基于多特征向量的语音情感识别

付丽琴[1] 王玉宝[2] 王长江[1]
(中北大学仪器科学与动态测试教育部重点实验室 太原 030051)[1] (中国电子科学研究院 北京 100041)[2]

摘 要 同一组特征向量对不同的情感,其识别能力不同。以HMM作为语音情感分类器模型,对不同情感状态选择不同的特征向量进行识别。系统分两个阶段完成:首先基于漏识率和误识率最小的决策原则,采用优先选择(PFS)算法分别为每种情感状态选择最优的特征向量,然后用这些特征向量分别建立对应情感状态的HMM模型。利用北航情感语音库(BHUDES)对算法进行验证,将所有实验样本分为训练样本集、特征选择样本集和测试样本集3组,采用交叉实验的方法对本算法进行验证,结果表明,与单特征向量HMM相比,多特征向量HMM可达到更高的识别精度。

关键词 多特征向量,优先选择算法,决策,漏识率,误识率
中图法分类号 TP391 **文献标识码** A

Speech Emotion Recognition with Multiple Feature Vectors

FU Li-qin[1] WANG Yu-bao[2] WANG Chang-jiang[1]
(National Key Laboratory for Electronic Measurement Technology, North University of China, Taiyuan 030051, China)[1]
(China Academy of Electronics and Information Technology, Beijing 100041, China)[2]

Abstract Same feature vector from speech may recognize different emotion state in different reliability. HMM was used as basic classifier and different feature vectors were chose as the input of HMM for different emotion. Firstly, based on the decision principle of the least miss-recognition rate and error-recognition rate, promising first selection (PFS) was adopted to choose the optimal feature vector for each emotion. Then, HMM for each emotion was set up using the selected feature vector. Cross experiments were implemented using Beihang University Database of emotion speech (BHUDES). All samples were divided into three groups: training sample set, feature selection sample set and test sample set. The experimental results show that HMM with multiple feature vectors can achieve better recognition precise than that with single feature vector.

Keywords Multiple feature vectors, Promising first selection (PFS), Decision, Miss-recognition rate, Error-recognition rate

语音中的情感之所以能够被识别与表达,是因为语音特征在不同情感状态下的表现具有差异,因此很多研究者对特征与情感类别之间的对应关系进行了深入探讨。如Murray和Amott完成的实验给出了基音包络、强度、音质等语音特征在不同情感状态下的定性描述[1];Cowie等人在其文献[2]中详细地列出了前人对14种情感状态下语音特征的研究结果;对于汉语情感语音,文献[3]研究了喜、怒、惊、悲4种情感的时间、振幅、基频和共振峰构造特征和分布规律,并列出9个语音特征的变化情况。然而,因为语音信号的复杂性,对语音特征情感表现的研究受到各种干扰因素的影响,学者们对于各种语音特征与情感表达的关联程度还未达成普遍的共识。

尽管各种模式识别算法都可用到语音情感识别领域,但HMM由于善于处理动态时序信号,在语音情感识别领域的研究一直受到广泛的重视。Schuller[4]等人在其情感识别试验中采用连续的HMM对7种情感状态进行识别,用德语语音库实验,识别率达到77.8%。Nwe等人在文献[5]中使用离散HMM模型进行情感分类,通过试验表明具有4个状态的遍历型HMM性能优于Left-Right结构的HMM。Nogueiras等人在文献[6]中采用了基音和能量特征以及一个半连续的HMM模型对7种情感状态进行分类,结果表明当HMM状态数为64时系统的识别性能最优。文献[7]利用Mel频率倒谱系数(MFCC)作为情感特征矢量,针对情感语音中不同类别的音素训练不同的HMM,试验结果表明这种方法优于一种情感对应单一HMM模型的方法,作者还对基于韵律特征的HMM和SVM分别做了识别试验,结果表明基于韵律特征的HMM方法优于SVM方法。此外,研究结果表明[8],对应于每组特征向量,同一种情感分类算法对不同

到稿日期:2008-07-18 返修日期:2009-03-10 本文受国家863计划资助项目(2006AA01Z135)和教育部博士点基金资助项目(20070006057)资助。

付丽琴(1971—),女,博士,副教授,研究方向为智能信息处理、机器视觉,E-mail:liq.f@nuc.edu.cn;王玉宝(1974—),男,硕士,工程师,研究方向为信号处理;王长江(1971—),男,硕士,副教授,研究方向为智能系统。

