



基于 DT-KSVM 的业务感知算法

董昊¹, 曲桦¹, 赵季红^{2,3}, 陈梁骏², 戴慧珺²

(1. 西安交通大学软件学院 西安 710049; 2. 西安交通大学电子与信息工程学院 西安 710049;
3. 西安邮电大学通信与信息工程学院 西安 710061)

摘要:提出一种新的基于 DT-KSVM(decision tree kernel support vector machine, 决策树支持向量机)的业务感知算法, 利用 ReliefF 特征选择算法提取特征, 提出样本间类别可分度计算方法排序不同业务感知难度, 优先感知易分业务。在实际网络业务数据集上与传统一对一(one-versus-one)SVM 感知方法进行对比, 结果表明该方法具有较高的业务识别准确率和更好的时间性能。

关键词: SVM; 决策树; 业务感知

doi: 10.11959/j.issn.1000-0801.2015227

Service-Aware Algorithm Based on DT-KSVM

Dong Hao¹, Qu Hua¹, Zhao Jihong^{2,3}, Chen Liangjun², Dai Huijun²

(1. School of Software, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Electronic & Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 3. School of Telecommunication & Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an 710061, China)

Abstract: A novel service-aware method based on decision tree kernel support vector machine (DT-KSVM) algorithm was proposed. A service-aware model was developed by using ReliefF algorithm to extract service characteristics, and proposing the separable degree between samples to simply the service-aware process. Through experiment comparison between this proposed model and traditional one-versus-one SVM method, it is shown that the proposed method has a better service-aware accuracy and time performance.

Key words: support vector machine, decision tree, service-aware

1 引言

日渐增长的网络规模和用户需求带给互联网越来越多的挑战, 其中 P2P 等网络共享业务更给网络造成了巨大的压力, 同时也降低了用户上网体验。为了保障高质量业

务的 QoS(quantity of service, 服务质量), 必须对业务加以区分^[1], 以此来对不同的网络业务进行差分服务。对网络中业务内容的区分可认为是对网络业务的感知过程, 此过程对管理网络性能、QoS 保障、准入控制等有很重要的意义。通过对业务的感知来对网络的各种资源进行合理的分配,

收稿日期: 2015-06-28; 修回日期: 2015-09-07

基金项目: 国家自然科学基金资助项目(No.61371087), 国家高技术研究发展计划(“863”计划)基金资助项目(No.2015AA015702)

Foundation Items: The National Natural Science Foundation of China(No.61371087), The National High Technology Research and Development Program of China(863 Program) (No.2015AA015702)

论文引用格式: 董昊, 曲桦, 赵季红等. 基于 DT-KSVM 的业务感知算法. 电信科学, 2015227

Dong H, Qu H, Zhao J H, et al. Service-aware algorithm based on DT-KSVM. Telecommunications Science, 2015227



实现异构网络的协同、网络负载的均衡,从而优化网络性能^[2]。这是当前研究的热点内容,也是运营商设计更好系统的依据和保障 QoS 的关键。

本文针对网络业务感知,提出了融合决策树算法与核 SVM(support vector machine,支持向量机)方法的 DT-KSVM 算法。结合决策树算法描述简单、分类速度快的特点以及 KSVM 算法基于结构风险最小原则,平面与核函数的优点,能有效地将难以解决的原始空间样本的非线性运算映射为线性空间的线性运算。通过对英国剑桥大学网络数据集的感知仿真,证实 DT-KSVM 算法具有感知速度快、识别率高、顽健性强的特点。

2 相关工作

目前,业务感知方法主要有:端口识别法、DPI(deep packet identify,深度分组检测)法^[3]、基于流量统计的流量识别方法和基于机器学习的识别方法。端口识别法效率低下,对网络中的大多数业务都无法感知;DPI 感知算法感知精度高,但对加密业务无法很好处理;基于流量统计的业务感知方法目前的感知精度不高;基于机器学习的方法^[4]利用业务流中存在的统计信息数据感知网络业务,但时间复杂度较高。

将 SVM 用于业务感知中,主要优势在于^[5]:首先是 SVM 利用高维映射,将非线性问题向线性问题转化,且在分类过程中不需要知道具体的业务特征;其次,SVM 的分类面个数取决于支持向量机数目,可以和其他算法进行结合来改变支持向量的数目,从而对分类面的数量进行调整,即它的时间复杂度可根据情况调整,泛化性好,顽健性高;最后,在求解过程中,SVM 将问题转化为二次规划类型的问题,得到的最优解是全局的,较其他算法分类准确率高很多。

