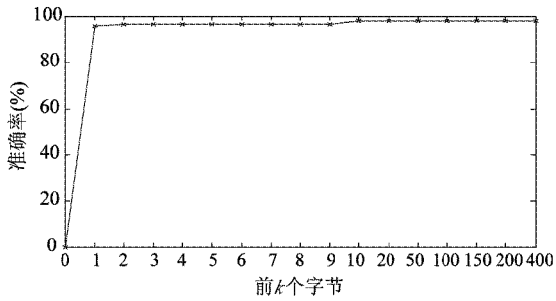

图 1　根据流前 k 个字节特征建立的 C4.5 分类器分类准确率

表 2　扩充特征集包含特征示例

特征名称
流持续时间
TCP 端口
包间隔时间(均值,方差,…)
载荷大小(均值,方差,…)
…
流载荷的第 1 个字节
…
流载荷的第 200 个字节

3　IG-C4.5 混合型特征选择算法

流分类特征集扩充后,呈现出高维特性。针对高维特征空间,简单的过滤型特征选择难以提供优质的分类特征子集,而封装型特征选择算法计算复杂度高,无法处理如此高维的数据,因此,本文提出

了一种混合型特征选择算法,该算法基于信息增益与 C4.5 分类器。信息增益源于 Shannon 提出的信息论,而 C4.5 算法是当前最著名的分类算法之一。根据文献[7]中对 5 种常用算法的流分类对比实验,各分类算法取得的准确率接近,而 C4.5 算法能够提供最快分类速度,选取它作为混合型特征选择算法的评估器,不仅可以提高流分类的速度,还可以减少特征选择过程中交叉验证的时间,3.1 与 3.2 节将对信息增益与 C4.5 分类算法进行简要介绍。

3.1 信息增益(information gain,IG)

信息增益又称为互信息,它可以用来衡量某个特征含有的信息量大小。特征的信息增益越大,代表其包含的分类信息量也越多,对分类的作用也越大。对于特征 A,它的信息增益[18]计算公式为

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \tag{1}$$

$I(s_1, s_2, \cdots, s_m)$是样本的总信息熵,其定义为

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} P(C_i)\log_2 P(C_i) \tag{2}$$

式中的 $P(C_i)$是任意样本属于 C_i 的概率,$P(C_i) = s_i/s$;m 表示样本类别数;s_i 是属于类 C_i 的样本数;s 是总的样本数。式(1)中的 $E(A)$是特征的信息熵,其定义为

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(s_{1j}, s_{2j}, \cdots, s_{mj}) \tag{3}$$

式中的特征 A 具有 v 个不同的值$\{a_1, a_2, \cdots, a_v\}$,可以用特征 A 将 s 划分为 v 个子集$\{s_1, s_2, \cdots, s_v\}$,其中 s_j 包含 s 中 A 值为 a_j 的那些记录。项$(s_{1j} + s_{2j} + \cdots + s_{mj})/s$ 是第 j 个子集的权值,等于子集 $A = a_j$ 中的样本个数除以 s 中样本总数。s_{ij} 是子集 s_j 中类 C_i 的样本数,且式(3)中有

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} P_{ij}\log_2 P_{ij} \tag{4}$$

其中,$P_{ij} = s_{ij}/s_j$ 是 s_j 中样本属于类 C_i 的概率。

3.2 C4.5 算法

Quinlan[16]提出的 C4.5 算法是当前的最著名的分类算法之一。该算法可以分为两个阶段:树的生成和树的剪枝。在第一阶段,算法根据信息增益最大的标准选取特征对训练集进行划分,递归调用直到每个划分中的所有样本都属于同一类;在第二阶段,算法通过计算信息增益率,对建立的树进行剪枝。信息增益率的计算公式为

$$Ratio(A) = Gain(A)/I(S_1, S_2, \cdots, S_v) \tag{5}$$

其中,v 是该节点分枝数,S_i 是第 i 个节点下的记录个数。就建模的计算复杂度而言,C4.5 算法也包含了建树与剪枝两部分,为 $O(mn\log n) + O(n(\log n)^2)$,$n$ 是训练样本集中的样本数量,m 是特征数量。

