文章编号: 1001-0920(2017)02-0368-05 DOI: 10.13195/j.kzyjc.2016.0270

多元时间序列相似性度量方法

李正欣[a†], 郭建胜[a], 毛红保[b], 高杨军[a]

(空军工程大学 a. 装备管理与安全工程学院，b. 航空航天工程学院，西安 710051)

摘 要: 现有的多元时间序列相似性度量方法难以平衡度量准确性和计算效率之间的矛盾.针对该问题,首先,对多元时间序列进行多维分段拟合;然后,选取各分段上序列点的均值作为特征;最后,以特征序列作为输入,利用动态时间弯曲算法实现相似性度量.实验结果表明,所提出方法参数配置简单,能够在保证度量准确性的前提下有效降低计算复杂度.

关键词: 多元时间序列；相似性度量；特征提取；动态时间弯曲；计算复杂度

中图分类号: TP311 **文献标志码:** A

Similarity measure for multivariate time series

LI Zheng-xin[a†], *GUO Jian-sheng*[a], *MAO Hong-bao*[b], *GAO Yang-jun*[a]

(a. Equipment Management and Safety Engineering College，b. Aeronautics and Astronautics Engineering College，Air Force Engineering University，Xi'an 710051，China)

Abstract: Existing similarity measure for multivariate time series can't calculate similarity accurately and rapidly. Firstly, multivariate time series are fitted with the multidimensional piecewise method. Then, average values of original points in every segment are computed as the feature pattern. Finally, inputted by feature series, dynamic time warping is used to measure the similarity of multivariate time series. The results of experiments show that the process of its parameter choice is simple, and the proposed method can guarantee the measure accuracy at relatively low computational cost.

Keywords: multivariate time series; similarity measure; feature extraction; dynamic time warping; computational complexity

0 引 言

时间序列是一种与时间相关的高维数据,它广泛存在于金融、经济、工程领域[1],例如:股市逐日的交易情况、"黑匣子"记录的飞行数据等,随着时间的推移,这类数据的存储规模呈现爆炸式增长.

按时间顺序获取的一系列观测值$x_t(j)$称为时间序列,其中$t(t = 1,2,\cdots,n)$表示第t个时刻,$j(j = 1,2,\cdots,m)$表示第j个变量,$x_t(j)$表示第j个变量在第t个时刻上的记录值[2].当$m=1$时,$x_t(j)$为一元时间序列(UTS);当$m > 1$时,$x_t(j)$为多元时间序列(MTS).

时间序列数据挖掘已成为数据挖掘领域中的研究热点[3].相似性度量是时间序列数据挖掘的核心技术之一,其度量精度直接影响着数据挖掘的效果[4].

与一元时间序列相比,多元时间序列相似性度量的研究相对较少,还有较多尚未解决的问题[5].现实世界中,多元时间序列更具普遍性.股票交易可以用开盘价、收盘价、最高价、最低价、交易量来描述.此外,多媒体数据(如音频、图像等)经过转换也可以形成多元时间序列.因此,研究多元时间序列的相似性度量具有重要的理论意义和广阔的应用前景.

1 相关研究

目前,多元时间序列相似性度量方法主要有欧氏距离[6](ED)、奇异值分解[7](SVD)基于点分布特征方法[5](PD)、动态时间弯曲(DTW)距离[8]和趋势距离[9](TD)等.

ED方法简单直观,但要求两条序列的长度必须相同,且无法处理序列在时间轴上的伸缩和弯曲.

SVD方法把时间序列中的变量理解为随机变量,以相关系数矩阵作为特征提取的基础,利用线性坐标变换建立相似性度量模型,能够有效体现变量间的相

收稿日期: 2016-03-07；**修回日期:** 2016-06-07.

基金项目: 国家自然科学基金项目(61502521).

作者简介: 李正欣(1982−), 男, 讲师, 博士, 从事信息系统工程与智能决策、数据挖掘、机器学习等研究；郭建胜(1965−), 男, 教授, 从事信息系统工程与智能决策等研究.

[†]通讯作者. E-mail: lizhengxin_2005@163.com

互关系.但它是一种基于统计的度量方法,不能描述观察值的时序关系,存在一定的误判风险.

