基于遗传非参数MDL-BW方法的HMM结构优化

徐佳伟[1,2],罗 倩[1,2]
(1. 北京信息科技大学信息与通信工程学院,北京 100101; 2. 北京信息科技大学光电测试技术及仪器教育部重点实验室,北京 100101)

摘 要: 隐马尔科夫模型(Hidden Markov Model,HMM)广泛用于语音信号等时序信号的建模.HMM的结构优化包括模型参数个数优化和参数值的优化.针对传统的用于训练HMM的鲍姆-韦尔奇(Baum Welch,BW)算法在寻求最优解时容易陷入局部极值以及无法优化HMM参数个数的问题,本文提出了遗传非参数MDL-BW方法.该方法通过结合遗传(Genetic Algorithm,GA)算法随机搜索的特点和自适应思想来扩大HMM参数值解的搜索空间,结合非参数思想帮助自动寻求HMM的合适参数个数,同时以最小描述长度MDL(Minimum Description Length,MDL)作为模型优化准则来寻求HMM在全局上的最优结构.仿真数据、语音数据以及人体动作数据的仿真结果表明遗传非参数MDL-BW方法相较BW方法等同类方法在HMM结构的寻求上具有更好的效果.

关键词: 随机搜索; MDL准则; 非参数; 结构优化; 全局优化; 隐马尔科夫模型; BW方法
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2022)11-2765-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210870

HMM Structure Optimization Based on Genetic Nonparametric MDL-BW Method

XU Jia-wei[1,2], LUO Qian[1,2]
(1. *College of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing* 100101, *China*;
2. *Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science & Technology University, Beijing* 100101, *China*)

Abstract: Hidden Markov model(HMM) is widely used for modeling time series signals such as speech signals. The structural optimization of HMM includes optimization of the number of model parameters and parameter values. Aiming at the problem that the traditional Baum Welch(BW) method used to train HMM is easy to fall into local maxima and the number of parameters cannot be optimized when seeking the optimal solution, genetic nonparametric MDL-BW method was proposed. This method expanded the search space of parameter values of HMM by combining the characteristics of stochastic search of genetic algorithm(GA) with adaptive ideas, and combined nonparametric ideas to help automatically find the appropriate number of HMM parameters, and used minimum description length(MDL) as optimization criterion to find the global optimal structure of HMM. Based on simulation data, speech data and human action data, the results show that the genetic nonparametric MDL-BW method has a better performance in searching for the structure of the HMM comparing with the BW method and other similar methods.

Key words: stochastic search; minimum description length(MDL) criterion; nonparametric; structural optimization; global optimization; hidden Markov model(HMM); Baum Welch(BW) method

1 引言

隐马尔科夫模型(Hidden Markov Model,HMM)是一种常用于语音、动作等时序信号建模的贝叶斯网络[1]. 对于HMM的训练通常采取鲍姆-韦尔奇(Baum Welch,BW)算法[2],即一种期望最大化(Expectation Maximization,EM)算法. BW算法是一种坐标上升式的迭代算法,该方法收敛速度快,但有一个不足之处在于估计模型参数时容易陷入局部解. 针对BW算法这一缺点,基于全局优化的算法也尝试用于HMM的参数求解. 文献[3]在基于粒子群算法的基础上,采用自适应

收稿日期:2021-07-09;修回日期:2021-12-02;责任编辑:覃怀银
基金项目:企业委托基金(No.9151524108)

参数调整方式，提高了算法的全局搜索能力，实现对HMM初始参数的优化. 文献[4]将遗传算法(Genetic Algorithm，GA)用于HMM的训练，通过模拟进化论中物种演变过程来寻求最优解. 算法运行速度慢是以上基于全局优化算法的一个共同特点. 为此，文献[5]将GA算法和BW算法结合而成的GA-BW算法用于HMM的参数估计，GA-BW算法利用GA算法随机搜索的特性和BW算法在局部解上的快速收敛的特点，既能保证算法的全局优化能力，又能在参数优化上提升一定的收敛速度. GA-BW算法在估计HMM的参数值上具有很好的效果，但由于模型初始状态数需要提前预设，无法对模型参数个数进行推断.

本文通过结合GA算法、非参数思想[6]以及自适应思想并以最小描述长度MDL[7](Minimum Description Length，MDL)作为优化准则来优化HMM结构. 首先初始化多个具有不同参数个数和不同参数值的HMM个体. 结合GA算法中的重组操作和自适应思想，产生具有新的参数值个体；通过GA算法的变异操作并结合非参数思想，产生具有新的参数个数的个体，从而将模型结构解的搜索空间进一步扩大. 个体的进化方向由MDL准则决定，模型结构朝着MDL值减小的方向优化，最优的HMM结构由求得最小的MDL值确定. 遗传非参数MDL-BW算法在传统的BW算法之上通过结合随机搜索与自适应重组方式使得模型在更大的解空间上寻求参数值的最优解，减小了初始值对模型参数值求解的影响，同时针对BW算法需要提前设定参数个数的问题，使用结合非参数的变异操作，使得算法能自动寻求合适的模型参数个数，避免了参数个数盲目设定的问题. 仿真结果表明遗传非参数MDL-BW算法在HMM结构优化上具有很好的表现.