3.3 IG-C4.5 特征选择算法

扩充后的特征集共包含 448 个流特征,可以计算出其特征子集的数目为 2^{448}。在如此巨大的空间中寻优,搜索策略对于产生特征子集的质量以及算法计算性能均有关键性的影响,而常用的随机搜索算法,比如遗传算法和模拟退火算法等,都无法避免搜索过程中的盲目性。为了降低搜索的盲目性,本文提出了一种变异的混合型特征选择算法。一般混合型特征算法可以分为两个阶段:一是根据数据自身的特性使用过滤型特征选择算法选取候选特征子集;二是在候选特征子集中使用特定的学习算法进行交叉验证,选取优化特征子集。所提出的 IG-C4.5 算法与传统的混合型特征选择算法不同,它在第一阶段不是利用过滤型算法产生候选特征子集,而是通过计算信息增益对所有特征进行排序,在第二阶段使用贪心前向搜索法(greedy forward search,GFS)作为搜索策略,C4.5 分类器作为评估器进行封装型特征选择。通过计算信息增益对特征排序,能够将对类别区分度高的特征排在前方,而 GFS 搜索法则能够优先将这些区分度高的重要特征选入特征子集,进而减少搜索的盲目性,提高特征选择的效率。为了方便表示,使用 $IG_{i,c}$ 表示特征 A_i 和类别 C 之间的信息增益,算法的具体过程如下:

(1) 用式(1)计算每个特征与类别的信息增益 $IG_{i,c}$,根据信息增益值从大到小对特征进行排序。

(2) 特征排序完成后,进入封装型特征选择循环。在循环中用 GFS 作为搜索策略产生特征子集,C4.5 分类器作评估器,选取优化特征子集。

算法使用 C4.5 分类器 5 折交叉验证获得的分类错误率均方差与分类错误率均值的比值作为特征选择循环的终止条件,该值的计算方法如下:

$$f_{\text{error}} = \text{MSE}(R)/\text{MEAN}(R) \tag{6}$$

其中,R 是 5 折交叉验证获取的 5 次分类错误率,$\text{MSE}(R)$ 用于计算分类错误率的均方差,$\text{MEAN}(R)$ 用于计算分类错误率的均值。当 f_{error} 值较小时,表明交叉验证的分类错误率接近,分类结果稳定。设定阈值 ε,当该值小于 ε 时,终止特征选择循环。文中经过多次实验设定 ε 值为 1%。

4　基于扩充特征集的流分类器

基于扩充特征集的流分类器建立在 IG-C4.5 特征选择算法基础上。它的详细流程见图 2。首先,将扩展特征集合和训练集作为输入,计算特征的信息增益,按信息增益值对特征进行排序。然后进入封装型特征选择的迭代循环,在每一次循环过程中,使用 GFS 搜索策略产生新的特征子集 S,以 C4.5 分类算法作评估器,通过 5 折交叉验证计算特征子集的平均分类正确率作为性能评估值 $F(S)$,比较 $F(S)$ 与 $F(S_{best})$,如果 $F(S)$ 小于 $F(S_{best})$,则 $F(S_{best}) = F(S)$,$S_{best} = S$。计算 f_{error} 作为终止条件,当 f_{error} 小于 ε 时,就终止循环。最后,在特征子集 S_{best} 上建立 C4.5 流分类器,并将它应用于流分类。