现有的基于 SVM 的感知方法主要有一对一(one-versus-one)分类方法^[6]、一对多(one-versus-rest)分类方法^[7]以及有向无环图(directed acyclic graph)SVM 分类方法^[7]等。其中一对一方法是最早出现也是应用较多的方法,在本文仿真部分中作为对比方法。其原理是在每两个类别之间都构造分类面,对样本进行分类判断时,需要将样本和每个分类平面进行判断,最后判为哪个类别的次数较多,该样本就属于哪类。对于有 k 个类别的样本,该方法需要构造的分类面数量为 $k(k-1)/2$ 。

由目前的研究现状可知,SVM 具有较好的推广性能和较好的分类准确率,本文在前人研究的基础上,提出

DT-KSVM 方法,结合决策树思想,通过核化的 SVM 方法对网络业务进行感知^[8]。

3 基于 DT-KSVM 的业务感知算法

DT-KSVM 算法构造了一个决策树的结构,首先对样本进行特征提取,计算样本间类别可分度,再对其中较易区分的样本先进行分类,最后按照所构造的树型结构感知出各个样本类别。

3.1 特征提取

在数据集中存在一些对感知没有影响甚至影响感知准确率率的属性,不适合用来作为感知的特征依据,需要将这些属性剔除。具体方法是对样本中的各个属性按照计算出来的属性权重平均值,设置阈值,将权重小于阈值的属性去除,剩下对感知有重大贡献的属性特征。针对网络业务样本数据,本文采用 ReliefF^[9]算法进行特征提取,该方法运行效率高,对数据类型没有限制,可处理多类别问题。

3.2 定义样本间类别可分度

特征提取完成后,为了解决决策树结构中的误差累积问题,本文提出样本间类别可分度概念,通过对样本类别间的分类难易程度进行量化,衡量样本类别分类难易程度。下面介绍本文中定义的样本间类别可分度。

在计算可分度之前,先计算样本的分离系数,本文中用样本的中心距离和样本的方差以及类别间的欧几里得距离(Eudidean distance)一起,作为样本分离系数的计算指标。单用欧几里得距离计算得到类别中心间的距离,并不能完全体现样本的可分程度。因此,计算欧几里得距离的同时引入权重,权重是计算样本的类别中心和方差的比值,方差能很好地体现样本的分布情况。样本越紧凑,方差则相对较小,相反,方差越大,则样本越离散。通过将类间中心距离和类的分布综合考虑进行类间分离系数的计算,准确率会提高很多。分离系数 ϕ_j 的计算如式(1)所示:

$$\phi_j = \omega_i \|d_i - d_j\| \quad (1)$$

其中, d_i 和 d_j 分别是第 i 类和第 j 类样本的类别中心,计算式如式(2)所示, ω_i 为第 i 类样本的欧几里得权重,它的计算方法如式(3)所示:

$$d_i = \frac{1}{n_i} \sum_{x \in X_i} x, i=1,2,\dots,l \quad (2)$$

$$\omega_i = \frac{d_i}{\phi_i} \quad (3)$$

其中, n_i 表示 X_i 的样本数量, X_i 为样本所属种类, l 表

示样本的类别数。 ϕ_i 表示第*i*类样本的方差,体现样本分布。其计算方法如式(4)所示:

$$\phi_i = \frac{1}{n_i - 1} \sum_{x \in X_i} \|x_i - d_i\|, i=1, 2, \dots, l \quad (4)$$

其中, x_i 表示第*i*类样本的属性值。将式(2)~式(4)代入式(1)中,即可求得样本间类别分离系数 ϕ_{ij} 。 ϕ_{ij} 的值越大,说明样本间的分离度越高,样本越易被分开。计算出各个样本的类间分离系数 ϕ_{ij} 后,统计各个样本类别的数量 n_i ,最终定义类间的可分度表达式如下:

$$D_i = n_i \sum_{j=1, j \neq i}^k \phi_{ij} n_j \sum_{j=1, j \neq i}^k \omega_j \|d_i - d_j\| \quad (5)$$

其中, k 为样本的类别数目,可以看出样本的可分度就是样本与和与它不同类的样本分离系数之和再和该类别样本的数量做乘积运算所得的结果。 D_i 是第*i*类样本的可分度。