2 隐马尔科夫模型与经典求解算法

2.1 基于高斯混合分布的隐马尔科夫模型

HMM由初始状态概率向量$\boldsymbol{\varphi}$，状态转移概率矩阵$\boldsymbol{A}$和观测概率矩阵$\boldsymbol{B}$组成. 其元素集合分别为$\{\varphi_i, i=1,2,\cdots,I\}$，$\{a_{i,j}, i=1,2,\cdots,I;\ j=1,2,\cdots,I\}$以及$\{b_{j,k}, j=1,2,\cdots,I;\ k=1,2,\cdots,K\}$. I为模型的状态变量$S=\{s_t, t=1,2,\cdots,T\}$中某一时刻状态$s_t$的种类数. T为时间序列的长度. K为每个状态变量对应的观测分布个数. 当HMM是基于高斯混合分布时[8]，K是指每个状态变量s_t对应的高斯混合分布的高斯分布数. 在时刻t状态变量$s_t=j$条件下的高斯混合分布如式(1)：

$$p(\boldsymbol{x}_{n,t}|s_t=j)=\sum_{k=1}^{K}\gamma_{j,k}\mathrm{N}(\boldsymbol{x}_{n,t}|\boldsymbol{u}_{j,k},\boldsymbol{\Sigma}_{j,k}),\sum_{k=1}^{K}\gamma_{j,k}=1 \quad (1)$$

其中，$\boldsymbol{x}_{n,t}$指第n个时间序列数据时刻t时的采样数据，在时刻t状态变量$s_t=j$条件下的高斯混合分布$p(\boldsymbol{x}_{n,t}|s_t=j)$由$K$个高斯分布$\mathrm{N}(\boldsymbol{x}_{n,t}|\boldsymbol{u}_{j,k},\boldsymbol{\Sigma}_{j,k})$组合而成，$\gamma_{j,k}$是状态$s_t=j$下的第$k$个高斯分布的权重. 在基于高斯混合分布的HMM中，每个状态变量$s_t=j$对应一个高斯混合分布，$\gamma_{j,k}$对应于观测概率矩阵$\boldsymbol{B}$的第j行b_j中的元素$b_{j,k}$. 模型参数θ包括$\boldsymbol{\varphi}$，$\boldsymbol{A}$，$\boldsymbol{B}$，$\boldsymbol{u}$以及$\boldsymbol{\Sigma}$，其中$\boldsymbol{u}$和$\boldsymbol{\Sigma}$的元素集合分别为$\{\boldsymbol{u}_{j,k}, j=1,2,\cdots,I;\ k=1,2,\cdots,K\}$和$\{\boldsymbol{\Sigma}_{j,k}, j=1,2,\cdots,I;\ k=1,2,\cdots,K\}$. 时间序列数据$\boldsymbol{X}$的元素集合为$\{x_{n,t,d}, n=1,2,\cdots,N;\ t=1,2,\cdots,T;\ d=1,2,\cdots,D\}$，其中$N$为时间序列数据的个数，$T$指每个时间序列的长度，$D$是时间序列在某一时刻采样的数据维度. HMM对应时序上的概率图模型如图1. 模型的隐变量包括图中的状态变量$S=\{s_t, t=1,2,\cdots,T\}$和隐含变量$C=\{c_t, t=1,2,\cdots,T\}$.

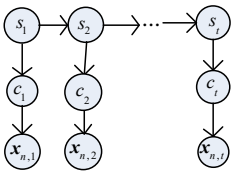

图1 高斯混合分布条件下的HMM

图1展示了第n个时间序列$\boldsymbol{x}_n$的生成过程如下. 在某一时刻t由状态$s_t=i$决定选取第i个高斯混合分布，再由此时刻的隐含变量$c_t=k$决定选取该高斯混合分布中的第k个高斯分布$\mathrm{N}(\boldsymbol{u}_{i,k},\boldsymbol{\Sigma}_{i,k})$，由此高斯分布$\mathrm{N}(\boldsymbol{u}_{i,k},\boldsymbol{\Sigma}_{i,k})$生成时刻$t$的样本$\boldsymbol{x}_{n,t}$. 即样本$\boldsymbol{x}_{n,t}$服从高斯分布$\mathrm{N}(\boldsymbol{u}_{i,k},\boldsymbol{\Sigma}_{i,k})$. 对于$t+1$时刻，由状态$s_t=i$根据状态转移矩阵$\boldsymbol{A}$跳转到时刻$t+1$的状态$s_{t+1}=j$，根据状态$s_{t+1}$选取第$j$个高斯混合分布，之后由$t+1$时刻的隐含变量$c_{t+1}$决定选取该高斯混合分布中的某个高斯分布，例如第$k+1$个高斯分布，来生成样本$\boldsymbol{x}_{n,t+1}$，此样本服从高斯分布$\mathrm{N}(\boldsymbol{u}_{i,k+1},\boldsymbol{\Sigma}_{i,k+1})$. 重复上述过程，可得到一个完整的时间序列$\boldsymbol{x}_n$.

