网络流量特征选择方法中的分治投票策略研究

高 文[1],钱亚冠[1,2],吴春明[1],郭 晔[3],朱 凯[1],陈双喜[4]
(1.浙江大学计算机学院,浙江杭州 310027; 2.浙江科技学院理学院,浙江杭州 310023;
3.浙江大学图书与信息中心,浙江杭州 310027;4.嘉兴职业技术学院,浙江嘉兴 314036)

摘 要: 特征选择作为机器学习过程中的预处理步骤,是影响分类性能的关键因素.网络流量具有数据量大,特征维度高的特点,如何快速提取特征子集,并提高分类效率对于基于机器学习的流量分类方法具有重要意义.本文提出基于分治与投票策略的特征提取方法,将数据集分裂为多个子集,分别执行特征提取算法,利用投票方法获得最后的特征子集.实验表明可有效提高特征提取的时间效率,同时使分类器取得良好的分类准确率.

关键词: 分治; 投票; 流量分类; 特征选择
中图分类号: TP393.4 **文献标识码:** A **文章编号:** 0372-2112 (2015)04-0795-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.04.024

The Divide-Conquer and Voting Strategy for Traffic Feature Selection

GAO Wen[1], QIAN Ya-guan[1,2], WU Chun-ming[1], GUO Ye[3], ZHU Kai[1], CHEN Shuang-xi[4]
(1. *The College of Computer Science, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;
2. *College of Science, Zhejiang University of Science and Technology, Hangzhou, Zhejiang* 310023, *China*;
3. *Library Information Center, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*;
4. *Jiaxing Vocational Technology College, Jiaxing, Zhejiang* 314036, *China*)

Abstract: Feature selection as a substantial preprocess step is a key factor for improvement of classification accuracy. The network traffic is characterized by huge volume and high dimensions. So how to extract the optimal feature subset in short time is practical for traffic classification based on machine learning. A novel method is proposed, which partitions the traffic dataset into several small subsets, and applies special feature selection algorithm to them respectively. Finally, the optimal feature subset is obtained by voting on these alternative feature subsets. The experiment results show that the proposed method has good time efficiency in searching optimal features and helps to improve classification accuracy efficiently.

Key words: divide and conquer; voting; traffic classification; feature selection

1 引言

通过网络流量识别出各种网络应用及网络攻击已成为当前互联网运行的核心任务,以统计学为基础的机器学习方法在网络流量分类中逐渐引起研究人员的关注[1,2].一个典型的有监督机器学习模型的建立过程包括:(1)训练集的建立;(2)分类标注;(3)特征选择;(4)模型构建;(5)模型评估.其中,特征选择是影响分类性能的一个关键因素.理论上,特征越多可以越有效的区分流量类型,但过多的特征将导致模型的建立时间过长、计算资源消耗过大、模型过于复杂等不利因素.同时,特征之间的相关性也可能降低模型的区分能力.因此,如何得到一个优化的特征子集成为机器学习应用于流量分类的一个重要课题[3].

互联网流量具有数据量大、特征变量多的特点,如Moore等[4]提出了248种流量特征.研究表明,训练集的大小与特征选择的优度成正比[5].因此,在实际应用中倾向于采用较大容量的训练集以取得最优的特征子集.然而,训练集容量的增大又将导致算法对计算资源的需求增加,甚至因主存容量的不足而长时间无法完成计算.为解决此问题,我们提出分治和投票的策略(Divide-conquer and Voting, DV),将原始训练集分割为多个较小容量的集合,在每个分割子集上利用基于相关的特征选择算法(Correlation based Feature Selection, CFS)[6]结合最

收稿日期:2013-11-05;修回日期:2014-03-03;责任编辑:梅志强
基金项目:国家973重点基础研究发展计划(No.2012CB315903);浙江省重点科技创新团队(No.2011R50010-21,No.2013TD20);国家自然科学基金(No.61379118);国家科技支撑计划(No.2014BAH24F01);国家863计划(No.2012AA01A507);浙江省网络媒体云处理与分析工程技术中心开放课题(No.2012E10023-14)

佳优先搜索算法(Best-First Forward Search, BF)[7]获得若干特征子集,再通过投票获得最终的特征子集.实验表明,DV 算法与直接在原始集合上应用 CFS + BF 算法相比的优势在于:(1)算法完成时间大大缩短,利用 CPU 的并行性可进一步缩短运行时间;(2)克服了特征选择的过拟合问题,使特征子集的泛化能力提高,分类的准确率得到有效提升.