3.3 DT-KSVM 算法

DT-KSVM 算法通过将决策树结构和 KSVM 算法结合,进行网络业务的感知,它是一种类似平衡二叉树的树型结构形态。其中,每个树节点上的样本类别和样本的可分度大小相关联。中间节点上至少有两个样本种类,叶子节点便是最终分出来的样本种类。DT-KSVM 算法的流程如图 1 所示。

经过自上而下的感知,到叶子节点时各个业务类别就能被很好地感知出来。本文通过对网络业务的特征提取、样本类别间可分度的计算,构建 DT-KSVM 算法训练模型

对样本进行感知。

算法的具体步骤如下。

步骤 1 通过 ReliefF 算法^[9]对样本的重要属性进行提取,去掉对样本区分度不大的属性。

步骤 2 特征选取后,根据式(5)计算每个样本类别的可分度,计算出每类样本的可分度,按照可分度从大到小进行排序,若遇到可分度相同的情况,则统计样本的数量,数量多的类别排在前面。

步骤 3 根据可分度排序结果,将排在前 1/2 的样本类别作为正类,排在后面的类别作为负类。

步骤 4 使用 DT-KSVM 算法构造分类平面,优化参数,并训练。

步骤 5 通过对 DT-KSVM 算法的训练,则样本的正类和负类作为决策树的两个子节点,由于样本类别较多,第一次分类决策后,每个子节点中可能含有两个以上的类别,若要将子类别分开,则转到步骤 3,对各个子节点的样本类别同样一半取正类,一半取负类,进入步骤 4 训练,进行再次决策。

步骤 6 不断执行上面的过程,直到节点中只含有一个样本类别时为止,即到达叶子节点,此时所有的样本类别将被分开。

4 仿真与分析

4.1 数据处理

本文数据是采用英国剑桥大学的网络数据集。该大学

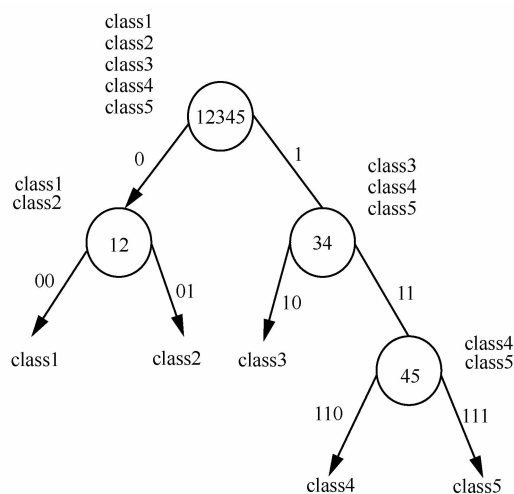
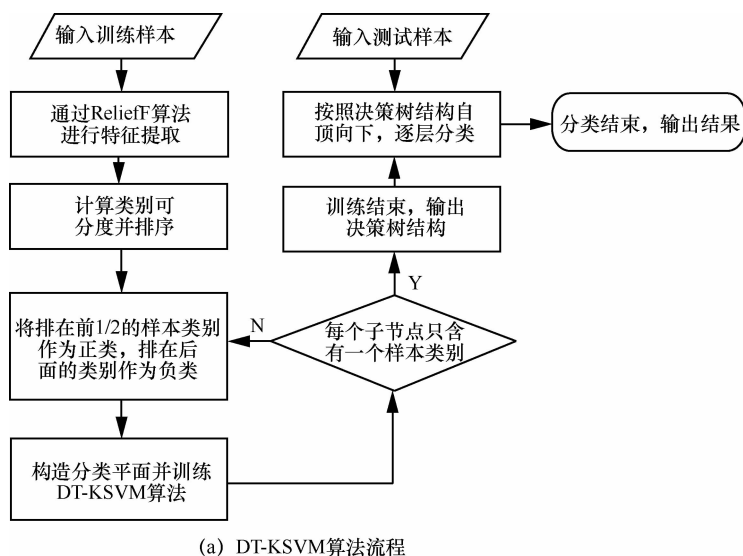


