一种基于改进隐马尔可夫的多媒体业务分类算法

王再见*①② 董育宁① 张 晖① 冯友宏②
①(南京邮电大学通信与信息工程学院 南京 210003)
②(安徽师范大学物理与电子信息学院 芜湖 241000)

摘 要：该文提出一种基于改进隐马尔可夫(Hidden Markov Model, HMM)的多媒体业务分类算法。改进后的算法保持典型 HMM 模型结构不变，通过区分包大小的位置信息，改变发射概率取值，提高了多媒体业务区分性能。理论分析表明，该文模型在计算量上低于高阶 HMM；实验结果表明，改进的 HMM 多媒体业务分类算法的区分效果优于现有的 HMM 多媒体业务分类方法。

关键词：多媒体业务流识别；隐马尔可夫；发射概率；包大小
中图分类号： TP393 **文献标识码**： A **文章编号**：1009-5896(2015)02-0499-05
DOI: 10.11999/JEIT140340

A Multimedia Traffic Classification Method Based on Improved Hidden Markov Model

Wang Zai-jian①② Dong Yu-ning① Zhang Hui① Feng You-hong②
①(*College of Communications and Information Engineering*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)
②(*The College of Physics and Electronic Information*, *Anhui Normal University*, *Wuhu* 241000, *China*)

Abstract: This paper proposes an improved Hidden Markov Model (HMM) based multimedia traffic classification method. This method preserves the classical HMM model structure, and improves the performance of multimedia traffic classification by changing the emitting probability value with the position information of packet size. Theoretical analysis indicates that the new model can reduce the computational complexity of the classical HMM model. Simulation results show that the proposed method can improve the classification performance compared with the existing HMM based classification method.

Key words: Multimedia traffic classification; Hidden Markov Model (HMM); Emitting probability; Packet size

1 引言

端到端 QoS(Quality of Service)保证领域的网络多媒体业务的分类，关注的是 QoS 特征而不是业务具体内容，希望用较少的 QoS 特征区分业务所归属的 QoS/业务类。典型的基于端口或深度包检测的业务分类方法在效率、安全等方面存在不足。一些机器学习方法由于分类的复杂度较高，影响了分类实时性。因此，近年来有大量工作，研究如何基于可观察的 QoS 特征状态，准确、高效地识别被隐藏的 QoS/业务类别[1]。

HMM(Hidden Markov Model)近年来取得较多的研究成果。文献[2–5]都基于典型 HMM 分类方法识别网络业务流，规避了传统识别/分类方法存在的法律问题等不足，但识别效果有待进一步提高，部分原因是由于 HMM 中对隐式序列的一阶假定有很大的局限性，发射概率是随机的。但事实上，网络多媒体业务中用于区分的特征，其所构成的观测序列不仅依赖于隐藏序列现在的状态，与过去的状态也有关，且常常存在高阶依赖关系。业务区分的重要特征包大小被很多研究人员用于业务识别和分类，但其取值并不是随机的，有一定的规律可循。如果放宽 HMM 的经典假设，将 HMM 直接拓展为高阶 HMM 则会引入相当大的计算量。因此，如何在不改变典型 HMM 结构的前提下，结合多媒体业务 QoS 特征，既不降低实时性要求，又提高识别准确性，是 HMM 模型在多媒体业务识别应用中面临的挑战。

本文的主要贡献如下：(1)通过分析典型网络多

2014-03-13 收到，2014-09-15 改回
国家自然科学基金(61401004, 61271233, 60972038, 61101105)，教育部高等学校博士学科点专项科研基金(20103223110001, 20113223120001)，工业与信息化部通信软科学课题(2011-R-70), 2011 年度江苏省研究生培养创新工程(CXZZ11_0396)和安徽师范大学科研培育基金(2013xmpy10)资助课题
*通信作者：王再见 wangzaijian@ustc.edu

媒体业务包大小特征，得出了与前人不同的新发现，发现了一组易于提取、计算复杂度低、处理速度快、可有效区分业务所属 QoS/业务类别的 QoS 特征；(2)根据这些新发现，通过融合特征上下文信息，调整发射概率，改进典型 HMM 模型，设计了一种基于改进 HMM 的多媒体业务分类算法。