2.2 GA算法与BW算法

GA算法[4]是一种模拟自然进化，在定义的搜索空间上进行全局搜索的随机优化方法. 该算法通过对亲代模拟遗传中基因重组和变异操作，得到子代个体，取评估较优的部分个体组成新一轮的亲代个体. 重复操作，直至达到终止条件，得到最优个体.

BW算法是EM算法在HMM的具体实现[9]. EM算法用于含有隐变量的概率模型参数估计. 同理，BW算法用于含有隐变量，即状态变量S和隐含变量C的HMM参数θ的估计. 与EM算法类似，BW算法通过计算期望值和求期望极大值两步迭代，最终收敛得到参数估计值.

3　遗传非参数MDL-BW方法

3.1　GA-BW算法

GA-BW算法是GA算法与BW算法的结合[5,10]. 对于HMM的参数估计,首先要对参数$\boldsymbol{\varphi}$,$\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{u}$和$\boldsymbol{\Sigma}$进行编码. 编码方式一般有二进制编码和实数编码方式. 由于参数$\boldsymbol{\varphi}$,$\boldsymbol{A}$和$\boldsymbol{B}$要满足如式(2)的限制条件,即$\boldsymbol{\varphi}$中元素之和,$\boldsymbol{A}$中某行元素之和以及$\boldsymbol{B}$中某行元素之和为1,因此采取实数编码方式较为方便.

$$\begin{cases}\sum_{i=1}^{I}\varphi_i=1\\ \sum_{j=1}^{I}a_{i,j}=1\\ \sum_{k=1}^{K}b_{i,k}=1\end{cases}\tag{2}$$

基于高斯混合分布的HMM的参数实数编码方式如图2. 参数$\boldsymbol{\varphi}$默认为单位向量,不参与编码. 图2也可看作编码后个体中的染色体,每个参数视为染色体中的一个基因.

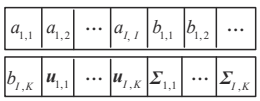

图2　参数实数编码方式

本文进一步要优化的HMM为自左向右型模型,即状态转移矩阵$\boldsymbol{A}$如下,每个状态只转换成自身的状态,或向下一个状态转换,向其他状态转换的概率为0;最后一个状态只转换成自身的状态,不再往其他状态转换. 自左向右模型减小了染色体的编码长度,也更加实际可用.

$$\boldsymbol{A}=\begin{bmatrix}a_{1,1}&a_{1,2}&0&\cdots&0\\0&a_{2,2}&a_{2,3}&\cdots&0\\0&0&a_{3,3}&a_{3,4}&0\\0&0&\cdots&a_{I-1,I-1}&a_{I-1,I}\\0&0&\cdots&0&a_{I,I}\end{bmatrix}$$

对于重组操作步骤,由于参数$\boldsymbol{A}$,$\boldsymbol{B}$要满足式(2)的限制,设两条染色体中$\boldsymbol{A}$、$\boldsymbol{B}$两部分基因之间交换的最小单位为$\boldsymbol{A}$、$\boldsymbol{B}$中的某行. 随机确定的交换位置也用于$\boldsymbol{u}$和$\boldsymbol{\Sigma}$交换位置的确定. 将两条染色体对应位置上的基因进行交换也就产生出含有新的染色体的个体. 变异操作根据变异概率将数据的均值和方差分别赋给被选定为变异集的$\boldsymbol{u}$和$\boldsymbol{\Sigma}$值.

3.2　遗传非参数MDL-BW算法

由3.1部分可知,重组和变异操作保证了模型产生新的参数值. 同理,可利用变异操作结合非参数做法对模型参数个数进行改变. 对于某个状态下的高斯混合分布,计算每两个高斯分布之间的相似度. 大于阈值e($0\le e\le 1$)的相似度所对应的两个高斯分布,随机选取其一作为变异候选集合中的一员,重复此操作得到每个状态下的变异候选集合,每个状态的变异候选集合的元素个数即该状态下的冗余高斯分布数. 找到变异候选集合元素数目最多的状态,分别计算该状态下删除冗余高斯分布数的MDL值MDL1和删除此状态后的MDL值MDL2,若MDL1<MDL2,则删除冗余高斯分布数,否则删除此状态. 非参数变异操作使得模型能根据MDL值去寻求模型合适的状态个数和高斯混合数.