图 1 DT-KSVM 算法流程



表 1 P2P 业务可分度

排序前			排序后		
序号	类别	可分度	序号	类别	可分度
1	Gnutella	4.911 687 8E12	1	Thunder	1.326 261 36E13
2	eDonkey	7.976 978 5E12	2	eDonkey	7.976 978 5E12
3	Thunder	1.326 261 36E13	3	BitTorrent	7.700 031 2E12
4	KaZaA	5.171 715 2E12	4	KaZaA	5.171 715 2E12
5	BitTorrent	7.700 031 2E12	5	Gnutella	4.911 687 8E12

的 Moore 等人的数据集收集了该学校的双工网络流量。每个文件含半小时统计的流,每个流由一行记录组,该数据集特征全面完整,有一定代表性^[10]。Moore 数据集中业务的属性是通过流的属性表现出来,其中统计了其 249 个属性值。对采集来的原始数据进行预处理,包括对数据集进行抽样、数据规整、数据归一化等过程。对归一化后的数据集采用 SVM 算法分类出样本集中的 P2P 业务,用于 DT-KSVM 算法的仿真。本文实验数据中的 P2P 业务主要有以下 5 个类别,分别为:Gnutella、eDonkey、Thunder、KaZaA 和 BitTorrent。

通过特征提取及样本间可分度计算,表 1 中列出各 P2P 业务的可分度值。

在构造各分类平面时需要 SVM 核函数中的参数和惩罚因子进行优化调整。在选取参数时采用交叉验证来建模,网格搜索来寻优。本文采集的数据集中,P2P 业务有 5 个类别,根据 DT-KSVM 算法可知,该样本集需要 $N-1$ 即 4 个分类平面就可以将所有的类别分开。表 2 列出了各个分类平面所对应的数据集以及各个分类平面的参数选择结果。其中 Originals.libsvm 中为所有的数据集,Original0s.libsvm 中在没分错的情况下应该包含 Thunder 和 eDonkey 的样本集,Original1s.libsvm 中应为 BitTorrent、KaZaA 和 Gnutella 的样本集,Original11s.libsvm 中为 KaZaA 和 Gnutella 的样本集。

表 2 各分类平面对应的数据集及最优参数

分类面	数据集	C	g
SVM1	Originals.libsvm	32 768	0.125
SVM2	Original0s.libsvm	8 192	0.031 25
SVM3	Original1s.libsvm	8 192	0.5
SVM4	Original11s.libsvm	2	0.5

4.2 仿真结果

本文通过 DT-KSVM 算法对提取出的 P2P 业务进行感知,感知结果见表 3。

表 3 P2P 业务识别结果

业务类型	识别数目	实际数目	识别数/实际数	召回率
Gnutella	371	390	95.1%	94.6%(369/390)
eDonkey	414	430	96.2%	95.8%(412/430)
Thunder	387	400	96.7%	96.7%(387/400)
KaZaA	440	420	104.7%	95.2%(400/420)
BitTorrent	482	454	106.1%	96.9%(440/454)

表 3 的识别结果是对本次采集的数据进行感知的结果,从结果中可以看出,依然存在少量的错分误分的现象,但该算法基本能较准确地感知出样本中不同的 P2P 业务,另外,除 Gnutella 外各个类别的召回率基本可达到 95%以上,其中召回率指算法正确决策的样本数和该类样本总数之比。将 DT-KSVM 算法与传统的 1vs1 多分类感知算法相比较,其感知准确率见表 4。

表 4 多分类感知算法结果对比

算法类别	准确率
1vs1	93.97%
DT-KSVM	96.08%

从表 4 中可知,DT-KSVM 算法比 1vs1 算法的准确率高很多,其准确率是指算法正确决策的样本数和样本总数之比。这是因为 DT-KSVM 算法比 1vs1 算法多了特征提取和类间可分度计算,另外,DT-KSVM 所构造的分类平面较 1vs1 算法少,因此,算法在时间性能方面也较 1vs1 算法要好。

图 2 为在抽取不同的样本数量时,1vs1 和 DT-KSVM 对样本训练所花费时间的对比。

从图 2 中可看出,样本数量越多,1vs1 算法与 DT-KSVM 算法的样本训练时间差距越大。这是由于 1vs1 算法需要构造更多的分类面,因此其花费的时间较长,DT-KSVM 算法较 1vs1 算法构造的分类面要少一些,因此时间性能更好。

