DOI:10.19304/j.cnki.issn1000-7180.2008.10.034

一种基于半监督学习的应用层流量分类方法

柳 斌,李之棠,涂 浩
(华中科技大学 网络与计算中心,湖北 武汉 430074)

摘 要:基于应用层的流量分类在用户行为识别、网络带宽管理等方面有着十分重要的应用.将机器学习应用到应用层流量分类问题中.首先提出了一种基于熵函数的组合式特征选择算法,提取了5种TCP连接的特征.针对监督学习中无法识别新流量类型的问题,提出了一种基于半监督学习的流量分类算法.实验结果表明,算法的检测率优于Kmeans方法.在少量标记样本的情况下,随着未标记样本数增加,算法的检测率在增加.

关键词:流量分类;半监督学习;特征选择;熵
中图分类号:TP393.08 文献标识码:A 文章编号:1000-7180(2008)10-0113-04

Network Application Classification Method Based on Semi-Supervised Learning

LIU Bin, LI Zhi-tang, TU Hao
(Network Centre, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Categorizing network traffic by application type is challenging because of the continued evolution of applications. In this paper, a feature selection method based on entropy is proposed firstly. Five TCP connection features are selected for classifier. Then, a semi-supervised classification method that allows classifiers to be designed from training data consisting of only a few labelled and many unlabelled is proposed. The experiment shows that the detection results precedes on Kmeans method.

Key words: traffic classification; semi supervised learning; feature selection; entropy

1 引言

应用层是TCP/IP网络体系架构中的最高层.按照应用层对流量进行识别和分类在用户行为分析、网络带宽管理、入侵检测等诸多方面有着十分重要的应用.应用层流量分类本质上是一个模式识别或模式分类问题,因而,可以将机器学习技术应用到流量分类领域[1-2].在机器学习领域,分为有监督学习和无监督学习.基于监督学习的方法如贝叶斯方法、决策树方法等,检测率高,但要求样本数据事先正确标记类别,因此无法发现未知的类别,且需要足够的训练数据,以生成具有良好的泛化性能的检测模型.非监督学习方法如聚类的方法,根据数据的相似性进行分组,克服了监督学习方法中需要标记数据的不足,但其检测精度明显低于有监督的分类方法.它只能对未标记数据建模,不利用标记数据的类别信息.文中提出了一种基于半监督学习的流量分类方法,利用少量的标记数据辅助聚类过程,确定簇与流量类型的映射关系,实现应用层流量分类.算法能发现未知的新的应用类型,提高了检测精度,同时,降低了对标记数据的需求.

2 基于熵函数的组合式特征选择方法

基于机器学习的分类方法首先要解决的一个问题就是特征提取与选择.一般情况,只有特征向量中包含足够的类别信息,才能通过分类器实现正确分

收稿日期:2008-06-18
基金项目:国家"八六三"计划项目(2007AA01Z420)

类,而特征中是否包含足够的类别信息却很难确定.为了提高识别率,总是最大限度地提取特征信息,如文献[3]中提取了 TCP 连接的 248 种特征.过高的特征维数给特征的进一步处理和分类器的实现带来了很大的困难.因此,需要进行特征选择,尽量降低特征空间维数.

特征选择的目的主要有两点,一是去除对于类别属性无关的特征.二是去除冗余的特征.文中提出了一种基于熵函数的组合式特征选择方法.首先计算所有特征的熵,前 d 个最优特征拿来构成候选的特征子集,这些特征对于分类而言都是有意义的,从而去掉了对于类别无关的特征.其次采用顺序后退搜索方法(SBS),以分类器本身的分类准确率为评估标准,依次考察去除特征对分类器准确率的影响,藉此去除冗余特征.

采用特征的后验概率分布衡量特征对分类的有效性,后验概率分布越集中,分类错误概率就越小,后验概率分布越分散接近均匀分布,分类错误概率就越大.从特征提取角度看,用具有最小不确定性的那些特征对分类是有利的.在信息论中,熵作为不确定性的度量,因此可以借助信息熵作为衡量后验概率分布的集中程度的定量指标.

对于 C 类问题,考查特征 X 的分类能力.将 X 的取值范围分为 k 段,样本落在第 j 端的概率为 $P(b_j)$.第 i 类样本落在第 j 端的概率为 $P(\omega_j \mid b_j)$.那么,熵定义为

$$H = -\sum_{j=1}^{k}\sum_{i=1}^{m} p(b_j)p(\omega_j \mid b_j)\log_m P(\omega_j \mid b_j) \tag{1}$$

熵越小,则特征 X 的分类能力越强.