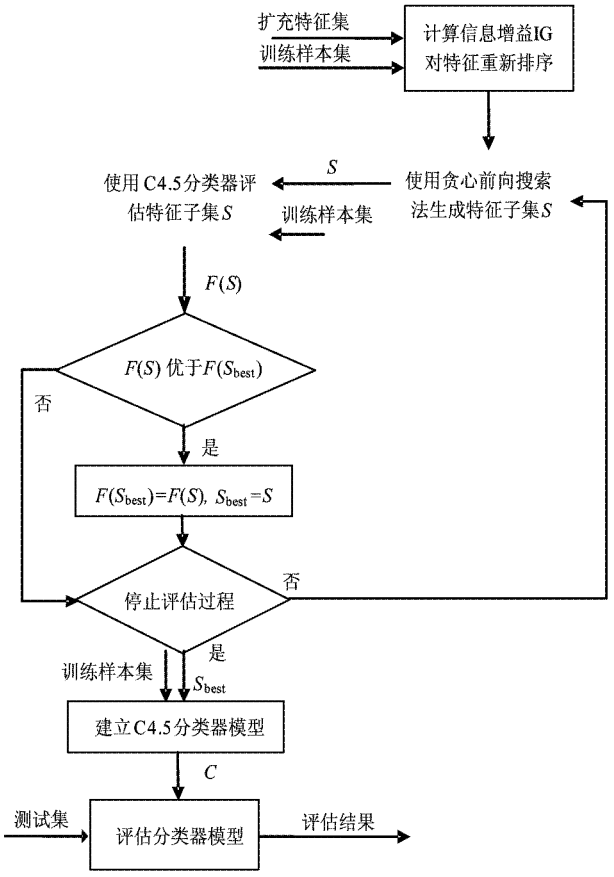

图 2　基于 IG-C4.5 特征选择的流分类流程

5　实验及分析

为了评估所提流分类方法的分类效果与计算性能,本节进行了大量实验。考虑到唯一与之相似的工作是文献[1]中使用 FCBF 结合 NBK 分类器基于载荷无关特征集的流分类研究,因此实验中使用了相同的数据集对两类方法进行评测,并分别比较了它们在分类效果、分类速度以及建模时间上差异。

5.1　数据集

实验中使用的数据集生成自骨干网,其原数据是在国内教育网的一处出口使用高性能网络监控设备捕获的 1 小时全载荷数据包,该点网络速度基本稳定在 200Mbit/s 到 250Mbit/s 之间,捕获数据总量约为 100G。为了方便处理,在保证流完整性不受损害的前提下将原数据文件切分成小块,然后清洗掉难以正确识别和完整性差的流,最后根据提取的特征和识别出的流类别生成数据集。流类别的标注使用到了开源工具 L7-filter[19]。

数据集中每一个样本代表一条流,流定义为具有相同源地址、目的地址、源端口和目的端口四元组的 TCP 包。每个样本都包含 448 个特征,其中有 248 个载荷无关的特征与 200 个载荷特征,文献[15]对载荷无关的流特征进行了详细说明,200 个载荷特征取自流的前 200 个字节,这些特征均取自流的前 5 个包。每个样本除了 448 个特征之外,还有一个标注好的流应用类型,如 WWW, P2P, BULK 等。表 3 列出了数据集中所有的类,如 P2P 类是由 edonkey, gnutella 等应用的流组成。

表 3　网络流的类型(每一个类型下含有多个应用流)

流类别	网络应用
WWW	www
MULTIMEDIA	rtsp, shoutcast
P2P	edonkey, gnutella, bittorrent, 100bao
BULK	xunlei
SERVICES	tsp, dns, ntp
GAME	armagetron
IM	aim

数据集中样本的统计信息详情见表4。可以看

表 4　数据集的流数目统计结果

流类别	流数目
WWW	146658
MULTIMEDIA	55
P2P	126640
BULK	145
SERVICES	182561
GAME	13
IM	3
Total Flows	456075

出在整个数据集中 3 类流的数目最多，其中，WWW 流的数目是 146 658，P2P 流的数目是 126 640，SERVICES 流的数目是 182 561，而整个数据集所有流的数目为 456 075。其余类别由于流数量太小，无法提供足够的分类信息，所以在实验中不予考虑。

5.2 实验方案设计

在实验中，将整个数据集划分成编号为由 1 到 5 的 5 个子数据集。实验分为训练与测试两个部分，首先在 1 个子数据集上选取特征并训练分类模型，然后将其它 4 个子数据集作为测试集来测试在前一个子数据集上建立的分类器的性能。这样在每个数据集上都建立分类器，并用其余 4 个数据集作为测试，就得到了 5 组评估数据。在每次评估中，实验首先用 FCBF 算法在载荷无关特征集上选取特征子集并建立 NBK 分类器，然后使用 IG-C4.5 算法在扩充特征集上选择特征子集并建立 C4.5 分类器，最后用测试集评估两种分类方法的性能。为了评价分类器的分类效果，实验中使用了三个标准指标：准确率(accuracy)：被正确分类的流数目与流的总数目的比值；召回率(recall rate)：对每一类，被正确分类的流数目与这一类流的总数目比值；精度(precision)：被分类器分好的一类流中，正确归类的流数目与被分类器分为这一类的流数目的比值。从两个指标的定义可以看出，准确率是针对所有类的评估指标，而召回率和精度是针对特定类分类性能的评估指标。为了评价流分类算法的计算性能，实验中使用了分类速度与分类器建模时间这两个指标。