论文具体安排如下：第 2 节分析典型网络多媒体业务的包大小特征，第 3 节介绍改进的模型，第 4 节描述了基于改进 HMM 的分类算法，第 5 节给出仿真结果，最后是结论。

2 典型网络多媒体业务的包大小特征分析

图 1 是实时捕获的土豆视频业务抓包截图，由图可见，当前包大小值与其在序列中位置有以下特点：(1)当某个包大小值首次出现时，其后包大小值以较大的概率与之相等或相近；(2)传输过程中，在一定时间内，包大小取值相同或相近；(3)当前包大小值与相邻的前一个包大小有一定的联系。这是由于协议、拥塞、丢包、排序、优先级、重传等因素的存在，使得包大小值具有阶段性的特点。

本文使用 Wireshark[6]在实验室公用网络收集即时通信、标清流媒体(592×252)、高清流媒体(768×326)和游戏 4 种较流行的多媒体业务。依据标清流媒体业务包大小累积分布曲线，本文将包大小值划为如下 4 个等级。4 种业务包大小值在 4 个等级之间的转化频率并不一致(见表 1)。与标清流媒体类似，其它 3 种业务包大小值转换频率也有变化，即当前包大小值和前一个包大小值有内在联系。限于篇幅，此处不再一一列出。

表 1 标清流媒体业务包大小值转化频率(%)

当前包等级	前一个包等级			
	1	2	3	4
1	78.62	40.63	66.67	2.51
2	7.59	9.38	6.67	5.33
3	8.97	0	0	0.63
4	4.83	50	26.67	91.54

3 改进的 HMM 模型

3.1 改进的模型

定义 1 $Q=\{q_1,q_2,\cdots,q_N\}$ 代表网络操作的有限状态集合，其中的状态数为 N，Q_t 表示在 t 时刻的状态。

定义 2 $V=\{v_1,v_2,\cdots,v_M\}$ 代表离散化情况下 QoS 特征取值的集合，其中 M 为 QoS 特征值总数。

将网络操作状态的集合 Q 看作 HMM 模型的隐藏状态集合，$\boldsymbol{O}=(O_1=o_1,\cdots,O_M=o_M)$ 是输出的可观测状态，M 是序列中观测值数目。网络操作状态相互转移的概率矩阵为 $\boldsymbol{A}=\{\alpha_{ij}\},1\le i,j\le N$ (N 为状态数目)，$\alpha_{ij}=p(q_j/q_i)$ 表示从状态 q_i 转移到状态 q_j 的概率，$\sum_{j=1}^{N}\alpha_{ij}=1,1\le i\le N$ 。$B=\{b_{im}\}$ 表示某个时刻因网络操作状态而得到相应 QoS 特征输出值的概率，$b_{im}=P(v_m/q_i)$ 描述在给定状态 q_i 输出 QoS 特征取值 v_m 的概率。随机产生的 $\boldsymbol{\pi}=\{\pi_i,i=1,\cdots,N\}$，$\sum_{i=1}^{N}\pi_i=1$ 为网络操作状态初始概率分布，那么由网络操作状态和 QoS 特征状态所构建的模型，形成 HMM 模型。

本文通过一个修改后的发射概率函数将包大小取值概率与其相邻的前一个包大小取值结合起来，调整发射概率，将混淆矩阵修改为 $\boldsymbol{B}=\{b_{im}(o_{t-1})\}$，$1\le i\le N,1\le m\le M$ ，其中 o_{t-1} 表示上一时刻输出的观察状态，$b_{im}(o_{t-1})=P(v_m/q_i,o_{t-1})$ 描述在 t 时刻时，给定的上一个观察特征状态为 o_{t-1} 和隐藏状态为 q_i 情况下，观察特征取值为 v_m 的概率。依据特征前后状态间依赖关系，通过设计的发射概率函数确定当前状态的发射概率 $b_{im}(o_{t-1})$，之后利用转移概率矩阵计算下一时刻的系统隐式状态的取值，转移概