情感状态识别的可靠性具有差异。

考虑到计算代价等问题，本文选择 6 状态的 Left-Right DHMM 作为基本分类器模型，研究汉语语音的情感识别问题，提出基于多语音特征向量的 HMM 分类方法，对不同情感状态采用不同的特征向量，并用汉语情感语音数据库进行实验。结果表明：与采用单一特征向量（包括特征搜索得到的最优特征组合）对所有情感状态识别的 HMM 分类器相比，本文算法可以得到更好的情感识别效果，该方法可为改善各类孤立 HMM 以及 HMM 混合模型的识别性能提供参考。

1 情感特征选择

有效情感特征的选择是语音情感计算领域的重要课题。目前公认的重要情感特征，如基频、能量以及倒谱系数等，都能在一定程度上区分某些情感状态。然而，由于不同情感状态的语音特征之间可能具有相似性，因此使用单一或少数几个特征难以区分出所有情感状态，如利用基频和平均振幅，可以很容易地将愤怒、喜悦、惊奇 3 种情感与悲伤、厌恶、恐惧区分开，但难以对愤怒、喜悦、惊奇 3 者之间进行区分。此外，由于语音特征的表现不仅仅受情感的影响，还与语音文本及说话人有关，因此，有效情感特征的选择始终是个难题。

HMM 语音情感识别是基于有监督的模式识别方法，即利用给定的训练样本集合进行分类模型的训练，然后用得到的分类模型对测试样本进行分类。样本在模式识别中以特征向量的形式出现，特征向量的选择对分类器性能有很大影响，如果不同情感类别的语音特征差异很大，就比较容易设计出具有较好性能的分类器。为此，研究者们一方面继续寻找有效表达情感的语音特征；另一方面则致力于在众多候选特征中搜索最有效的特征组合，比较常用的启发式搜索算法有序列前向选择（SFS）、序列后向选择（SBS）、优先选择算法（PFS）和逐步判别分析（SDA）法等[8]，每种算法都有各自的优缺点和适用范围。然而，由于特征对各种情感的分类性能不同，所选择的最优特征子集并不能保证对所有情感状态都取得最好识别率。

2 多特征向量的 HMM 识别

2.1 基本 HMM 分类模型

HMM 是一种利用特征矢量序列作为输入训练得到的统计信号模型，可用五元组 $\lambda=(N,M,\pi,A,B)$ 来描述，其中，N 表示状态数目；M 表示每个状态可能的观察值数目；A 为与时间无关的状态转移概率矩阵；B 是给定状态下，观察值的概率分布；π 为初始状态空间的概率分布。

HMM 分类器的建立过程，实际上就是为不同类别情感语音建立其对应 HMM 的过程。对于 K 种情感状态，可得到 K 个 HMM 模型。设观察值序列用 $O=o_1,o_2,\cdots,o_T$ 表示（T 为观察值序列的长度），由此模型产生此观察值序列的概率表示为 $P(O/\lambda)$。

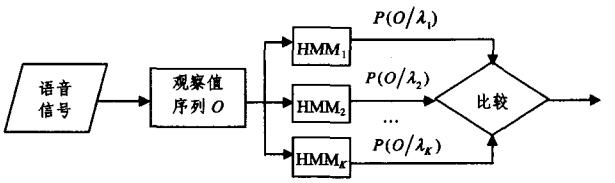

图 1 HMM 语音情感识别过程

图 1 所示为 HMM 语音情感识别过程。对每条情感语音，首先通过分帧和特征参数提取等步骤，得到观察值序列 O，并将 O 作为离散 HMM 的输入；然后用 Viterbi 算法[9]求出各个模型的 $P(O/\lambda_i)$ 值，λ_i 为对情感状态 i 的建模；最后选择最大 $P(O/\lambda_i)$ 值对应的情感状态作为识别结果。

2.2 多特征向量的 HMM 识别

2.2.1 基于线性加权法的多指标决策模型

选择不同的特征组合进行训练；然后用测试集中的部分待测语音进行实验，考察每一组特征向量的识别性能。我们发现，各组特征向量对不同情感状态识别的能力不一致，例如，有的向量对愤怒的识别率最高，而对其他情感的识别率则可能较低。因此，综合考虑各组特征向量的情感识别优势，可以得到性能更好的新的 HMM 分类器。