PD方法抽取多元时间序列在三维空间上的局部重要点作为特征,依据重要点的分布特征进行相似性度量,对小规模的多元时间序列具有较好的匹配效果.但它也是一种基于统计的度量方法.

DTW距离支持不同长度时间序列的相似性度量,支持序列在时间轴上的伸缩和弯曲,具有较好的度量精度和鲁棒性,因此被广泛采用[10];但由于计算复杂度高,限制了其在海量时间序列中的应用.

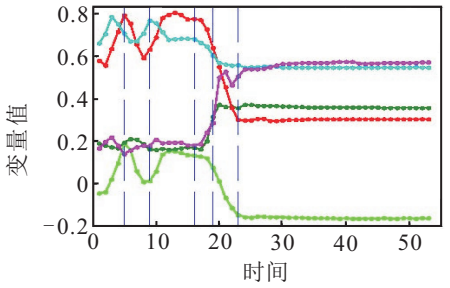
TD方法以多元时间序列的倾斜角和时间跨度作为特征,利用DTW算法实现特征序列的对齐匹配,与DTW方法相比,有效降低了计算复杂度;但模型参数较多、参数优化配置环节较为复杂.

针对以上问题,本文首先对多元时间序列进行多维分段拟合;然后选取各分段上序列点的均值作为特征;最后利用DTW算法度量相似性,并通过实验验证所提出方法的有效性.

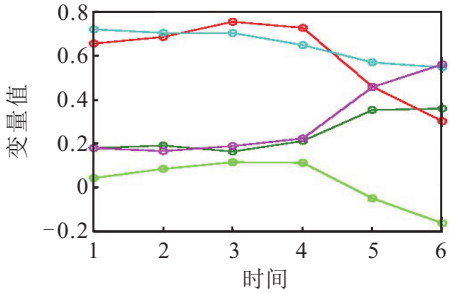
2　多元时间序列特征提取

特征提取是使用简单、突出的特征对多元时间序列进行描述.最直观的方法是分别提取每一变量维度上的特征,然后将这些特征依次排列构成特征向量.然而,这种方法没有考虑变量间的相关性.

本文采用多维分段拟合方法将多元时间序列分割为多个序列段,即同时对全部变量维度进行分段操作,如图1(a)所示.然后,将每个分段上原始点的均值作为该段序列的特征,如图1(b)所示.



(a)　多维分段拟合



(b)　分段平均

图1　多元时间序列特征提取示意图

多维分段保持了变量间的相关性,分段平均实现了时间维度上的降维.一个拟合段在第i个变量维度上的拟合误差定义为

$$e_i=\sqrt{\sum_{j=1}^{p}d_j^2},\ i=1,2,\cdots,m. \tag{1}$$

其中:d_j为第j个原始点与拟合直线的垂直距离,p为该拟合段中包含的原始点数量.对全部变量维度上的拟合误差求和,得到一个拟合段的总误差

$$e_{\text{seg}}=\sum_{i=1}^{m}e_i. \tag{2}$$

多维分段拟合采用滑窗分段策略,用一组直线拟合一个分割窗口内的数据点,当窗口内各线段拟合误差之和小于最大分段误差maxSegError,即

$$e_{\text{seg}}<\text{maxSegError} \tag{3}$$

时,进一步增加窗口宽度;否则,开启一个新窗口.直至序列完全被窗口包容,结束分段操作.

将各分段上序列点的均值作为特征.用$m\times n$的矩阵表示多元时间序列,m表示变量个数,n表示时间点数量,矩阵的行代表变量维,列代表时间维,则多元时间序列转换为特征序列$m\times n'$,且有$n>n'$.

3　多元时间序列相似性度量

设时间序列$X=(x_1,x_2,\cdots,x_n),Y=(y_1,y_2,\cdots,y_m)$,DTW距离$D_{\text{dtw}}(X,Y)$定义为[8]