实验对基于载荷无关特征集、扩充特征集建立的 NBK 分类器和 C4.5 分类器在各项性能上进行了测试与对比，以评估基于扩充特征集的 IG-C4.5 流分类方法的性能。所有的实验在同一平台下完成，该平台的配置为：Intel processor 2.8GHz，1.00GB RAM，Windows XP 操作系统。

5.3 特征选择

实验使用 FCBF 算法对载荷无关特征集进行特征选择，使用 IG-C4.5 算法对扩充特征集进行特征选择。表 5 列出了对 5 次特征选择的结果。可以看出，IG-C4.5 算法在扩充特征上获取了更多的分类特征，而 FCBF 算法在载荷无关特征集中选取的特征都不超过 4 个，少于文献[1]中的实验结论，造成这种情况的原因一方面可能是数据集中的特征取自流的前 5 个包，导致特征的描述能力降低；另一方面是原数据采自骨干网，流的特性相较于局域网更不稳定。由于详细特征选择结果过于冗长，难以逐一列出，本文仅在表 6 与表 7 中列出了第 3 次评估中选取特征的详情作为示例。

表 5 每个训练集特征选择后的特征数目

训练数据集编号	载荷无关特征数目	扩充特征数目
1	3	5
2	2	6
3	3	5
4	2	4
5	3	7

表 6 第 3 次评估在载荷无关特征集中选取的特征

载荷无关特征集中选取的特征
初始窗口大小（client→server）
完整 RTT 样本总数（server→client）
IP 包的最小字节数（server→client）

表 7 第 3 次评估在扩充特征集中选取的特征

扩充特征集中选取的特征
端口号
流的理论长度（server→client）
分片的包数量（server→client）
流按包数计 3/4 位置处的字节数
流载荷的第一个字节

5.4 分类效果比较

为了验证基于 IG-C4.5 混合型特征选择的流分类方法的分类效果，实验在选择后的载荷无关特征集与扩充特征集上分别建立 NBK 分类器与 C4.5 分类器，并对分类准确率、召回率和精度进行比较。

图 3 在准确率上对两种分类器进行了比较。从图中可以看出，基于扩充特征集的 C4.5 分类器准确率显著高于基于载荷无关特征集的 NBK 流分类器。这说明基于扩充特征集的 IG-C4.5 分类方法获得了更完备的分类信息，能够更准确地完成分类任务。

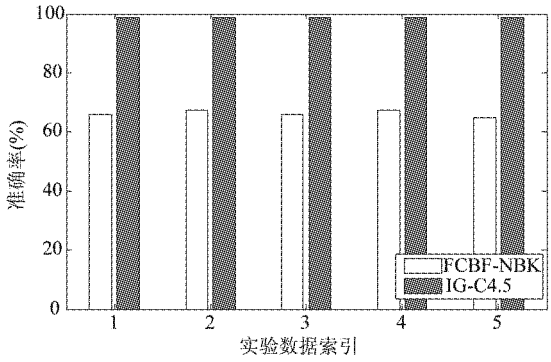

图 3 基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器分类准确率对比

图 4 和图 5 分别比较了分类器在 WWW 流类别上的召回率与精度。可以看出两种分类方法均能较好地区分 WWW 类型的流,在 5 次评测中,基于扩充特征集的 IG-C4.5 流分类方法在召回率上都略优于基于载荷无关特征集的 FCBF-NBK 方法。这表明基于扩充特征集的 C4.5 流分器能够更好地处理 WWW 类型的流。