用 j 表示特征向量的序号，假设特征向量的个数为 J，则 j 的取值范围是 $1\sim J$；用 i 表示情感状态的序号，i 的取值范围是 $1\sim K$，K 是情感状态的个数。第 j 组特征向量对第 i 种情感状态的识别性能的优劣可从两个方面来权衡，一是模型能对该情感正确识别；二是模型不易对其它情感误识。因此，为了合理选择子 HMM，有两类错误需要考虑：

（1）Ⅰ类错误概率 $\varepsilon_{i,1}$

$\varepsilon_{i,1}$ 也叫漏识率，表示第 i 种情感没被正确识别的概率。设第 j 组特征向量对第 i 种情感的正确识别率为 p_i，则：

$$\varepsilon_{i,1}=1-p_i \tag{1}$$

（2）Ⅱ类错误概率 $\varepsilon_{i,2}$

$\varepsilon_{i,2}$ 也叫误识率，表示其它情感被误识为第 i 种情感的概率。

只有对应于每种情感的分类器模型发生上述两类错误的概率都很低，整个分类系统才有可能获得较高的平均识别率。因此，利用线性加权的方法构造决策函数如下：

$$\mu(C_{i,j})=1-(\varepsilon_{i,1}+\varepsilon_{i,2}) \tag{2}$$

其中 $C_{i,j}$ 表示第 j 组特征向量对第 i 种情感的识别性能。用 C_i^* 表示能最佳识别第 i 种情感状态的子 HMM，则应满足：

$$F(C_i^*)=\max_{1\leqslant j\leqslant L}(\mu(C_{i,j})) \tag{3}$$

对每种情感状态重复以上决策，就得到对应每种情感的最优特征向量。

2.2.2 多特征向量 HMM 识别

多特征向量的 HMM 分类器用 C 表示，则

$$C=\{C_1^*,C_2^*,\cdots,C_K^*\} \tag{4}$$

其中，C_i^* 是由式（1）到式（3）计算得到的子 HMM，它以能最佳识别第 i 种情感状态的特征向量作为输入。

整个 HMM 分类系统中，每个子 HMM 所用特征向量不一致，其识别模型如图 2 所示。对每条情感语音，通过分帧和特征参数提取等得到 K 个观察值序列 $O_1,O_2,\cdots,O_K$，分别作为 HMM_1 到 HMM_K 的输入，用 Viterbi 算法求出各个模型的 $P(O_i/\lambda_i)$ 值，然后选择最大 $P(O_i/\lambda_i)$ 值对应的情感状态作为识别结果。

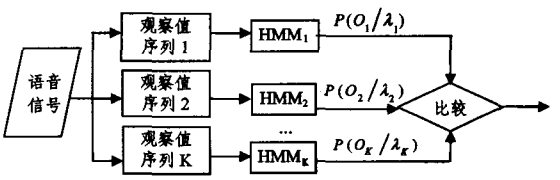

图 2 多特征向量 HMM 语音情感识别过程

3 实验与实验结果

3.1 数据库说明

本文采用北京航空航天大学情感语音工作组录制的诱发语音库(BHUDES)。该语料库采用了 Ekman 等建立的情感分类标准[10],包括悲伤、愤怒、惊奇、恐惧、喜悦、厌恶 6 种情感语句,说话人为 4 名男性和 3 名女性,采用了诱导型录音方式,说话人年龄在 20～30 岁之间,文化程度在本科以上。录音设备统一采用配备 sigmaTel C-major 声卡的华硕 M2413N-DR 笔记本电脑;头戴式麦克风;使用 GOLDWAVE 完成录音工作;采样率为 11025Hz,双通道、16Bit 量化,格式为 PCM。

20 句录音脚本集合基本覆盖汉语语音的主要元音和辅音,脚本长度控制在短句的范围内,在 3～12 字之间;采用口语化的陈述句,每句录音脚本均适于用各种情感进行表达,每条语句的每种情感被录制 3 次。录制工作完成后,利用专门的语音情感评价系统对所录语音进行评价,综合多人评价的结果得出每个语句的情感可信度,将可信度大于 0.7 的挑选出来作为本文所用的实验数据,共 714 条。