组合式特征选择方法算法描述如下:

算法:基于熵函数的组合式特征选择算法.

输入:训练数据集 T,过滤阈值 H,初始特征集 $S=\{F_i,\ i=1,\cdots,n\}$,C 为学习方法.

输出:特征子集.

1. 按照式(1)计算各个特征的熵 $W=\{W_i,\ i=1,\cdots,n\}$;
2. 保留大于阈值 H 的特征构成新的特征集 $S'=\{F_i \mid W_i > H,\ i=1,\cdots,n\}$;
3. $J = C(T, S')$
4. $J_{best} = J$, changed = true
5. while $|S'| > 0$ & Changed {
6. changed = falsed
7. for $i = 1, 2, \cdots, |S'|$ {
8. $M = \{S' - F_i\}$;
9. $J' = C(T, M)$;
10. if($J' < J_{best}$){
11. $J_{best} = J'$ changed = true
12. $F_{worst} = F_i$}
13. }
14. $S' = \{S' - F_{worst}\}$
15. }// end while
16. End if

3 基于半监督学习的流量分类算法 STC

3.1 基本思想

在机器学习中,应用层流量分类问题可描述如下:抽取每条流特征构成特征集合 $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$,其中 $\boldsymbol{X}$ 为特征向量,X_{ij} 表示第 i 个特征向量的第 j 个特征.流特征可以是流的持续时间,流的包数,流的字节数等参数.$Y = \{Y_1, Y_2, \cdots, Y_q\}$ 为应用流量类型,如 HTTP, Mail, P2P 等.应用层流量分类就是寻找 $X_i \rightarrow Y_j$ 的映射关系.有监督的机器学习方法依赖于先验的标注训练样本集,利用大规模标注过的训练数据提高学习算法结果的准确度,但是标记必须由人手工完成,这是一项费时费力的工作.另一方面,由于网络应用的分布情况随时间变化,往往存在着未知的新的应用流量类型,根据标记数据训练的检测模型,无法对未知流量类型进行分类.在文中提出了一种基于半监督的应用层流量分类算法 STC.STC 算法主要有三个部分:

(1)利用少量标记数据,确定 Kmeans 的初始中心.在 Kmeans 算法中,参数为各簇的中心 $\mu_1, \mu_2, \cdots, \mu_k$,簇中心初始值的选择将会严重影响到算法的性能.标记数据能够在一定程度上反映出真实网络数据的分布情况.在拥有少量标记数据的情况下,据此选择适当的模型参数的初始值,改善 Kmeans 性能.

(2)建立映射关系.确定每个簇的应用类型.计算 $P(Y = y_j \mid C_k)$,$j = 1, \cdots, q$,j 为应用类型.$k = 1, \cdots, K$,K 为聚类数目.利用标签数据对 $P(Y = y_i \mid C_k)$ 进行估计.

$$P(Y = y_j \mid C_k) = \frac{n_{jk}}{n_k},$$

式中,n_{jk} 为分配到类 K 标记为 j 的流的数目,n_k 为总的流数目.簇的类型为

$$\text{label}_i = \arg\max_{y_1 \cdots y_q} P(Y = y_i \mid C_k).$$

(3)发现未知应用.利用(2)确定每个簇的应用类型.对于无法标记的簇,利用快速的 k 近邻法[4]搜索距其最近的 k 个数据,并利用多数投票的方式

确定其标记, 此时若仍无法标记数据, 则该数据为新流量类型.

3.2　半监督流量分类算法 STC

输入: 标记数据集合 $D_i=\{(x_i, y_i)\mid i=1, \cdots, n\}$, 非标记数据集和 $D_u=\{x_i \mid i=1, \cdots, n)\}$, 初始类别数 m, K 近邻个数 g.

输出: $x_i \in D_u$ 的流量类型.

算法:

(1) 对标记数据集 D_i 进行聚类. 使得每个簇中仅包含相同类型的数据. 簇中心为 μ_k, $\mu_k=\frac{1}{n}\sum x_i$, $i=1, \cdots, n$, $k=1, \cdots, K$. 对 D_u 划分 m 组, m 组的中心为 μ'_k, $\mu'_k=\frac{1}{n}\sum x_i$, $i=1, \cdots, n$, $k=1, \cdots, K$. n 为簇中流的数目, k 为类别数.