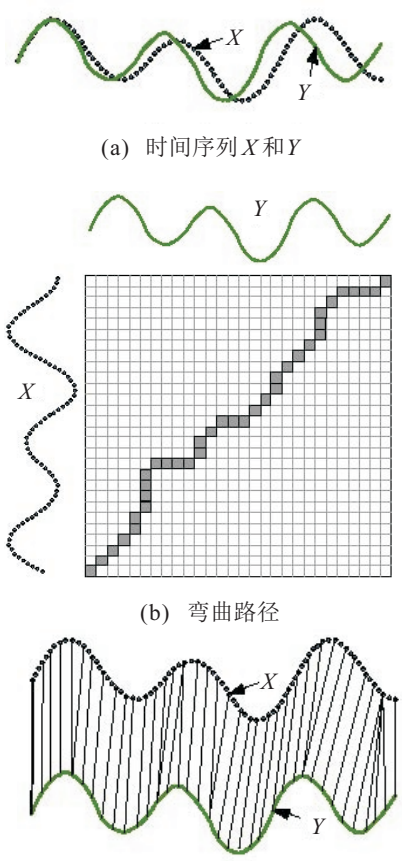
$$D_{\text{dtw}}(X,Y)=D_{\text{base}}(x_1,y_1)+\min\begin{cases}D_{\text{dtw}}(X,Y[2:-]);\\ D_{\text{dtw}}(X[2:-],Y);\\ D_{\text{dtw}}(X[2:-],Y[2:-]).\end{cases} \tag{4}$$

其中$D_{\text{base}}(x_i,y_j)$表示向量x_i和y_j之间的基距离,通常使用欧氏距离.

本质上,DTW距离用于确定序列X和Y上每个点之间的对齐匹配关系,如图2(c)所示,每种匹配关系可以用一条弯曲路径表示,如图2(b)所示.

弯曲路径必须满足3个基本条件[11]:1)边界条件,路径必须起始于点(x_1,y_1)、终止于点(x_n,y_m),它表示两个序列的起始点和结束点对应匹配;2)连续性,路径上的任意两个相邻点(x_{i1},y_{j1})和(x_{i2},y_{j2})须满足条件$0\leqslant|i_1-i_2|\leqslant1,0\leqslant|j_1-j_2|\leqslant1$;3)单调性,若$(x_{i1},y_{j1})$和$(x_{i2},y_{j2})$为路径上前后两个点,则须满足$i_2-i_1\geqslant0,j_2-j_1\geqslant0$.

满足上述条件的弯曲路径有很多,每一条弯曲路径都代表一种点对匹配关系.在所有的点对匹配关系中,点对基距离之和的最小值即为DTW距离,对应的



(a) 时间序列X和Y

(b) 弯曲路径

(c) 点对匹配结果

图 2 DTW距离的弯曲路径

弯曲路径称为最佳路径,可用动态规划法求解.本文所提出方法是在分段线性表示(PLR)的基础上使用DTW算法,因此可简记为PLR-DTW方法.

4 实验与分析

4.1 实验数据与实验方法

选用Australian Sign Language[12]、EEG[13]、Robot Execution Failure[14]和Japanese Vowels[15]4组分类结果已知的公开多元时间序列数据集进行实验.

Australian Sign Language(ASL)是包含22个变量的手语信号数据集,左、右手动作各用11个变量描述.选用前8种语意(alive, all, answer, boy, building, buy, change-mind, cold)对应的216个序列作为实验数据集.EEG是对两类人群测得的一组脑电图数据集,包含64个变量,选取前2位测试者的前11次测试作为实验数据.Robot Execution Failure是机械故障监控数据集,包含6个变量,选取第1个子数据集LP1进行实验.Japanese Vowels(JV)是日语元音发音数据集,包含12个变量,共270个样本.

实验采用k-近邻与留一交叉验证法.设数据集中有n个多元时间序列,从中任取一个输入序列X.对该数据集中的所有序列进行特征提取,然后采用某种相似性度量方法找出与X最相似的k个序列(k分别取1、5和10).在找出的k个序列中,计算与X同类的序列个数n_0,计算准确率

$$e=\frac{n_0}{k}. \tag{5}$$

对于数据集中其他多元时间序列,依次作为输入序列,可以得到n个相似性度量的准确率.计算平均准确率

$$e^*=\frac{1}{n}\sum_{t=1}^{n}e_t, \tag{6}$$

并将其作为度量有效性的比较依据.

实验环境为: Matlab R2010a, Windows7, Intel(R) Core(TM) i7-3770CPU, 4.00 G RAM.