将所有实验数据分为 3 组,对应语句的第一次录音(254 句)、第二次录音(286 句)和第 3 次录音(174 句),分别作为训练样本,特征选择样本和测试样。

3.2 基于最优特征组合的 HMM 识别实验

首先提取常用情感特征,韵律特征包括:瞬时能量、过零率、能频积[11]、基频以及它们的一、二阶差分;声学特征包括:10 阶线性预测系数(LPC)、10 阶线性预测倒谱系数(LPCC)、10 阶 Mel 倒谱系数(MFCC)和第一共振峰频率及其一、二阶差分等,共 45 维。

基于分类器正确率判据对所选特征进行评估是最直接的方法,本文用优先选择(PFS)算法对数据集中提取的 45 维动态特征进行选择,评估模块选择离散 HMM 交叉验证正确率判别准则,由于 HMM 运算量较大,即使采用 PFS 算法仍然非常耗时,考虑到 45 维动态特征中 LPC,LPCC 和 MFCC 可分别视为一个整体,因此将其分别与其余 15 维特征中的每一维组合进行可分性判别,从而大大减少运算量。实验表明,对所有情感平均识别最优的特征组合为:基频的一、二阶差分+能量的一、二阶差分+10 阶 MFCC;当采用该组特征向量时,平均识别率达到 71.9%。

3.3 基于最优特征组合的 HMM 识别实验

用上述方法,以 HMM 对每种情感的识别性能作为依据,得到分别对应 6 种情感状态最佳识别性能的特征向量,如表 1 所列。其中,悲伤和惊奇对应同一组特征向量。

表 1 各分类器特征矢量

情感	特征向量
悲伤	第一共振峰及其一、二阶差分+10 阶 MFCC
愤怒	过零率的一阶差分+10 阶 LPCC
惊奇	第一共振峰及其一、二阶差分+10 阶 MFCC
恐惧	基音频率的一、二阶差分+10 阶 LPCC
喜悦	能频积及其一、二阶差分+10 阶 LPCC
厌恶	瞬时能量的一、二阶差分

分别以这 6 组特征向量作为输入得到 6 个 HMM 分类器,然后对第 2 组样本(特征选择样本)进行识别,结果如表 2 所列。其中,第一列标号 1～6 分别表示悲伤、愤怒、惊奇、恐惧、喜悦和厌恶 6 种情感状态,7 表示平均识别率。

表 2 各单特征向量 HMM 对第 2 次录音语句的识别(%)

	分类器 1,3		分类器 2		分类器 4		分类器 5		分类器 6	
	识别率	误失率	识别率	误失率	识别率	误失率	识别率	误失率	识别率	误失率
1	100	10.2	76.7	19.1	60.0	3.04	76.7	12.9	76.7	8.39
2	35.6	2.64	63.8	9.91	35.6	3.55	35.6	0.82	51.3	3.55
3	75.2	7.39	40.7	11.8	64.8	3.85	61.4	5.62	44.1	13.6
4	27.7	0.55	43.1	2.88	73.9	9.08	43.1	3.66	35.4	2.88
5	54.8	4.25	37.4	3.40	54.8	5.08	72.2	9.29	37.4	4.24
6	73.3	15.5	0.00	9.96	80.0	21.0	60.0	13.9	93.3	17.1
7	64.6	—	48.5	—	58.9	—	59.6	—	56.8	—

从表 2 的实验数据可以看出,6 个分类器对不同情感的识别性能有差异,如分类器 6 的平均识别率很低,但对厌恶的识别率却远高于其它 7 个分类器,因此可通过采用多特征向量 HMM 来提高系统整体识别性能。

对每种情感,选择对应其最优识别性能的特征向量进行建模,利用图 2 所示模型对第 2 次录音语句进行识别,其结果如表 3 所列,表中数值表示百分比。

表 3 多特征向量 HMM 对第 2 次录音语句的识别(%)

悲伤	愤怒	惊奇	恐惧	喜悦	厌恶	平均识别
100	70.0	79.1	83.6	65.0	95.0	80.9

3.4 识别实验

将第 3 次录音语句作为测试样本,分别用单特征向量分类器和多特征向量 HMM 进行情感状态的识别,得到结果如表 4 所列,表中数值表示百分比。

表 4 各单特征向量分类器对第 3 次录音语句的识别(%)

分类器	悲伤	愤怒	惊奇	恐惧	喜悦	厌恶	平均
1,3	94.4	45.9	64.4	32.5	45.0	77.1	60.0
2	55.6	68.2	57.0	32.5	20.0	28.6	47.4
4	83.3	38.5	53.3	45.0	80.0	48.6	58.7
5	77.8	23.7	68.2	32.5	70.0	91.4	57.8
6	88.9	34.8	42.2	32.5	30.0	91.4	48.3
本文算法	100	76.5	71.5	59.3	63.8	91.4	75.1