2 特征选择算法

假设原始的特征集合 $A=\{f_1,f_2,\cdots,f_N\}$,特征子集 $A_M\subset A$ 且 $M\ll N$,那么特征选择问题就是寻找特征子集 A_M^*,满足 $F(A_M^*)=\mathrm{Max}\ F(A_M),\forall\ A_M\subset A$,其中 $F(\cdot)$为评价函数.大量的研究工作集中在对特征子集空间的搜索方法和评价函数这两方面上[8~11].而本文则是从训练集合的角度研究如何加快评价函数的计算效率及减少特征选择的过拟合问题.

理论上可通过穷举法搜索所有可能的特征组合,但 n 个特征的搜索空间将达到 2^n.本文采用最佳优先搜索 BF 算法[7],它基于贪心爬山策略完成特征空间的搜索,实践证明具有良好的搜索效率.在评价函数上,本文采用基于相关性的 CFS 算法.CFS 算法将相关性作为评价函数,因此利用下述相关系数作为评价函数:

$$r_{zc}=\frac{k\bar{r}_{zi}}{\sqrt{k+k(k-1)\bar{r}_{ii}}} \tag{1}$$

其中,z 是类标签,c 是特征子集,k 是特征子集中的特征个数,$\bar{r}_{zi}$表示特征与类之间的平均相关度,$\bar{r}_{ii}$表示特征之间的平均相关度.

本文在 BF 搜索算法与 CFS 特征提取算法的基础上,提出基于分治与投票策略的特征选择方法.具体方法是将大容量训练集合通过分类抽样的方法划分为若干较小的数据子集,并在这些子集上应用 CFS + BF 算法,通过投票的方式得出最终特征子集,整个特征选择算法的框架如图 1 所示.

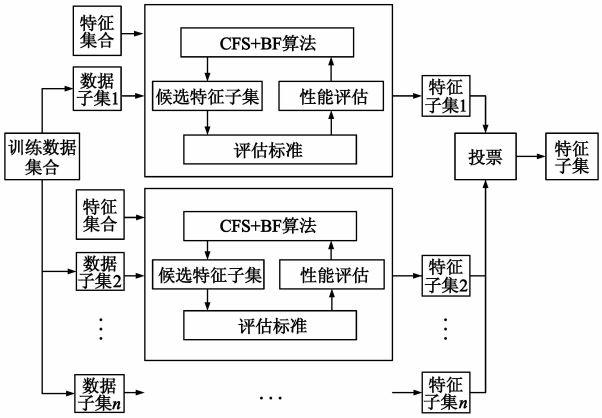


图1 基于投票的特征选择算法的框架

2.1 训练数据子集划分

DV 算法的第一步是将原始的数据集 T 分割为规模较小的数据集 $T_1,T_2,\cdots,T_n$.但随机的划分 T 会造成不同的数据子集 T_i 的类分布不均衡.为了消除这种不均衡性,我们采用分类抽样的方法获得 $T_1,T_2,\cdots,T_n$.

假设数据集 T 中有 m 种流量类,每种流量类的样本数为 $\Omega(C_i)$,划分为不相交的 n 个集合 $T_1,T_2,\cdots,T_n$.对每个类 C_i,每次在 T 中按不放回抽样随机抽取 $\lfloor\Omega(C_i)/n\rfloor$个样本放入子集 T_i.这样划分的好处是使得 $T_1,T_2,\cdots,T_n$ 的大小基本相同,且各种流量类型在上述子集中的比例也基本相同,避免特征提取过程中对某些流量类产生偏倚.算法的时间复杂度依赖于样本数,为线性阶 $O(n)$.具体的划分算法如算法 1.

算法 1 Divide_Dataset
输入:T 是已标注过的流量数据;n 是划分子集的数目;
输出:$T_1,T_2,\cdots,T_n$
(1)初始化每个数据子集为空集 $T_i\leftarrow\varnothing$
(2)统计每个流量类的样本数 $\Omega(C_i)$
(3)for 每个数据子集 T_i do
(4) for 每个流量类 C_j do
(5) 从 T 随机抽取$\lfloor\Omega(C_i)/n\rfloor$个样本放入 T_i
(6) end for
(7)end for
(8)return $T_1,T_2,\cdots,T_n$

2.2 基于分治投票策略的算法(DV)

尽管采用分类抽样的方法可均衡各流量类在不同子集中的分布,但各子集中提取的特征集合仍会存在一定差异.为此,通过投票策略来筛选出各特征子集上得票数最多的一些特征作为最终的特征子集.这种方法的优点是可以避免在特征提取过程中对某个训练集合的过拟合,提高所选取的特征子集的泛化能力.算法复杂度依赖于特征数与划分的子集数,为 $O(n^2)$.