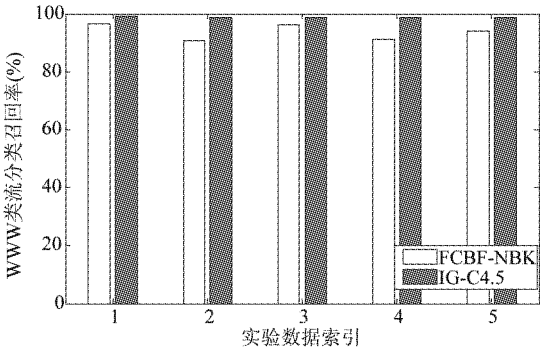

图 4　基于载荷无关特征集与基于扩充特征集 C4.5 分类器在 WWW 类上召回率对比

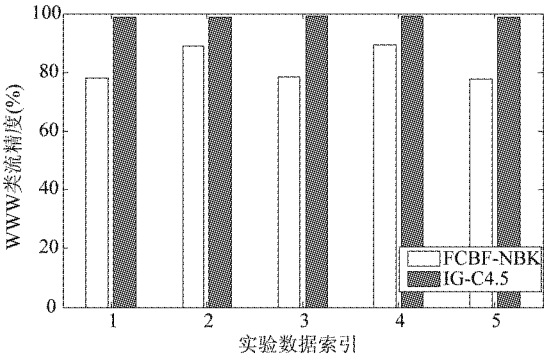

图 5　基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器在 WWW 类上精度对比

图 6 和图 7 是两种分类器在 SERVICE 流类别上的召回率与精度比较。不难看出,两种分类方法对

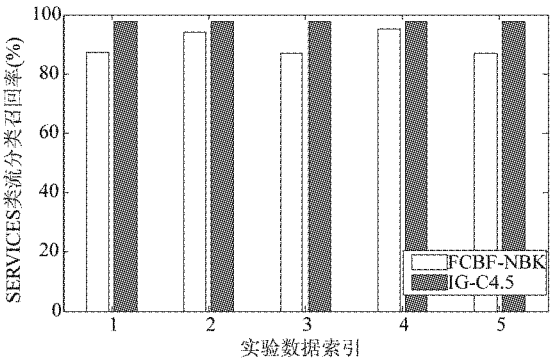

图 6　基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器在 SERICE 类上召回率对比

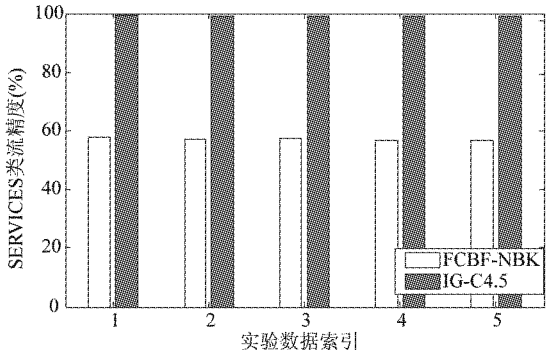

图 7　基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器在 SERICE 类上流精度对比

于 SERVICE 类型的流也均能获得较好的召回率,而基于载荷无关特征集的 FCBF-NBK 分类方法精度较差,它将大量 P2P 类型的流误分为 SERVICE 类型。在 5 次评测中,基于扩充特征集的分类方法在召回率与精度方面都优于基于载荷无关特征集的分类方法。

图 8 与图 9 中比较了分类器在 P2P 流类别上的召回率与精度。从图中可以看出,基于扩充特征集的IG-C4.5方法都能够很好地识别P2P类别的流,

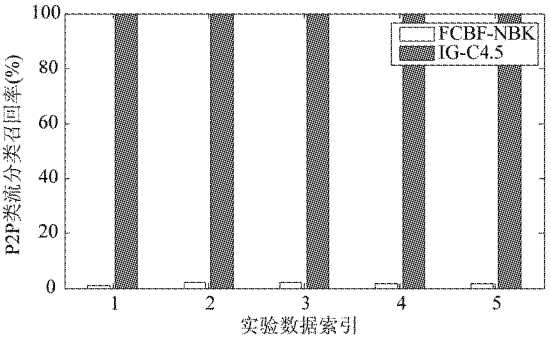

图 8　基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器在 P2P 类上召回率对比

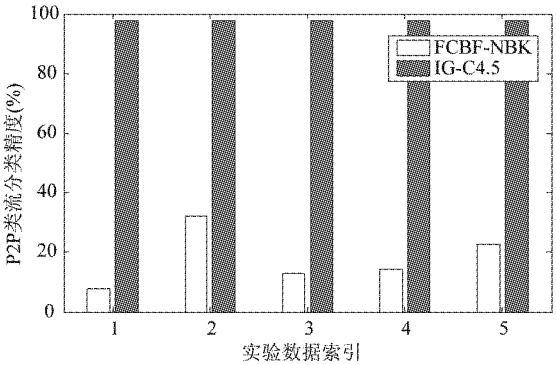

图 9　基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器在 P2P 类上精度对比

识别率接近 100%,而基于载荷无关特征集的 FCBF-NBK 方法分类效果极差,无论在召回率和精度方面都无法满足实际应用需求。造成这种情况的原因是 P2P 流的行为较复杂,并且在主干网络环境下,载荷无关的特征更不稳定,难以依赖它们有效区分 P2P 流类别,而载荷特征对于 P2P 流类别有较好的区分度,基于扩充特征集的 IG-C4.5 流分类方法获取载荷信息,并能够在其中进行有效的筛选,从而可以更准确地识别 P2P 类别的流。