(2) 以 $\mu_k+\mu'_k$ 为初始中心对 $D=D_u\cup D_i$ 进行 Kmeans 聚类, 对 $x_i \in D$ 重复以下步骤, 直至收敛.

对于 $x_i \in D$,

计算 $h=\arg\min_h \| x_i-\mu_h \|$ 修改簇中心, $\mu_k=\frac{1}{n}\sum x_i$, $i=1, \cdots, n$, $k=1, \cdots, K$ 其中 n 为簇中流的数目, k 为类别数.

(3) 确定簇的标记.

计算 $P(Y=y_i \mid C_k)=n_{jk}/n_k$, n_{jk} 为分配到类 K 标记为 j 的流的数目, n_k 为总的流数目. 簇标记 $\text{label}_i=\arg\max_{y_1 \cdots y_q} P(y_i \mid C_k)$.

(4) 对于未标记簇 CU, 重复以下步骤, 直至 CU 集合为空. 计算 $x_i \in$ CU 与已标记簇中心 $C(x_j)$ 及簇半径 $r(x_j)$ 的差值 $d_{\min}$.

$d_{\min}=\min(d(x_i, C(x_j))-r(x_j))$.

若 x_i, $x \in$ CU, $d(x, x_i)<d_{\min}$ 近邻个数 $<g$, 将 x_i 加入到 x 近邻集合中. 如果 $d(x, x_i)>d_{\min}$, 则将距离 x 最近的已标记簇加入到 CU 中, 重复执行(4).

(5) 根据近邻集合进行多数投票, 确定数据的流量类型 label(x):

$$\text{label}(x)=\begin{cases}\text{label}_j & \text{label}_i<\text{label}_j\\ \text{新流量类型} & \text{label}(x)\text{未知}\end{cases}.$$

4　实验与分析

4.1　特征选择

采用了文献[3]中 Andew Moore 数据集作为原始数据集. Andew Moore 数据集利用 TCPTRACE 等工具提取了不同应用的 248 种 TCP 连接特征. 数据集一共有32 000条记录. Andew Moore 数据集将应用分成了 6 类: (1)BULK 类. 如 FTP, xunlei 应用等. (2) Interactive 类. 如 telnet 应用等. (3)WWW 类. 如 http, https 应用. (4)Service 类. 如 Dns 应用. (5)P2P 类. 如 bittorrent, skybe, edonkey 应用. (6)Mail 类. 如 POP3, SMTP 应用等.

首先人工剔除掉一些明显无关的特征, 然后按照式(1), 分别计算剩余 108 种特征的熵值, 熵值从小到大, 前 15 名如表 1 所示. 其次, 在表 1 的基础上, 采用朴素贝叶斯分类器(NBC)以及 SBS 搜索方法, 去除特征间的冗余, 按序保留了 5 个特征作为 STC 算法的最优特征子集, 结果为序列 1, 2, 4, 6, 14. 这时 NBC 的检测率为 92.2%.

表 1　特征熵

序号	特征名称	描述	熵值
1	Serverport	服务器的监听端口	0.12
2	Clinetport	客户端的监听端口	0.14
3	Protocol	协议	0.15
4	Duration	连接持续时间	0.15
5	meanIAT	平均到达间隔	0.16
6	Synpkts(a—b)	a 向 b 发送的 SYN 包数	0.18
7	Ackpkts(a—b)	a 向 b 发送的 Ack 包数	0.19
8	Actualdata(a—b)	a 向 b 发送的实际字节数	0.21
9	Totalpacket(a—b)	a 向 b 发送的总包数	0.21
10	Totalpacket(b—a)	b 向 a 发送的 SYN 包数	0.21
11	Meandataip	平均字节数	0.29
12	Synpkts(b—a)	b 向 a 发送的 SYN 包数	0.43
13	Ackpkts(b—a)	b 向 a 发送的 Ack 包数	0.46
14	Actualdata(b—a)	B 向 a 发送的实际字节数	0.46
15	Avgwin(a—b)	a 向 b 发送的平均窗口大小	0.48

4.2　STC 算法分析

为了测试 STC 算法的性能, 采用检测率衡量算法的检测效果.

检测率=正确标记的流数目/总的标记流数目.