算法 2 DV
输入:T 是已标注过的流量数据;A 是全部特征的集合;
输出:特征子集 $A_M{}^*$
(1)调用 Divide_Dataset(T,n),将训练集合 T 划分为 n 个子集 $T_1,T_2,\cdots,T_n$
(2)for 每个子集 $T_i\subset T$ do
(3) $A_i{}^*=\mathrm{CSF+BF}(T_i,A)$
(4)end for
(5)for 每个特征 $f_i\in A$ do
(6) for 每个特征子集 A_j^* do
(7) if $f_i\in A_j^*$ then
(8) $C_i\leftarrow C_i+1$//统计投票数
(9) end if
(10) end for
(11)end for
(12)sort(C_i) //根据投票数排序
(13)取前 $M\ll|A|$个特征为最终特征子集 A_M^*
(14)return A_M^*

3 实验评估

首先,我们选择目前已应用于流量分类的 Naïve Bayes[12]和决策树[13]算法作为分类性能的评估算法.其次,我们通过比较上述分类算法在原始特征空间、CFS + BF 算法获得的特征空间及 DV 算法获得的特征空间之间,在召回率、精度与 F-Measure 这三个分类性能指标上的差异来分析 DV 算法的性能.最后,通过特征递减的方法来进一步证明投票法的合理性.

本文采用 Moore 等提供的公开流量数据集[14],该数据集共有 248 个数据流特征[4],这些特征采用数字索引,如 4 表示特征为服务器端口,50 表示服务器重传给客服端的数据包数,具体特征含义参见文献[4].应用类型包括 WWW、P2P 等 12 种,共计 377526 条数据流.

3.1 实验结果分析

集合 T 为原始流量集合,利用 CFS + BF 算法获得的特征子集为 $A=\{4,50,72,91,109,113,137\}$.利用 DV 算法将 T 划分为 10 个数据子集 $T_1,T_2,\cdots,T_{10}$,在 10 个数据子集上分别采用 CFS + BF 算法提取特征,获得的特征子集如表 1 所示.DV 算法根据不同特征的得票数,最终确定特征集为 $A^*=\{4,137,113,78,72,91,148\}$.由此,我们可获得三个不同特征空间数据集合:$T$(248 个原始特征)、$TA$(7 个特征)、$TA^*$(7 个特征).

表 1　BF 算法提取到的特征子集,表内为特征的数字索引

T_1	4	50	72	91	108	113	155	202
T_2	4	71	72	81	149	152		
T_3	4	72	78	108	113			
T_4	4	78	109	137	148			
T_5	4	51	109	113	137	180		
T_6	4	49	78	91	137	203		
T_7	4	29	78	83	113	137		
T_8	4	66	78	113	137	148		
T_9	4	35	51	137	151	166		
T_{10}	4	91	110	113	137	148	182	

3.1.1 分类性能比较

我们通过比较 T、TA、TA^* 在 Naïve Bayes、决策树这两种分类方法下的准确率与时间效率,来进一步证明 DV 算法的可行性.我们选择 WWW、MAIL、P2P 及 FTP-DATA 这几种典型应用进行分析.

图 2 显示了三种特征子集下的 Naïve Bayes 方法分类准确性(真阳性率、精度和 F-Measure 这三种指标).对于 WWW,CFS + BF 特征选择后反而不如原始特征空间上的分类性能.而 DV 方法获得的特征子集使这三个指标均得到了有效提升(95% 以上),可见 DV 方法识别 WWW 的效果明显.对于 MAIL,CFS + BF 特征选择后也导致三种指标的大幅下降,而 DV 算法在召回率上有显著提升,精度有所下降,F-Measure 略低于原始数据.对于 P2P,CFS + BF 与 DV 算法特征选择后明显提升了分类准确率,DV 算法效果更为明显.FTP-DATA 与 MAIL 类似,特征选择后召回率提高,但精度下降.

Naïve Bayes 模型在原始 248 个特征下训练需要 35.88s,而用 DV 算法获得的 7 个特征训练仅需 1.16s,缩短了近 31 倍的时间.在特征提取效率上,CFS + BF 需要将近 10min,而 DV 算法只需 46.32s.由此可见,DV 算法在特征提取的时间效率上明显优于 CFS + BF,同时 DV 算法获得的特征空间在分类性能上优于 CFS + BF.