采用交叉实验的方法,从 3 组语音样本中分别选择原始训练集、重组训练识别集和测试集,共有 6 种方案进行实验,得到单分类器识别平均率和多特征向量 HMM 分类系统的平均识别率,如图 3 所示。其中,X 坐标的序号 1—6 表示第 1 到第 6 个分类器,序号 7 表示新分类器。Y 坐标代表各分类器的平均识别率。从图 3 可以看到,新分类器的识别率高于基于单一特征向量分类器。

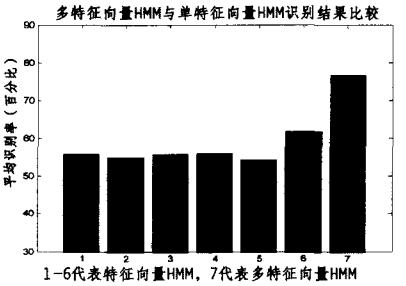

图 3 单特征向量 HMM 和多特征向量 HMM 的识别率比较

结束语 HMM 在情感计算领域具有重要地位。本文利用语音特征矢量对不同情感状态辨识可靠性的差异,基于漏识率和误识率最小的决策原则,为每种情感状态选择最优的特征向量,然后用这些特征向量建立对应情感状态的 HMM 模型。实验表明,在仅仅采用几种常用情感特征的条件下,新

分类系统仍可获得较高的识别精度。

实验中还发现：原始训练样本及重组训练用测试样本越多，得到对应每种情感的最优子 HMM 的可靠性越高，重组模型也越稳定；否则，当训练样本数量不是很多但达到一定数量时，分类系统中可能用到某些情感的次优特征向量，识别效果不是最理想，但相对于单特征向量分类器来说，整体性能仍得到提高。

参 考 文 献

[1] Murray I，Arnott J L. Towards the Simulation of emotion in Synthetic Speech：A review of the Literature on Human Vocal Emotion[J]. Journal of the Acoustic Society of American，1993，93 (2)：1097-1108

[2] Cowie R，Douglas-Cowie E，Tsapatsoulis N，et al. Emotion Recognition in Human-Computer Interaction [J]. IEEE Signal Processing magazine，2001，18(1)：32-80

[3] 赵力. 语音信号处理[M]. 北京：机械工业出版社，2003

[4] Schuller B，Rigoll G，Lang M. Hidden Markov Model - Based Speech Emotion Recognition[C]// ICASSP'03. 2003(2)：1-4

[5] Nwe T L，Foo S W，Silva L C D. Speech Emotion Recognition Using Hidden Markov Models [J]. Speech Communication，2003，41(4)：603-623

[6] Nogueiras A，Moreno A，Bonafonte A，et al. Speech Emotion Recognition Using Hidden Markov Models [A]// Eurospeech 2001 [C]. Seandinavia，2001

[7] Lee C M，Yildirim S，Bulut M，et al. Emotion Recognition Based on Phoneme Classes [A]// ICSLP 2004[C]. 2004：889-892

[8] 谢波，陈岭，陈根才，等. 普通话语音情感识别的特征选择技术[J]. 浙江大学学报：工学版，2007，41(11)：1816-1822

[9] Mao Xia，Zhang Bing，Luo Yi. Speech emotion recognition based on a hybrid of HMM/ANN[C]// The 7th WSEAS International Conference. 2007：181-184

[10] Ortony A，Tunen T J. What's Basic About basic Emotions[M]. Psychological Reviews，1997，3：315-331

[11] Chen Guanghua，Liu Junhai，Ye Jun. An improved method of endpoints detection based on energy-frequency-value[C]// IEEE Proceedings of HDP'06. 2006：9-11

(上接第 216 页)

{5}，{6}，{7}。

(4) 数据处理

首先分别按 3 种推理方式对 DBNs 模型分别执行 BK 推理算法和 1.5 片联合树算法(JT)12 次，记录下各自的运行时间；去掉运行时间的最大值及最小值，将剩下的 10 组数据取其平均值，如表 1 所列。由表 1 可知，BK 算法在滤波、平滑和固定步长平滑的推理中的时间性能要明显好于 1.5 片联合树算法。由图 3 可知，1.5 片联合树算法可以看成是 BK 算法的特殊情况，即当 BK 算法中将所有接口结点分为一个团时即成为 1.5 片联合树算法；且随着分团个数增加，误差逐步增加。