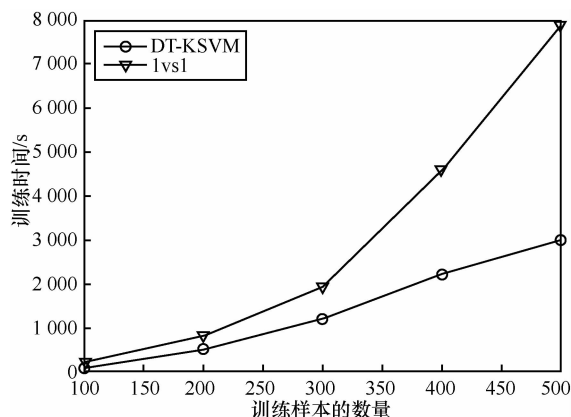


图2 样本训练时间对比

5 结束语

本文主要通过和传统的多分类算法相比较,提出 DT-KSVM 方法来感知网络业务。在感知过程中采用 ReliefF 算法对网络业务的主要特征进行提取,通过本文中的类别可分度来衡量样本的易分程度,并对样本进行排序。按照可分度从大到小排序结果,运行 DT-KSVM 算法对网络业务进行感知。最后,通过仿真,与 1vs1 算法进行对比,可知 DT-KSVM 算法的感知准确率和时间性能均较好,是有效的网络业务感知算法。这对网络性能的管理、服务质量的提升、网络资源的分配等都有重要意义。

参考文献

- 1 Kawahara R, Kamei S, Kamiyama N, *et al.* A Method of Constructing QoS Overlay Network and Its Evaluation. Technical Report of IEICE Cq, 2009
- 2 Lu X M, Pei J, Zhou Y J. A new P2P traffic identification model based on node status. Proceedings of the 2010 International Conference on Management and Service Science (MASS 2010), Wuhan, China, 2010:1~4
- 3 Song H, Han I Y, Kim Y, *et al.* The NADPH oxidase inhibitor DPI can abolish hypoxia-induced apoptosis of human kidney proximal tubular epithelial cells through Bcl2 up-regulation via ERK activation without ROS reduction. Life Sciences, 2015(126): 69~75
- 4 Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques. Amstel dam: Elsevier, 2011
- 5 Yuan R X, Li Z, Guan X H, *et al.* An SVM-based machine learning method for accurate internet traffic classification. Information Systems Frontiers, 2010, 12(2):149~156
- 6 Rocha A, Goldenstein S K. Multiclass from binary: expanding

one-versus-all, one-versus-one and ecoc-based approaches. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(2): 289~302

- 7 Ng S S Y, Tse P W, Tsui K L. A one-versus-all class binarization strategy for bearing diagnostics of concurrent defects. Sensors, 2014, 14(1): 1295~1321
- 8 戴慧珺. 基于业务分类识别流量预测的覆盖网 QoS 路由算法(博士学位论文). 西安: 西安交通大学, 2015
Dai H J. Research on QoS routing based on traffic identification and flow prediction in overlay network(doctor dissertation), Xi'an: Xi'an Jiaotong University, 2015
- 9 Zhou Q, Ding J, Ning Y, *et al.* Stable feature selection with ensembles of multi-reliefF. Proceedings of the 10th International Conference Natural Computation (ICNC 2014), Xiamen, China, 2014: 742~747
- 10 Moore A W, Zuev D. Internet traffic classification using bayesian analysis techniques. Proceedings of 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, USA, 2005: 50~60

[作者简介]



董昊,男,西安交通大学硕士生,主要研究方向为深度学习与网络流量分类。

曲桦,男,博士,西安交通大学教授、博士生导师,信息与通信技术研究院院长、未来网络与新业务研究中心主任,主要研究方向为宽带无线移动通信网、未来网络、无线接入技术、移动互联网技术、智能光网络技术、IP 网络技术、泛在网技术与应用、网络的管理与控制技术、网络的异构和融合技术、自适应信号处理与自适应网络技术。

赵季红,女,博士,西安交通大学教授、博士生导师,赴美高级访问学者,ITU 中国组专家,下一代网络研究中心主任,主要研究方向为宽带通信网络、移动通信、未来网络、网络的管理与控制 and 光通信网络。

陈梁骏,男,西安交通大学博士生,主要研究方向为网络流量分类、深度学习和信息理论学习。

戴慧珺,女,博士,西安交通大学电子与信息工程学院工程师,主要研究方向为业务分类与流量预测。