Source	Destination	Protocol	Length	Info
10.10.145.89	224.0.0.22	IGMPv3	60	Membership Report / Leave group 224.0.0.252
WistronI_32:21:03	Broadcast	ARP	60	Who has 10.10.145.20? Tell 10.10.145.89
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
10.10.145.5	122.227.237.14	TCP	54	cisco-ipsla > http [ACK] Seq=1 Ack=209793 Win=65535 Len=0
fe80::ad24:e311:2aeff02::16		ICMPv6	90	Multicast Listener Report Message v2
10.10.145.89	224.0.0.22	IGMPv3	60	Membership Report / Join group 224.0.0.252 for any sources
fe80::ad24:e311:2aeff02::1:3		LLMNR	87	Standard query 0xea93 ANY zhuy-PC
10.10.145.89	224.0.0.252	LLMNR	67	Standard query 0xea93 ANY zhuy-PC
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
10.10.145.5	122.227.237.14	TCP	54	cisco-ipsla > http [ACK] Seq=1 Ack=214017 Win=65535 Len=0
fe80::ad24:e311:2aeff02::1:3		LLMNR	87	Standard query 0xea93 ANY zhuy-PC
10.10.145.89	224.0.0.252	LLMNR	67	Standard query 0xea93 ANY zhuy-PC
10.10.145.56	234.34.23.234	UDP	62	Source port: awg-proxy Destination port: 33674
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic
122.227.237.14	10.10.145.5	HTTP	1462	Continuation or non-HTTP traffic

图 1 实时捕获的抓包截图

率矩阵和时间无关，保证了典型 HMM 模型的结构不变，提高识别准确度。相比其它模型，本文模型既简化了 QoS/业务类区分隐式过程存在的高阶依赖关系，又由于发射概率源于历史信息，确保了准确性。此外，本文模型保留了更多训练数据的统计信息，有利于解决训练和分类上的困难。

3.2 改进的发射概率

由于 HMM 的参数主要包括转移概率和发射概率，包大小在状态中的位置信息只能通过参数体现出来，由于与包大小有关，因此只能通过包大小发射概率体现出来。为此，本文参考文献[7]根据包大小所处位置的不同，采用不同的发射概率计算方法。具体说来，就是根据包大小在状态中的位置的不同以及前一个包大小的不同，选择不同的发射概率，公式为

$$b_{im}(o_{t-1})=\begin{cases}\dfrac{c(P(q\nwarrow\sigma))}{\sum\limits_{\rho\in\Sigma}c(P(q\nwarrow\rho))}, & \sigma\neq\sigma_{-1}\\ \dfrac{c(P(q\uparrow\sigma_{-1}\sigma))}{c(P(q\uparrow\sigma_{-1}))}, & \sigma=\sigma_{-1}\end{cases}\tag{1}$$

式中 σ_{-1} 表示 σ 的前一个包大小特征。$c(x)$ 表示事件 x 在训练集中出现的次数，为了计算 $c(x)$，需要在模型合并过程中保存状态的一些必要信息，包括状态中各个特征值出现的频率以及顺序关系。

4 基于改进 HMM 的分类算法

基于易获取、区分效果好、复杂度低的考虑，本文选取包大小作为区分特征，针对不同类型的 QoS/业务类网络多媒体业务，训练相应的基于改进 HMM 的分类器。基于获得的新分类器，对未知网络多媒体业务流进行分类，区分依据是相应分类器产生的概率。这样只需调整发射概率，就可通过一个 HMM 模型，简单地表述、解决复杂的网络多媒体业务流分类问题。上述过程的详细描述如下：

(1)数据预处理：获取包大小特征及位置信息，采用K-means聚类算法聚类。依据聚类中心，生成训练和识别的特征序列；

(2)改进的 HMM：根据数据库中包含的业务类型数目设置 N，观测值数目为 M；随机产生初始概率向量 $\boldsymbol{\pi}$，A 中的每个元素用业务在数据库中分布概率代表各个状态转移出现的概率，B 采用改进的观测值发生概率(见式(1))，初始化该类型业务的 HMM 为 λ；