4.2 相似性度量的有效性比较

针对ASL数据集,分别采用SVD、PD、DTW、TD和PLR-DTW等5种方法进行相似性度量,平均准确率见表1,其中3种方法涉及参数设置:PD方法分割形式为$X[i-1:i+1,\ j-1:j+1]$;TD方法中maxError$=0.03,\varepsilon=0.8,\lambda=0.2$;PLR-DTW方法中maxSegError $=$ 0.1.PLR-DTW方法的准确率最高,DTW和TD方法也能得到较好的度量效果.与DTW方法相比,PLR-DTW方法虽然对时间序列进行了降维,但却提高了度量准确性.这表明相似性度量中,使用特征序列可能比原始序列更有效.

表 1 ASL数据集上的相似性度量结果

	SVD	PD	DTW	TD	PLR-DTW
$k=1$	0.6944	0.5787	0.9769	0.9722	0.9954
$k=5$	0.6463	0.5704	0.9287	0.9582	0.9806
$k=10$	0.5616	0.5565	0.9287	0.9056	0.9718

进一步地,分别统计各个准确率上的k-近邻查询次数,见表2.PLR-DTW方法在准确率较低的情况下,对应的次数最少;准确率较高的情况下,对应的次数最多.整体上看,针对ASL数据集,DTW、TD和PLR-DTW方法优于SVD和PD方法.

用ASL数据集中第48个序列(ASL_48)作为输入,分别采用上述5种相似性度量方法,找出最相似序列,结果如图3所示.

DTW、TD和PLR-DTW三种方法找到的最相似序列与输入序列在形状上比较接近,且属于同一类,而SVD、PD方法找到的序列与输入序列的形状差别较大.主要原因在于,SVD、PD方法基于统计方式进行相似性度量,而没有考虑序列的对齐匹配关系.

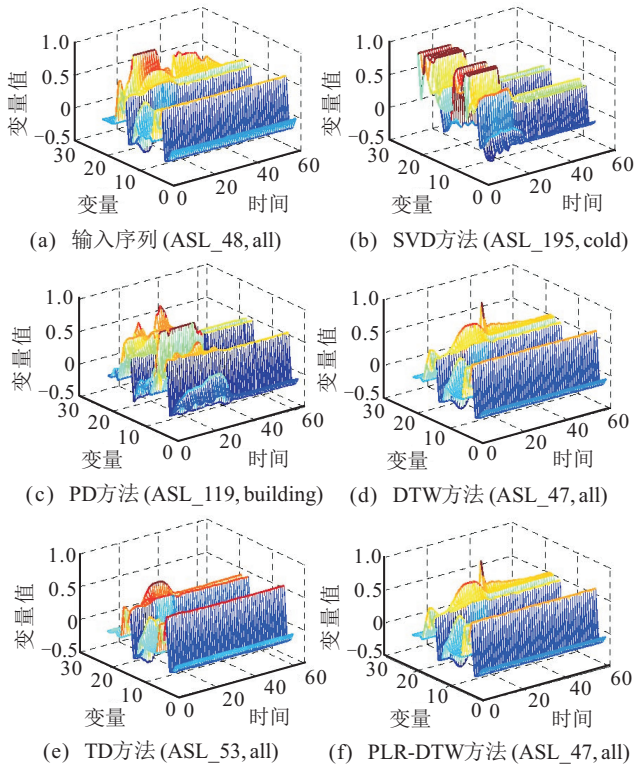
下面分别在EEG、LP1和JV数据集上,比较几种方法的有效性,结果见表3,对应的参数设置见表4.

表 2　不同准确率对应的k-近邻查询次数

准确率/%	SVD			PD			DTW			TD			PLR-DTW		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
$e=0$	66	19	15	91	33	14	5	0	0	6	0	0	1	0	0
$e=10$	—	—	15	—	—	13	—	—	0	—	—	0	—	—	0
$e=20$	—	25	19	—	25	17	—	0	0	—	0	0	—	0	0
$e=30$	—	—	15	—	—	28	—	—	1	—	—	2	—	—	0
$e=40$	—	34	24	—	36	18	—	13	0	—	4	3	—	2	0
$e=50$	—	—	29	—	—	26	—	—	6	—	—	9	—	—	0
$e=60$	—	30	18	—	34	17	—	12	15	—	9	8	—	2	3
$e=70$	—	—	6	—	—	13	—	—	15	—	—	7	—	—	6
$e=80$	—	25	10	—	23	6	—	14	11	—	15	22	—	11	9
$e=90$	—	—	6	—	—	12	—	—	30	—	—	30	—	—	13
$e=100$	150	83	59	125	65	52	211	177	148	210	188	135	215	201	185