表 1 BK 算法与 1.5 片联合树算法时间性能比较表(时间单位：秒)

时间片		10	20	30	40	50	60	70	80	90	100
过滤	JT	0.187	0.397	0.591	0.803	0.952	1.196	1.398	1.547	1.908	1.962
	BK	0.129	0.250	0.382	0.528	0.645	0.788	0.924	1.063	1.211	1.289
平滑	JT	0.156	0.329	0.496	0.664	0.822	1.008	1.146	1.336	1.510	1.695
	BK	0.122	0.240	0.350	0.474	0.586	0.710	0.841	0.965	1.077	1.218
固定步长平滑	JT	0.215	0.469	0.685	0.963	1.183	1.426	1.659	1.908	2.156	2.373
	BK	0.170	0.363	0.557	0.736	0.940	1.134	1.317	1.519	1.707	1.914

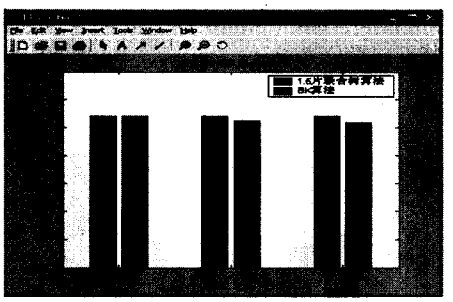

图 3 BK 算法精度分析

时间性能分析：

① 1.5 片联合树算法的时间复杂度 $O(M^I T)$。其中，M 表示状态变量最多能取到的值的个数，I 表示接口中所含的结点的个数，T 表示时间步骤。

② 引入分割团的 BK 推理算法的时间复杂度最少为 $O(T(M)^{\max(S(C_i))+\max(F_{C_i})+1})$。其中，$M$ 表示状态变量最多能取到的值的个数，F_{C_i} 是团 C_i 中变量的父结点数，$S(C_i)$ 表示 C_i 中所含结点数，T 表示时间步骤。

结束语 由于 DBNs 的推理是一个 NP 问题[10]，因而近似推理是主要的推理方式，BK 算法是 DBNs 的一种主要近似推理算法。BK 算法人工地在弱交互的子系统之间强加独立性，生成相对独立的团，并将其转换成联合树，进而通过联合树上结点之间的信息传播进行推理。为了减小 BK 算法的推理误差，本文给出一种引入分割团的新 BK 算法，并针对 Robocup 中的两个球员配合射门问题构建 DBNs，分别利用引入分割团的 BK 算法和 1.5 片联合树算法对 DBNs 进行推理，结果表明引入分割团使 BK 算法在精度损失较小的情况下，时间性能有显著提高。

参 考 文 献

[1] Murphy K. Dynamic Bayesian networks：representation，inference and learning[D]. University of California，Berkeley，2002

[2] Jensen F V，Jensen F. Optimal junction trees[C]// Proc. of UAI-94. 1994：360-366

[3] 周本达，王浩，姚宏亮. 1.5 片联合树算法在动态贝叶斯网精确推理中的应用[J]. 计算机工程与应用，2005，41(14)：81-84

[4] Boyen X，Kollen D. Tractable inference for complex stochastic processes[C]// Proc. of UAI-98. San Francisco：Morgan Kaufmann，1998：33-42

[5] 张润梅，王浩，姚宏亮. 一种基于影响图的决策方法及在 RoboCup 中的应用[J]. 系统仿真学报，2005，17(1)：134-137

[6] Dechter R. Bucket elimination：A unifying framework for probabilistic inference[C]// Proc. of UAI-96. San Francisco：Morgan Kaufmann，1996：75-104

[7] Draper D. Clustering without(thinking about) triangulation[C]// Proc. of UAI-95. San Francisco：Morgan Kaufmann，1995：125-133

[8] Kjaerulff U. Reduction of computational complexity in Bayesian networks through removal of weak dependences[C]// Proc. of UAI-94. 1994：374-382

[9] Paskin M A. Thin junction tree filters frontier for simultaneous localization and mapping[C]// Proc. of IJCAI-03. 2003：1157-1164

[10] Dagum P，Luby M. Approximating probabilistic inference using Bayesian networks is NP-hard[J]. Artificial Intelligence，1993，60(1)：141-151