重组操作通过结合随机搜索与自适应重组,来帮助产生具有新参数值的子代,以扩大解的搜索空间. 为了保证两个亲代参数$\boldsymbol{A}$或参数$\boldsymbol{B}$中的行与行之间能进行交换,仅对$\boldsymbol{A}$和$\boldsymbol{B}$具有相同维度的两个亲代之间进行重组操作. 随机选取亲代θ_1中参数$\boldsymbol{A}_1$的某行a_i和亲代θ_2中参数$\boldsymbol{A}_2$的某行a_j. 重组后得到的新的子代θ_1'对应的行a_i'和子代θ_2'对应的行a_j'的值如式(3).

$$\begin{cases}a_i'=\eta\cdot a_i+(1-\eta)a_j\\ a_j'=(1-\eta)a_i+\eta\cdot a_j\end{cases}\tag{3}$$

其中,η定义如式(4):

$$\eta=\frac{\mathrm{MDL}(\theta_2)}{\mathrm{MDL}(\theta_1)+\mathrm{MDL}(\theta_2)}\tag{4}$$

式(4)中,$\mathrm{MDL}(\theta_1)$和$\mathrm{MDL}(\theta_2)$分别为两个亲代的MDL值. 由式(3)可以看到重组后的两个子代参数a_i'或a_j'的值不是直接复制亲代a_i或a_j的值,而是由亲代的MDL值加权而定,同理,子代参数b_i'和b_j'的值也由亲代的MDL值加权而定. 由于MDL值越小,表示模型结构越准确,所以式(4)中的加权系数η是用$\mathrm{MDL}(\theta_2)$比上$\mathrm{MDL}(\theta_1)$与$\mathrm{MDL}(\theta_2)$之和. 若$\mathrm{MDL}(\theta_1)$大于$\mathrm{MDL}(\theta_2)$,则对子代θ_1'的a_i'的组成而言,亲代θ_1的比例小于亲代θ_2的比例,子代θ_1'的值朝着MDL减小的方向改变;相反,对于MDL值较小的亲代θ_2而言,子代θ_2'朝着MDL值增大的方向改变. 总之,结合自适应的重组操作保证了参数解空间的扩大. 遗传非参数MDL-BW算法步骤,如算法1.

算法1　遗传非参数MDL-BW算法

输入:数据$\boldsymbol{X}$,初始化变量(包括个体数num_pop,子代数num_off,最大状态数Max_S,最大高斯分布混合数Max_C,相似度阈值e,β值)

随机初始化多个亲代个体$\theta_1,\theta_2,\cdots,\theta_{\mathrm{num_pop}}$

WHILE 某个个体没有连续5次评为最优个体 DO

　BW算法优化个体θ值

　由式(3)、式(4)进行自适应重组操作,得到num_off个子代个体

　BW算法优化子代个体θ'值

　由式(5)计算所有个体的MDL值,得出最优个体,选取num_pop个较优的个体组成新一轮的亲代,对新一轮亲代中非最优个体进行如下非参数变异操作,设某个个体的状态数为num_S

　FOR $i=1$: num_S

根据阈值e计算每个状态i的高斯分布变异候选集合

END FOR

选出变异候选集合元素最多所对应的状态，分别计算删除冗余高斯分布数的MDL1和删除此状态后的MDL2

IF MDL1 < MDL2

删除冗余的高斯分布数

ELSE

删除此状态

END IF

得到具有新的参数个数的个体

END WHILE

输出：最优个体的状态转移矩阵$\boldsymbol{A}$、观测概率矩阵$\boldsymbol{B}$、$\boldsymbol{u}$和$\boldsymbol{\Sigma}$

在算法1中，随机初始化后的个体是num_pop个具有不同参数个数以及不同参数值的个体. BW算法步骤优化每个个体的参数值，自适应重组操作步骤产生具有新的参数值的个体，进一步将解的搜索空间扩大. 非参数变异操作根据MDL准则决定是改变高斯混合数还是改变状态数，使得模型朝着MDL值减小的方向改变模型参数个数.

3.3 非参数模型选择准则

非参数指模型的参数个数可以随着数据的变化而进行改变，非参数技术能根据数据去寻找合适的模型结构. 本文利用MDL准则来寻求模型最优结构. MDL准则如式(5).

$$\mathrm{MDL}=-\log P(\boldsymbol{X}|\theta)+\beta\left[H(L+1)+Q\right]\log N \tag{5}$$

在基于混合高斯分布的HMM中，式(5)中的H指模型高斯分布的个数，即状态S的个数与每个状态下的高斯混合数的乘积；L是每个高斯分布的参数个数，维度为D的多元高斯分布的参数个数为$D+D(D+1)/2$个；N为时间序列数据$\boldsymbol{X}$中的时间序列$\boldsymbol{x}_n$的个数；Q为状态转移矩阵$\boldsymbol{A}$中不为0的元素个数.