(3)采用 Baum-Welch 算法[8]求解 HMM 参数 λ；

(4)判决输出：计算各业务在每个模型下的产生概率，由最优判决输出结果。

不同于基于典型 HMM 模型的多媒体业务分类方法，本文通过改进发射概率，简化高阶依赖关系降低实现难度，基于历史信息提高准确性，同时实现保持典型 HMM 模型结构不变，不增加训练和识别阶段的计算复杂度，这是由于本文不涉及联合概率、条件概率等概率公式的求解，也不涉及特征高阶量的计算，其计算复杂度远小于高阶 HMM。此外，与典型 HMM 模型相比，本文模型通过改进发射概率随机取值的不足，较好地考虑了多媒体业务 QoS 特征的内在规律，有利于提高区分效果。

5 实验仿真

本文使用 Wireshark 从实际网络中捕获 4 类流行的多媒体业务，作为样本流，通过人工识别，将流分为训练样本和测试样本(表 2)，用于评估算法性能，通过与文献[2–5]中相关算法比较，验证本文方法的有效性。

表 2 样本统计信息

	训练集		测试集	
	流(条)	字节(M)	流(条)	字节(M)
标清类流媒体	1892	593	952	369
高清类流媒体	1486	857	938	529
即时通信类	497	375	169	126
游戏类	1525	477	938	247
其它	948	394	872	369

通过多次实验，使用记录完整的分组信息，并保存在流文件中，下面取 2 组不同时间段进行的典型实验数据进行分析，其中“其它”包括 HTTP 等业务和未知流量，作为干扰数据，其中 HTTP 为主，约占“其它”流量的 75%以上。然后基于 HMM 工具箱[9]，将本文方法与典型方法[2–5]进行对比，结果如表 3 和表 4 所示。

由表 3 可见，本文方法对 4 种业务的区分准确度较高，而基于典型 HMM 多媒体业务分类方法区分业务效果欠佳。文献[2–5]方法同时忽略了业务内在 QoS 需求对区分特征的影响，将发射概率设为随机值，且演化过程中保持不变，违背了业务内在结构稳定性特点，导致区分效果欠佳。由于重传、网络资源限制、业务类型特点等因素的存在，相近的特征取值之间有较紧密的联系，输出观察值概率与多个状态有关。因此，似乎应该基于特征取值的上下文信息对业务区分，这体现业务自身内在规律性的要求。

本文方法依据位置信息改变发射概率，反映了特征取值的位置信息，一定程度上体现了特征值位

表 3 不同方法测试样本集的区分结果

识别率(%)		标清类流媒体	高清类流媒体	即时通信类	游戏类
标清类流媒体	本文方法	99. 83	0.15	0.01	0.01
	文献[2]方法	90.87	5.65	2.12	1.36
	文献[3]方法	59.94	38.79	1.06	9.21
	文献[4]方法	96.58	1.91	0.49	0.72
	文献[5]方法	89.94	9.43	0.36	0.27
高清类流媒体	本文方法	0.11	99.87	0.01	0.01
	文献[2]方法	6.01	90.05	2.07	1.87
	文献[3]方法	40.52	56.64	1.81	1.03
	文献[4]方法	1.52	97.39	0.51	0.58
	文献[5]方法	8.91	90.36	0.39	0.34
即时通信类	本文方法	0.07	0.05	97.67	2.21
	文献[2]方法	0.32	0.23	91.19	8.26
	文献[3]方法	1.22	1.15	76.76	20.87
	文献[4]方法	0.49	0.53	93.69	5.29
	文献[5]方法	0.42	0.37	92.78	6.43
游戏类	本文方法	0.07	0.02	1.79	98.12
	文献[2]方法	0.29	0.27	6.61	92.38
	文献[3]方法	1.70	1.07	41.62	55.61
	文献[4]方法	0.72	0.49	5.73	93.06
	文献[5]方法	0.44	0.48	8.11	90.97