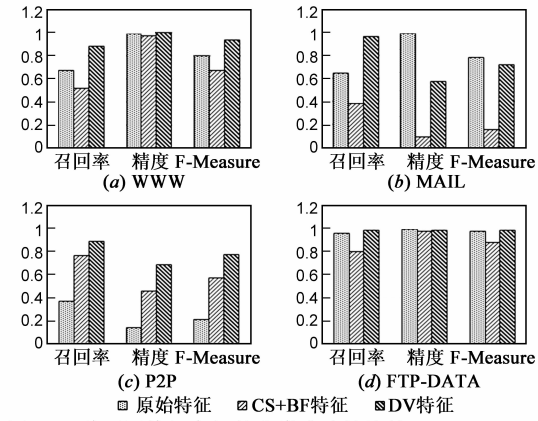


图2　三种不同特征空间的分类准确性比较(Naïve Bayes)

再通过决策树方法来验证 DV 算法的合理性与有效性.表 2,3 显示特征选择后对 WWW 与 MAIL 分类性能几乎没有影响,三个性能指标均接近 100%.从表 2 ~ 5 可以发现,不管是否采取特征提取,决策树均有非常优秀的分类性能,几乎接近 100%,误分类率极低.但特征选择的优势是使模型的训练时间缩短,复杂度降低.如果不进行特征提取,构建 248 个特征的决策树需要 283.13s,节点数多达 161 个,叶子结点多达 81 个.而采取 DV 算法获得的决策树,构建时间仅需 6.83s,节点只有 85 个,叶子结点 43 个.

表 2　WWW(%)

	召回	精度	F-Measure
原始特征	100	99.8	99.9
CFS + BF	99.9	99.8	99.8
DV	100	99.7	99.9

表 3　MAIL(%)

	召回	精度	F-Measure
原始特征	100	100	100
CFS + BF	100	100	100
DV	100	100	100

表 4　P2P(%)

	召回	精度	F-Measure
原始特征	97	96.4	96.7
CFS + BF	89.9	97.6	93.6
DV	88.6	97.7	93

表 5　FTP-DATA(%)

	召回	精度	F-Measure
原始特征	95.9	99.6	97.7
CFS + BF	99.9	99.9	99.9
DV	99.9	99.5	99.7

3.1.2 特征子集分析

按得票数由高到低,逐个裁减特征的方式来分析得票数与其区分能力的相关性,进而证明得票数的合

理性.先从特征子集中去除得票数最高的特征 4,进行分类测试;接着再去除得票数次高的特征 137,进行分类测试;如此递减测试.

图 3 显示了随着特征子集按得票数由高到低递减后的 Naïve Bayes 分类的性能比较情况.X 轴表示特征数逐一减少的过程.去掉得票数最多的特征 4 后,WWW 的假阳性率增大到 60%,而分类精度降低;之后的特征减少使这些指标的变化渐趋平缓.同样,P2P 对于去除特征 4 后的分类性能变化明显,召回率从 76.2% 下降到接近 0.从 MAIL 看,召回率与假阳性率呈递减趋势.FTP-DATA 也是在去除特征后分类性能变化最大.采用决策树分类方法进行同样的实验,实验结果与 Naïve Bayes 方法类似.由此,我们可以得出得票数越多的这些特征对分类性能的影响越大,也证明了利用得票数构建特征子集的合理性.

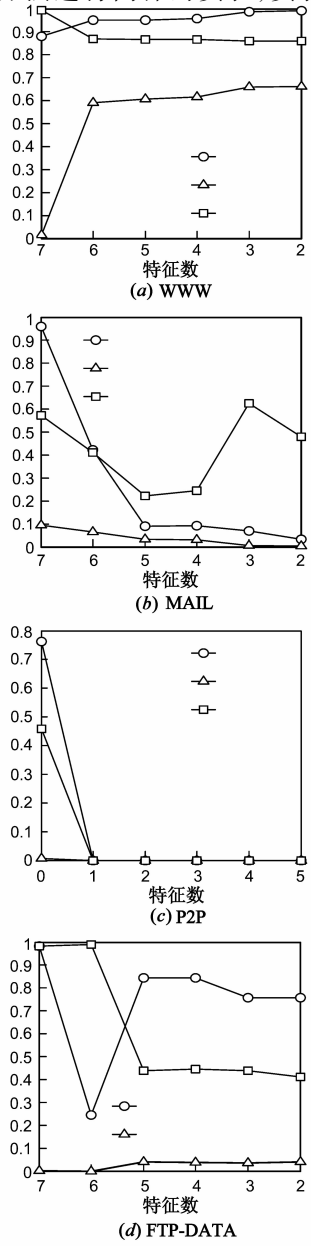


图3 特征子集按得票数递减后的分类性能比较(Naïve Baye)

4 结束语

互联网流量数据本身具有特征维度高,数据量大的特点.我们采用将训练集合分解,在规模变小的数据子集上采用特征提取算法,再通过投票方式获得最终特征集合.实验证明在分类效率上比直接在原始大数据集上提取特征具有更大的优势,并可利用目前处理器的并行性,大大缩短算法的运行时间.本文提出的方法具有简单有效的特点,可利用现有的各种特征提取方法,对于实际部署具有很好的实际意义.