例如对于一个状态数为5，每个状态下的高斯混合数为2，每个高斯分布的维度$D=2$的HMM. 模型高斯分布个数H为状态S的个数与高斯混合数的乘积，即$H=5\times2=10$；每个高斯分布的参数个数L由均值向量元素个数加上协方差矩阵中有效元素个数. 每个高斯分布均值向量的元素个数与高斯分布维度数一致，为D个. 由每个高斯分布协方差矩阵的对称性可知，协方差矩阵有效元素个数只取矩阵上三角元素个数即可，为$D(D+1)/2$个. 所以$D=2$的每个高斯分布的参数个数$L=D+D(D+1)/2=2+2(2+1)/2=5$；由3.1节自左向右型状态概率转移矩阵$\boldsymbol{A}$的结构可知，除最后一行代表的状态只有1个不为0的元素，其余行均有2个不为0的元素. 所以状态数为5的转移概率矩阵$\boldsymbol{A}$中不为0的元素个数$Q=5\times2-1=9$. 给定β值即可计算MDL的第二项，此项用于约束模型的参数个数. MDL第一项$-\log p(\boldsymbol{X}|\theta)$是对数似然函数的负值，表达模型对数据$\boldsymbol{X}$的拟合程度. 该项计算过程如下.

设$\boldsymbol{x}_n$是$\boldsymbol{X}$中第n个时间序列数据，即$\boldsymbol{X}$的元素集合为$\{\boldsymbol{x}_n, n=1,2,\cdots,N\}$，所以

$$\log p(\boldsymbol{X}|\theta)=\sum_{n=1}^{N}\log p(\boldsymbol{x}_n|\theta) \tag{6}$$

进一步需要计算在模型参数θ下，每个时间序列数据$\boldsymbol{x}_n$的概率$p(\boldsymbol{x}_n|\theta)$，具体可参见文献[9,11]中利用前向后向算法计算得到该值，其中递推式如式(7)，$\boldsymbol{x}_{n,1:t}$表示第n个时间序列数据的第1个至第t个采样样本. $\alpha_t(j)$表示第n个时间序列的第1个至第t个样本与时刻t时状态$s_t=j$的联合概率.

$$\begin{cases}\alpha_t(j)=p(\boldsymbol{x}_{n,1:t}, s_t=j)\\ \alpha_{t+1}(j)=\left[\sum_{i=1}^{I}\alpha_t(i)\cdot a_{i,j}\right]\cdot p(x_{n,t+1}|s_{t+1}=j)\end{cases} \tag{7}$$

需要注意的是，式(7)的$p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j)$是服从如式(1)的高斯混合分布. 要计算此项，需先求得概率$p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j, c_{t+1}=k)$，由$s_{t+1}=j, c_{t+1}=k$条件下$\boldsymbol{x}_{n,t+1}$的高斯分布

$$\begin{aligned}&p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j, c_{t+1}=k)=\mathrm{N}(\boldsymbol{x}_{n,t+1}|\boldsymbol{u}_{j,k}, \boldsymbol{\Sigma}_{j,k})\\ &=(2\pi)^{-\frac{d}{2}}\left|\boldsymbol{\Sigma}_{j,k}\right|^{-\frac{1}{2}}\exp\left[\frac{1}{2}(\boldsymbol{x}_{n,t+1}-\boldsymbol{u}_{j,k})^{\mathrm{T}}\boldsymbol{\Sigma}_{j,k}^{-1}(\boldsymbol{x}_{n,t+1}-\boldsymbol{u}_{j,k})\right]\end{aligned} \tag{8}$$

可知当给定采样$\boldsymbol{x}_{n,t+1}$和θ时，由式(8)可以计算出$t+1$时刻样本$\boldsymbol{x}_{n,t+1}$概率值$p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j, c_{t+1}=k)$. 又由

$$\begin{aligned}p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j)&=\sum_{k=1}^{K}p(\boldsymbol{x}_{n,t+1}, c_{t+1}=k|s_{t+1}=j)\\ &=\sum_{k=1}^{K}p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j, c_{t+1}=k)\cdot p(c_{t+1}=k|s_{t+1}=j)\end{aligned} \tag{9}$$

其中，$p(c_{t+1}=k|s_{t+1}=j)$为观测概率矩阵$\boldsymbol{B}$的第j行中的元素$\{b_{j,k}, k=1,\cdots,K\}$，可计算得到$p(\boldsymbol{x}_{n,t+1}|s_{t+1}=j)$.. 但通常为了防止值下溢，会对进一步计算出的$\alpha_t(j)$进行归一化，归一化处理具体可参见文献[11]. 由前向后向算法计算得到$p(\boldsymbol{x}_n|\theta)$. 进一步在给定数据集$\boldsymbol{X}$和初始参数$\theta$下，由式(6)可以计算出$\log p(\boldsymbol{X}|\theta)$，求得MDL准则的第一项.