表 4 不同方法的计算时间

方法类比	本文方法	文献[2]方法	文献[3]方法	文献[4]方法	文献[5]方法
计算时间(s)	20.738080	30.309623	15.978385	43.548917	45.685027

置间的上下文信息，较好地反映了其内在规律，取得较好的区分效果。由于包大小等多媒体业务特征受网络协议、业务类型等因素影响较大，其出现的位置具有内在的规律性。这是由于典型 HMM 多媒体业务分类方法中隐式序列为一阶马尔可夫链的约束条件具有内在局限性，侧重于反映区分特征的某个侧面，不能反映业务区分特征全部信息，特征考虑的全面程度不足，决定不利于提高区分效果，因此，有必要更深入研究区分特征的上下文信息。

表 4 的数据在 MATLAB R2008a 环境下获得，由表 4 可见，本文方法耗时较少，这是由于本文方法选择的区分特征最少，而计算复杂度随着状态个数的增加而增长。文献[2]方法选择两个区分特征，通过加权将两个特征联系起来，有较高的计算复杂度。文献[3]方法仅适合区分基于 TCP 协议的业务，不适合多媒体业务流的区分。文献[4]方法使用的区分特征需要较高的预处理成本，计算时间依然较大。文献[5]针对无线业务识别，计算复杂度较高且联合分布获取很困难。

6 结论

本文针对典型 HMM 多媒体业务分类方法使用包大小特征区分业务的不足，分析了典型多媒体业务包大小分布特点，通过引入包大小位置信息，改进典型 HMM 发射概率的取值，提高识别效果。仿真结果验证了本文方法的有效性。由于多媒体业务识别相关研究处于不断发展的阶段，一些其它关键问题还亟待解决。此外，发射函数的引入也改变了 HMM 中马尔可夫过程的性质，是否有负面效果的存在，也尚待深入展开研究。

参 考 文 献

[1] Michael F, Chris R, Eduardo R, *et al*.. A survey of payload-based traffic classification approaches[J]. *IEEE Communications Surveys & Tutorials*, 2014, 16(2):

1135–1156.

[2] Mu Xue-feng and Wu Wen-jun. A parallelized network traffic classification based on hidden Markov model[C]. IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Beijing, 2011: 107–112.

[3] El Khadimi A, Lmater M A, Eddabbah M, *et al*.. Packet classification using the hidden Markov model[C]. IEEE International Conference on Multimedia Computing and Systems, Ouarzazate, Morocco, 2011: 1–5.

[4] 许博, 陈鸣, 魏祥麟. 基于 HMM 模型的 P2P 流识别技术[J]. 通信学报, 2012, 33(6): 55–63.
Xu Bo, Chen Ming, and Wei Xiang-lin. Hidden Markov model based P2P flow identification technique[J]. *Journal on Communications*, 2012, 33(6): 55–63.

[5] Maheshwari S, Mahapatra S, Kumar C S, *et al*.. A joint parametric prediction model for wireless internet traffic using Hidden Markov Model[J]. *Wireless Networks*, 2013, 19(6): 1171–1185.

[6] Gold S. Hacking on the hoof[J]. *Engineering and Technology*, 2012, 7(3): 80–83.

[7] 张铭, 银平, 邓志鸿, 等. SVM+BiHMM:基于统计方法的元数据抽取混合模型[J]. 软件学报, 2008, 19(2): 358–368.
Zhang Ming, Yin Ping, Deng Zhi-hong, *et al*.. SVM + BiHMM: a hybrid statistic model for metadata extraction[J]. *Journal of Software*, 2008, 19(2): 358–368.

[8] Baggenstoss P M. A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces[J]. *IEEE Transactions on Speech and Audio Processing*, 2001, 9(4): 411–416.

[9] Kevin Murphy. HMM Toolbox. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm_download.html,2014.

王再见：　男，1980 年生，博士生，副教授，研究方向为端到端 QoS 保证技术、业务流识别.

董育宁：　男，1955 年生，博士生导师，教授，研究方向为多媒体通信与信息处理、绿色通信.

张　晖：　男，1982 年生，博士，副教授，研究方向为无线网络、云计算.