表 3　其他数据集上的相似性度量结果

方法	EEG			LP1			JV		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
SVD	0.909 1	0.836 4	0.700 0	0.477 3	0.470 5	0.460 2	0.492 6	0.437 0	0.384 8
PD	0.818 2	0.663 6	0.513 6	0.909 1	0.818 2	0.737 5	0.570 4	0.474 8	0.450 0
DTW	0.727 3	0.590 9	0.563 6	0.886 4	0.747 7	0.614 8	0.955 6	0.900 7	0.854 8
TD	1.000 0	1.000 0	0.995 5	0.920 5	0.836 4	0.735 2	0.611 1	0.577 0	0.535 6
PLR-DTW	0.818 2	0.718 2	0.627 3	0.920 5	0.804 5	0.668 2	0.933 3	0.878 5	0.837 0



(a) 输入序列 (ASL_48, all)　(b) SVD方法 (ASL_195, cold)

(c) PD方法 (ASL_119, building)　(d) DTW方法 (ASL_47, all)

(e) TD方法 (ASL_53, all)　(f) PLR-DTW方法 (ASL_47, all)

图 3　ASL数据集中最相似序列匹配结果

最大分段误差maxSegError是PLA-DTW方法涉及的唯一参数,它决定了对多元时间序列的刻画粒度,进而影响着相似性度量精度.其取值与时间序列数据集自身的特点密切相关.

DTW、PLR-DTW方法在3组数据集上都具有较高的度量精度.TD方法针对EEG和LP1有着较高的精度,原因在于TD方法以序列的倾斜角和时间跨度作为特征,对序列特征的刻画更为精细;代价是涉及的参数较多、参数配置较为复杂,且对某些时间跨度较短的数据集(JV)的度量准确性较低.

表 4　不同数据集上的参数配置

数据集	PD	TD			PLR-DTW
	分割形式	maxEr	ε	λ	maxSegEr
EEG	$X[i-6:i+6,\ j-10:j+10]$	250	0	1	60
LP1	$X[i-3:i+3,\ j-3:j+3]$	180	0.2	0.8	160
JV	$X[i-3:i+3,\ j-1:j+1]$	5.5	0.5	0.5	5

4.3　计算复杂度比较

处理4组数据集时,TD、DTW和PLR-DTW方法具有较好的稳定性,它们都以DTW算法为基础,不同之处在于特征提取方式.DTW方法直接以原始序列作为特征, TD方法用序列的倾斜角和时间跨度作为特征,PLR-DTW方法在多维分段表示的基础上,以每一段上的原始序列点的均值作为特征.

下面比较3种方法的计算复杂度,DTW算法的计算复杂度为$O(m \times n)$,m和n表示参与运算的两个序列的长度.因此,3种方法的计算复杂度主要取决于特征序列长度.用特征序列压缩率的平方近似比较TD、PLR-DTW与DTW方法的计算复杂度,见表5.

表5 TD、PLR-DTW与DTW计算复杂度比较

数据集	均长	TD		PLR-DTW	
		特征均长	压缩率^2	特征均长	压缩率^2
ASL	59	32	0.54^2	20	0.34^2
EEG	256	62	0.24^2	186	0.73^2
LP1	15	2	0.13^2	2	0.13^2
JV	16	1	0.06^2	1	0.06^2

为了更加直观地比较计算复杂度,消除实验环境引起的偏差,用TD、PLR-DTW两种方法的计算时间分别除以DTW方法的计算时间,见图4.与DTW方法相比,TD、PLR-DTW方法都能较大幅度地降低相似性度量的计算复杂度.

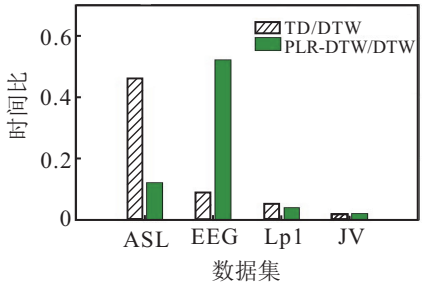


图4 TD、PLR-DTW与DTW方法计算时间的比较

5 结 论

本文对多元时间序列进行了多维分段拟合,以每一段序列的均值作为局部特征,提出一种基于DTW算法的相似性度量,并通过实验验证了所提出方法的有效性.