MDL值越小，说明模型的结构越适合于数据所隐含的模型结构. 由于第一项是关于数据$\boldsymbol{X}$的对数似然函数的负值，要使MDL减小则要增大对数似然函数值. 第二项是关于模型参数个数项，要减小MDL的值则需要减少模型的参数个数. 式(5)的参数β用于调节第二项对于MDL值的影响程度. MDL准则的物理意义是希望构建的模型具有尽量少的参数个数，即越简单，又同时尽量满足高似然值，即模型能更准确地描述数据$\boldsymbol{X}$. 本文将MDL准则作为遗传非参数MDL-

BW算法的优化准则.

4　数据分析

4.1　仿真数据分析

随机生成序列个数$N=50$,每个序列长度$T=25$,维度$D=2$的仿真数据$\boldsymbol{X}$,其元素集合为$\{x_{n,t,d}, n=1,\cdots,N;\ t=1,\cdots,T;\ d=1,\cdots,D\}$,如图3,$y$、$z$分别表示仿真数据在平面上的位置,每一种颜色表示一种状态,在某种状态下的数据是服从高斯分布数为2的高斯混合分布.数据服从的状态转移矩阵$\boldsymbol{A}$和观测概率矩阵$\boldsymbol{B}$如下,遗传非参数MDL-BW算法的初始值设置如下,$\beta=0.667$, $e=0.4$, num_pop=10,num_off=6,Max_S=8以及Max_C=5.

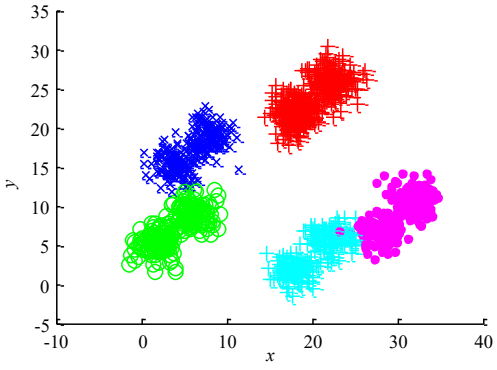

图3　仿真数据

$$\boldsymbol{A}=\begin{bmatrix}0.75 & 0.25 & 0 & 0 & 0\\ 0 & 0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.85 & 0.15 & 0\\ 0 & 0 & 0 & 0.9 & 0.1\\ 0 & 0 & 0 & 0 & 1\end{bmatrix},\quad \boldsymbol{B}=\begin{bmatrix}0.5 & 0.5\\ 0.5 & 0.5\\ 0.5 & 0.5\\ 0.5 & 0.5\\ 0.5 & 0.5\end{bmatrix}$$

由遗传非参数MDL-BW算法得到的状态转移矩阵$\boldsymbol{A}$和观测概率矩阵$\boldsymbol{B}$如下.由矩阵$\boldsymbol{A}$可以看到对于状态个数的推断是正确的,为5个,对应的转移概率元素值与原仿真数据所用的$\boldsymbol{A}$中元素基本相等.由矩阵$\boldsymbol{B}$可以看出,虽然推断出来的每种状态对应的高斯混合数为4,但是权重都集中在两个高斯分布上,而且比例基本为1∶1,与原观测概率矩阵$\boldsymbol{B}$一致.例如$\boldsymbol{B}$中首行所代表的第1个状态的混合高斯分布的权重集中在第1列代表的高斯分布,权重为0.5377,和第4列代表的高斯分布,权重为0.4531,权重之比基本为1∶1.

$$\boldsymbol{A}=\begin{bmatrix}0.7699 & 0.2301 & 0 & 0 & 0\\ 0 & 0.8142 & 0.1858 & 0 & 0\\ 0 & 0 & 0.8303 & 0.1697 & 0\\ 0 & 0 & 0 & 0.8989 & 0.1011\\ 0 & 0 & 0 & 0 & 1\end{bmatrix},$$

$$\boldsymbol{B}=\begin{bmatrix}0.5377 & 0.0046 & 0.0047 & 0.4531\\ 0.4911 & 0.0048 & 0.0223 & 0.4818\\ 0.0038 & 0.4920 & 0.0037 & 0.5004\\ 0.5013 & 0.4952 & 0.0035 & 0\\ 0 & 0.4844 & 0.5075 & 0.0081\end{bmatrix}$$

图4为遗传非参数MDL-BW算法估计出的均值$\boldsymbol{u}$与方差$\boldsymbol{\Sigma}$在原仿真数据上的图示,y、z分别表示仿真数据在平面上的位置.图中的"*"点表示混合高斯分布中的某个高斯分布的均值,轮廓线显示出方差的大小.可以看出,估计出的高斯分布的参数$\boldsymbol{u}$和$\boldsymbol{\Sigma}$与原始仿真数据分布基本一致.因此,由仿真数据的矩阵$\boldsymbol{A}$、$\boldsymbol{B}$和基于遗传非参数MDL-BW算法得出的矩阵$\boldsymbol{A}$、$\boldsymbol{B}$对比以及图3与图4的对比,表明该方法在基于高斯混合分布的HMM结构优化上具有良好的效果.