第一个实验分析在标记样本固定的条件下, 未知样本数的变化对检测率的影响, 图 1 给出了在固定标记样本集1 000和10 000两种条件下, 未知样本数变化时的算法检测的准确率. STC 算法的初始类数 $m=200$, 近邻数 $g=30$.

从图 1 中可以看出在标记样本一定的情况下, 增加未标记样本数, 算法准确率可以提高. 这个结果在实际中很有价值, 因为与标记样本或者错误标记

的代价比较,未标记样本是很容易获得的,且代价低廉,可以通过增加未知样本的数来提高学习的检测率.

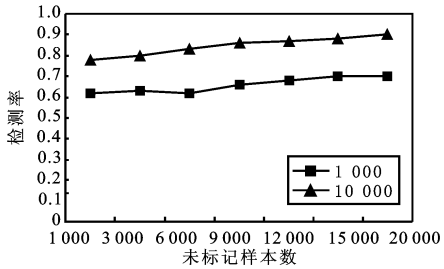

图 1 算法检测率

将STC 算法与典型的有监督学习算法 NBC,无监督的 Kmeans 进行检测率比较.总的样本集为 32 000 条,分成两个子集,从 BULK,INTERACIVE,P2P,MAIL,WWW5 个已知类别的数据集中随机选择若干样本构成已标记样本集(SERVICE 类中的样本作为未知样本),其余为未知样本子集,改变已标记样本集的个数,分析各个算法的检测率.在 NBC 算法中用已标记样本集进行训练,未知样本子集作为测试集,Kmeans 中只对已标记样本集进行测试(取 $K=100$,随机选择簇中心).检测率结果如表 2 所示.

从结果可以看出,STC 算法采用了少量标记数据来确定初始聚类中心,指导聚类过程,其检测精度明显高于随机选择簇中心的 Kmeans 聚类算法,

表 2 结果比较

已标记样本数	STC/%	NBC/%	Kmeans/%
100	75	63	67
1 000	77.2	72	69
5 000	78.2	80.2	70.4
10 000	82.2	90.2	72

STC 在标记样本数低于5 000时比 NBC 算法具有更高的检测率.这是因为 STC 算法能发现新的流量类型,减少了样本错分的情况,提高了检测率.

参考文献:

[1] Auld A W M T, Gull S F. Bayesian neural networks for internet traffic classification[J]. IEEE Transactions on Neural Networks, 2007(18): 223—239.

[2] Zander S, Nguyen T T T, Armitage G. Automated traffic classification and application identification using machine learning[C]// Proceedings of IEEE LCN. Australia, 2005.

[3] Moore A W, Zuev D. Discriminators for use in flow—based classification[R]. Cambridge: Intel Research, 2005.

[4] 俞研, 黄皓. 一种半聚类的异常入侵检测算法[J]. 计算机应用, 2006, 26(7): 64—66.

作者简介:

柳 斌 男,(1971—),博士研究生,讲师.研究方向为网络管理、网络安全.

(上接第 112 页)

机过程与模糊过程的理论与方法,进一步探讨船舶调度问题.

参考文献:

[1] 刘志勤. 集装箱班轮航线调度的随机规划模型[D]. 上海: 上海海运学院, 2005.

[2] 谢新连. 船舶调度与船队规划方法[M]. 北京: 人民交通出版社, 2000.

[3] 王勇. 船岸集成信息系统方案设计[J]. 开发应用, 2003, 12(8): 23—28.

[4] 张方炳, 程正标. 基于 GIS/GPS、移动通信技术的船舶调度、监控系统[J]. 水运程, 2003, 358(11): 15—19.

[5] 谢小良. 随机规划下投资决策优化模型[J]. 统计与决策, 2006, 217(7): 12—46.

[6] 符卓. 开放式车辆路径问题及其应用研究[D]. 长沙: 中南大学, 2003.

[7] Zhuo Fu, Eglese R, Li L. A new tabu search heuristic for the open vehicle routing problem[J]. Operational Research Society, 2005, 56(3): 267—274.

[8] Oguz C, Cheung B. A genetic algorithm for flow—shop scheduling problems with multiprocessor tasks[C]// Proceeding of 8th International Workshop in Project Management and Scheduling. Spain, 2002.

作者简介:

谢小良 男,(1964—),博士研究生,副教授.研究方向为物流工程、优化与决策.

符 卓 男,(1960—),博士,教授,博士生导师.研究方向为交通运输规划与管理、物流配送优化、运筹学.