实验表明,基于扩充特征集的 IG-C4.5 流分类方法获得了更丰富的分类信息,并能够在其中有效地筛选高质量的分类特征,以获取更好的分类效果,而基于载荷无关特征集 FCBF-NBK 的方法获取的分类信息不足,难以处理特定的流类别。

5.5 计算性能比较

为了评估基于扩充特征集的流分类方法对计算性能的影响,本节对两种分类器的分类速度与建模时间进行了详细比较。其中,分类速度决定了流分类器的在线吞吐量,是主要的评估指标,而建模时间是指根据训练样本集建立分类器模型的耗时,是次要的计算性能评估指标,只需要处于可接受的范围即可。图 10 提供了两种分类器在分类速度上的对比,为了方便比较,分类速度经过归一化处理,即最大的分类速度每秒 29061 个流作为 1,其它分类速度通过计算每秒处理流的数目与 29061 的比值获取。可以看出,在 5 次评测中基于扩充特征集的分类器分类速度明显优于基于载荷无关特征集的分类器,这是因为基于扩充特征集的流分类方法获取了更优质的特征子集,生成的分类器模型的结构更简单,进而提升了分类速度。

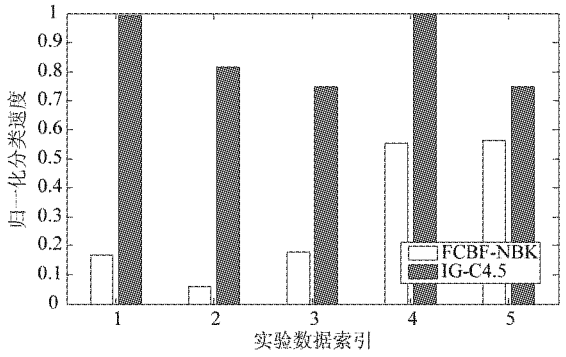

图 10 基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器分类速度归一化对比

在文献[8]中的实验已经对各类算法的建模时间进行了对比,结果表明,NBK 分类器建模耗时最短,C4.5 分类器次之。图 11 是本文实验中 5 次评测的建模时间对比,可以看出评测中 C4.5 分类器的建模耗时都多于 NBK 分类器,与文献[8]中的实验结论一致。这主要是由于 NBK 分类器建模的计算复杂度优于 C4.5 分类器,并且 C4.5 分类器建模中使用了更多的特征,对建模时间也产生了负面影响。虽然 C4.5 分类器需要的建模耗时相对较长,但在 5 次评测中也都不超过 16s。实验表明,与基于载荷无关特征集的 FCBF-NBK 流分类方法相比,基于扩充特征集的 IG-C4.5 方法能够更有效地提高分类速度,虽然该方法的建模时间稍长,但在评测中均不超过 16s,处于可接受的范围。

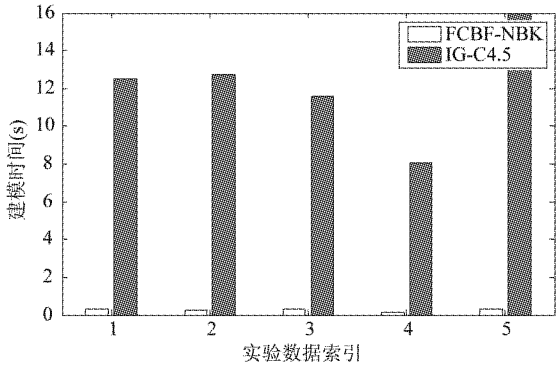

图 11 基于载荷无关特征集的 NBK 分类器与基于扩充特征集的 C4.5 分类器训练时间对比图

6 结 论

当前流分类研究多集中使用载荷无关特征的基础上,而载荷无关的特征往往无法提供足够的分类信息,这造成了该类方法分类准确率不佳,难以在实际网络环境中使用。为了解决该问题,本文提出了一种基于扩充特征集的流分类方法,该方法首先提取载荷特征扩充流分类特征集合。然后提出了一种混合型特征选择算法处理高维特征空间,并基于该算法构造流分类器。实验表明,所提出的方法不仅能够大幅改善分类效果,同时还能够有效提升分类速度,相对于已有方法更适用于实际网络环境。