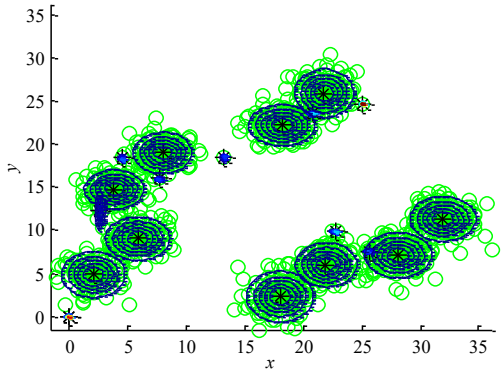

图4　参数$\boldsymbol{u}$和$\boldsymbol{\Sigma}$在原始数据上的图示

4.2　语音数据分析

语音识别是HMM的典型应用领域,本文进一步进行孤立词语音识别仿真.采用的语音数据集为FSDD(https://github.com/Jakobovski/free-spoken-digit-dataset),其中,包含数字"0"~"9"共10个数字的英文录音,每个数字有300个录音,一共3000个录音.本文选取每个数字的一半录音数据作为训练集,对应另一半录音数据作为测试集.对训练集数据,先提取每类数字的13个梅尔频率倒谱系数MFCC(Mel Frequency Cepstrum Coefficient, MFCC),再结合MFCC利用遗传非参数MDL-BW算法对每类数字建立其对应类型的HMM,该算法初始值设置如下,$\beta=0.018$, $e=0.4$, num_pop=10, num_off=6, Max_S=8以及Max_C=5.用作比较的BW算法的状态数设置为5,高斯混合数为2;基于似然函数GA算法[4]的状态数设置为5,高斯混合数为3;基于MDL准则改进的GA算法参数值设定与本文提出方法一致;GA-BW算法[5]的状态数和高斯混合数与BW算法设置一致.不同算法的数字语音识别准确率如表1.

由表1可知,遗传非参数MDL-BW算法的数字孤立

词语音识别准确率在同类算法中的识别准确率最高，为82.40%，识别准确率相较BW算法提升了33.93%，相较基于似然函数GA算法[4]提升了23.53%，相较GA-BW算法[5]提升了17.2%. 对比BW算法，本文提出方法性能提升的主要原因是本文方法通过扩大模型结构的搜索空间，使得算法不再受制于预先设定的模型结构；同时以MDL作为优化准则，保证算法在较大的解空间上能找到准确的模型结构. 对比基于似然函数GA算法和GA-BW算法，本文方法在扩大参数解搜索空间的基础上，针对上述方法无法优化模型参数个数的问题，将非参数的做法融入变异操作中，并利用MDL准则作为优化准则，使得模型参数个数也得到优化，进而得到更为准确的模型结构，提升了模型性能. 此外，表1中基于MDL的GA算法性能优于基于似然函数的GA算法性能，也是因为前者在后者的基础上，将模型参数个数进行非参数化，使得模型参数个数能进一步优化. 由此可见，HMM参数个数优化对模型性能提升是十分重要的.

表1 不同算法的数字语音识别准确率

方法	准确率/%
BW算法	48.47
基于似然函数GA算法[4]	58.87
基于MDL的GA算法	77.53
GA-BW算法[5]	65.20
遗传非参数MDL-BW算法(本文)	82.40

由于BW算法受初始设定参数个数的影响，以及基于似然函数的GA算法和GA-BW算法无法优化模型参数个数的问题，本文进一步根据遗传非参数MDL-BW算法推断出的模型状态数在5至8个之间，高斯混合数基本为5个的条件下，进行在统一预设高斯混合数为5，不同状态数下的HMM数字孤立词语音识别仿真，识别准确率如表2. 由表2可知，在与遗传非参数MDL-BW算法具有相似状态数和高斯混合数的情况下，BW算法的数字孤立词语音识别准确率最高为80.60%，基于似然函数的GA算法准确率最高为81.40%，GA-BW算法准确率最高为81.93%，仍都低于本文方法的准确率82.40%，可见本文提出方法对模型参数值优化的效果更好. 另一方面，对比表1中各个算法的准确率可知，

表2 不同算法在统一预设高斯混合数为5以及不同状态数下的识别准确率 单位:%

方法	状态数				
	5	6	7	8	9
BW算法	73.00	74.80	78.20	80.60	79.00
基于似然函数GA算法[4]	61.53	71.73	81.40	75.87	72.80
GA-BW算法[5]	79.87	81.93	80.27	80.20	78.07

利用遗传非参数MDL-BW算法推断出的模型参数个数，能使算法性能进一步提升，同时避免了模型参数个数的预先盲目设置.