参考文献

[1] Dainotti A, Pescapè A, Claffy K C. Issues and future directions in traffic classification[J]. Network, IEEE, 2012, 26(1): 35 – 40.

[2] Yaguan Q, Chunming W, Qiang Y, et al. Network traffic anomaly detection based on maximum entropy model[J]. Chinese Journal of Electronics, 2012, 21(3): 579 – 582.

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

[4] Moore A, Zuev D, Crogan M. Discriminators for Use in Flow-Based Classification[R]. London: Queen Mary and Westfield College, Department of Computer Science, 2005.

[5] Jain A, Zongker D. Feature selection: Evaluation, application, and small sample performance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(2): 153 – 158.

[6] Hall M A. Correlation-based feature selection for machine learning[D]. Waikato, New Zealand: The University of Waikato, 1999.

[7] E Rich, K Knight. Artificial Intelligence[M]. New York, US: McGraw-Hill, 1991.

[8] Sun M, Chen J, Zhang Y, et al. A new method of feature selection for flow classification[J]. Physics Procedia, 2012, 24: 1729 – 1736.

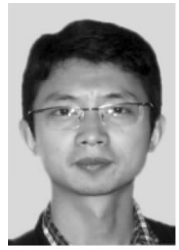
[9] 宁卓,孙知信,龚俭,等.利用流量特征的 GIDS 报文分类优化算法[J].电子学报,2012,40(3):530 – 537.
NING Zhuo, SUN Zhi-xin, GONG Jian, ZHANG Wei-wei. An improved GIDS packet classification algorithm using the characteristic of the traffic[J]. Acta Electronica Sinica, 2012, 40(3): 530 – 537. (in Chinese)

[10] Jamil H A, Zarei R, Fadlelssied N O, et al. Analysis of features selection for P2P traffic detection using support vector machine[A]. Information and Communication Technology (I-

CoICT)[C]. Bandung: IEEE, 2013. 116 – 121.
[11] Alazab A, Hobbs M, Abawajy J, et al. Using feature selection for intrusion detection system[A]. Communications and Information Technologies (ISCIT)[C]. Gold Coast, QLD: IEEE, 2012. 296 – 301.
[12] Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques[A]. ACM SIGMETRICS Performance Evaluation Review[C]. Banff, Alberta, Canada: ACM, 2005. 50 – 60.
[13] Zhang Y, Wang H, Cheng S. A method for real-time peer-to-peer traffic classification based on C4.5[A]. 2010 12th IEEE International Conference on Communication Technology (ICCT)[C]. Nanjing: IEEE, 2010. 1192 – 1195.
[14] Andrew W Moore. Dataset[OL]. http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/index.html, 2013.8.

作者简介

高　文　男,1987 年生,广东广州人. 2009 年毕业于上海交通大学,获学士学位. 现为浙江大学计算机学院博士研究生,主要研究方向为软件定义网络、可重构网络、网络仿真平台.
E-mail: gavingao@zju.edu.cn



钱亚冠(通信作者)　男,1976 年生,浙江嵊州人. 2005 年毕业于浙江大学计算机学院,获工学硕士学位. 现为浙江大学博士研究生. 主要研究方向为互联网流量分类、下一代互联网、机器学习与数据挖掘.

吴春明　男,1967 年生,浙江萧山人,博士. 现为浙江大学计算机学院教授、博士生导师,主要研究方向为未来互联网络、可重构网络技术与网络虚拟化、网络服务质量.
E-mail: wcm@zju.edu.cn

朱　凯　男,1989 年生,山东枣庄人. 2011 年毕业于山东大学,获学士学位. 现为浙江大学计算机科学与技术学院博士研究生,主要研究方向为软件定义网络、可重构网络、网络虚拟化计算.

陈双喜　男,1980 年生,安徽安庆人,硕士,现为嘉兴职业技术学院信息技术分院讲师,浙江大学计算机系统结构与网络安全研究所科研聘岗教师,主要研究方向:未来网络体系架构、网络内容安全、机器学习、分布式计算.