这项研究还有很多工作需要去完成,比如,使用更高效的特征选择算法和分类算法改善分类性能,研究 UDP 流的分类方法,获取更多的网络数据充实实验内容等。下一步的工作将继续从这些方面展开研究。

参考文献

[1] Moore D, Zuev D. Internet traffic classification using Bayesian analysis techniques. In: Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York, USA: ACM, 2005. 50-60

[2] Moore D, Keys K, Koga R, et al. The CoralReef software suite as a tool for system and network administrators. In: Proceedings of the 15th USENIX Conference on System Administration. California, USA: USENIX Association Berkeley, 2001. 133-144

[3] Logg C, Cottrell L. Characterization of the traffic between SLAC and the Internet. http://www.slac.stanford.edu/comp/net/slac-netflow/html/SLAC-netflow.html: Stanford Linear Accelerator Center, 2003

[4] Moore A W, Papagiannaki D. Toward the accurate identification of network applications. In: Proceedings of the 6th Passive and Active Measurement Workshop. Berlin: Springer, 2005. 41-54

[5] McGregor A, Hall M, Lorier P, et al. Flow clustering using machine learning techniques. In: Proceedings of the 5th Passive and Active Measurement Workshop. Berlin: Springer, 2004. 205-214

[6] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification. In: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM New York, 2004. 135-148

[7] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Proceedings of the 2006 ACM CoNEXT Conference. New York, USA: ACM, 2006

[8] Williams N, Zander S, Armitage G. A preliminary comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review*, 2006, 36(5): 5-16

[9] Zender S. NetMate meter. http://sourceforge.net/projects/netmate-meter: SourceForge, 2006

[10] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2003. 856-863

[11] Liu H, Yu L. Towards integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(4): 491-502

[12] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 2004, (5):1205-1224

[13] Das S. Filters, wrappers and a boosting based hybrid for feature selection. In: Proceedings of the 8th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 74-81

[14] ErHeng Z, Sihong X, Wei F. Graph-based iterative hybrid feature selection. In: Proceedings of the 8th IEEE International Conference on Data Mining. California, USA: IEEE Computer Society, 2008. 1122-1138

[15] Moore A W, Zuev D. Discriminators for use in flow-based classification. http://www.cl.cam.ac.uk/~awm22/publications/moore2005discriminators.pdf: University of London, 2005

[16] Quinlan J R. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann, 1993

[17] Bernaille L, Teixeira R. Traffic classification on the fly. *ACM SIGCOMM Computer Communication Review*, 2006, 36(2):23-26

[18] Baglioni M, Furletti B, Turini F. DrC4.5: Improving C4.5 by means of prior knowledge. In: Proceedings of the 2005 ACM Symposium on Applied Computing. New York, USA: ACM, 2005. 474-481

[19] Strait M, Sommer E. Linux layer 7 packet classifier. http://sourceforge.net/projects/l7-filter/: SourceForge, 2007

Traffic classification based on extended feature set

Dai Lei[*,**], Yun Xiaochun[*], Xiao Jun[*,**], Chen You[*,**]

([*] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**] Graduate University of Chinese Academy of Sciences, Beijing 100049)

Abstract

In consideration of the problem that the present studies on traffic classification are all based on the use of payload-independent features, but the payload-independent features often do not contain sufficient information to allow for an accurate methodology, the paper proposes a traffic classification method based on the extended feature set. The method extends the flow feature set with payload, and after extension, the feature number is significantly increased. For high dimension data, the paper then proposes a hybrid feature selection algorithm for traffic classification and builds the classifier with the selected features. The experimental results demonstrate that the proposed method can not only guarantee a high classification accuracy but also a better performance in terms of classification speed. Therefore, it is more suitable for the real network applications than the traditional approaches.

Key words: traffic classification, feature selection, information gain