PLR-DTW方法的优势在于: 1) 支持序列在时间轴上的伸缩和弯曲,具有较好的度量准确性; 2) 与SVD、PD方法相比,体现了时序关系,能够消除误判风险; 3) 与DTW方法相比,实现了原始序列的降维,提高了计算效率; 4) 与TD方法相比,减少了参数数量,降低了参数优化配置的复杂度.

所提出方法用分段序列的均值刻画多元时间序列的局部特征,降维效果明显、模型参数少,但对各分段序列的跨度和局部形状特征缺少足够描述,有待展开后续研究.

参考文献(References)

[1] 李海林,郭崇慧. 基于多维形态特征表示的时间序列相似性度量[J]. 系统工程理论与实践, 2013, 33(4): 1024-1034.
(Li H L, Guo C H. Similarity measure based on multidimensional shape feature representation for time series[J]. Systems Engineering — Theory & Practice, 2013, 33(4): 1024-1034.)

[2] 李正欣,张凤鸣,李克武,等. 一种支持DTW距离的多元时间序列索引结构[J]. 软件学报, 2014,25(3): 560-575.
(Li Z X, Zhang F M, Li K W, et al. Index structure for multivariate time series under DTW distance metric[J]. J of Software, 2014, 25(3): 560-575.)

[3] Tak-chung Fu. A review on time series data mining[J]. Engineering Applications of Artificial Intelligence, 2011, 24: 164-181.

[4] 李海林, 郭韧, 万校基. 基于特征矩阵的多元时间序列最小距离度量方法[J]. 智能系统学报, 2015, 10(3): 442-447.
(Li H L, Guo R, Wan X J. A minimum distance measurement method for a multivariate time series based on the feature matrix[J]. CAAI Trans on Intelligent Systems, 2015, 10(3): 442-447.)

[5] 管河山,姜青山,王声瑞. 基于点分布特征的多元时间序列模式匹配方法[J]. 软件学报, 2009, 20(1): 67-79.
(Guan H S, Jiang Q S, Wang S R. Pattern matching method based on point distribution for multivariate time series[J]. J of Software, 2009, 20(1): 67-79.)

[6] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases[C]. Proc of the 4th Int Conf on Foundations of Data Organization and Algorithms. Chicago, 1993: 69-84.

[7] Kiyoung Yang, Cyrus Shahabi. An efficient k nearest neighbor search for multivariate time series[J]. Information and Computation, 2007, 205: 65-98.

[8] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[C]. Proc of the Workshop on Knowledge Discovery in Databases. Seattle, 1994: 229-248.

[9] 李正欣,张凤鸣,李克武. 多元时间序列模式匹配方法研究[J]. 控制与决策, 2011, 26(4): 565-570.
(Li Z X, Zhang F M, Li K W. Research on pattern matching method for multivariate time series[J]. Control and Decision, 2011, 26(4): 565-570.)

[10] Arash Gharehbaghi, Per Ask,Ankica Babic. A pattern recognition framework for detecting dynamic changes on cyclic time series[J]. Pattern Recognition, 2015, 48: 696-708.

[11] Keogh E, Ratanamahatana C. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2005, 7(3): 358-386.

[12] Mohammed Waleed Kadous. High-quality recordings of australian sign language signs[EB/OL].(2015-09-25). http://kdd.ics.uci.edu/databases/High-quality Australian Sign Language/High-quality Australian Sign Language.html.

[13] Henri Begleiter. EEG Database[EB/OL]. (2015-09-25). http: //kdd.ics.uci.edu/databases/eeg/eeg.html.

[14] Luis Seabra Lopes, Luis M Camarinha-Matos. Robot execution failures[EB/OL]. (2015-09-25). http://kdd.ics.uci.edu/databases/robotfailure/robotfailure.html.

[15] Mineichi Kudo, Jun Toyama, Masaru Shimbo. Japanese Vowels[EB/OL]. (2015-09-25). http://kdd.ics.uci.edu/databases/Japanese Vowels/Japanese Vowels.html.

(责任编辑: 孙艺红)