4.3 人体动作数据分析

为进一步验证本文提出算法对模型结构优化的能力，本文利用人体动作数据进行分析. 相对语音数据而言，人体动作数据本质上也是时间序列. 文献[12]指出对于人体动作可以用状态空间模型建模，例如HMM. 本文采用的人体动作数据为MSRAction3D数据集[13]. 该数据集是关于人体动作的三维骨架数据. 一个动作在某一时刻的姿态是由20个关节点采样构成，每个关节点由3维空间直角坐标组成. 骨架数据的可视化如图5，可以看到人体骨架由20个节点组成，因此，第n个动作在第t时刻的采样$\boldsymbol{x}_{n,t}$的特征维度$D=60$.

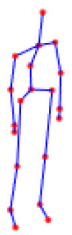

图5 可视化骨架数据

本文采用该数据集中的8种动作作为训练集，非训练集中的参与者所做的对应动作作为测试集. 动作以数字标签进行分类，选取的动作类型为“horizontal arm wave”、“hammer”、“forward punch”、“high throw”、“hand clap”、“bend”、“tennis serve”和“pickup throw”. 对应的数字标签为2，3，5，6，10，13，18和20. 为了方便动作类别叙述，以下皆用数字标签替代动作类型. 用于仿真的动作训练集样本个数$N=120$，对应的测试集动作样本个数$N=106$，每个动作序列的长度T介于30至100不等. 对于人体动作数据的建模，遗传非参数MDL-BW算法的初始值设置如下，$\beta=0.013$，$e=0.4$，num_pop=10，num_off=6，Max_S=8以及Max_C=5. 各个动作的识别准确率如表3.

表3 基于遗传非参数MDL-BW算法的动作识别准确率

动作类别	状态数	高斯混合数	准确率/%
2	7	5	91.67
3	6	4	33.33
5	4	5	72.73
6	7	5	45.45
10	7	5	93.33
13	6	5	100.00
18	7	4	100.00
20	7	5	66.67

从表3可以得出遗传非参数MDL-BW算法在人体动作数据上的平均识别率达77.36%. BW算法、基于似然函数GA算法[4]以及GA-BW算法[5]在统一预设下不同状态数和高斯混合数为5条件下的识别准确率如表4. 表4的3种同类算法的识别准确率都低于本文方法的识别准确率77.36%,说明本文方法相较同类算法对HMM结构的寻求更为有效. 在参数值的求解上,可以看到不同状态数下的BW算法的识别准确率整体低于基于似然函数的GA算法、GA-BW算法以及本文方法的平均识别率77.36%. 这主要是因为后三者通过在GA算法随机搜索的框架下,将解的搜索范围扩大,减小了参数初值对参数求解的影响. 此外,本文方法相较基于似然函数的GA算法以及GA-BW算法,利用自适应重组操作进一步扩大了解的搜索空间.

表4　不同算法在统一预设高斯混合数为5以及不同状态数下的识别准确率　　单位:%

方法	状态数				
	4	5	6	7	8
BW算法	73.58	70.75	66.98	66.98	71.70
基于似然函数GA算法[4]	70.75	75.47	71.70	69.81	69.81
GA-BW算法[5]	71.70	70.75	76.42	74.53	72.64

另外,模型参数个数是影响模型效果的重要因素. 由表3可以看到,遗传非参数MDL-BW算法对每一类动作的HMM推断出适合模型的状态数和高斯混合数,而表4中各个算法的状态数是对所有动作类别模型的统一设定. 由准确率可以看到,统一预设参数个数是很难较快找到合适的模型参数个数,需要多次重复设定不同参数个数进行探寻. 而遗传非参数MDL-BW算法依据MDL值减小的方向自动推断模型参数个数,避免了盲目预设模型参数个数的问题.

5　MDL准则对模型的影响

式(5)所示的MDL准则由数据的对数似然概率值的负值和模型参数个数编码长度两部分组成. 以MDL作为优化函数实质上是在模型拟合数据的程度和模型复杂度之间做平衡. 目标是希望训练出来的模型能很好地拟合数据但同时又要防止过拟合. 但是,两者的比例如何分配是个值得商榷的问题. 为此,本文进一步将人体动作数据动作识别率与参数β的关系显示出来,如图6所示. 由图6可知,当$\beta=0$时,即优化函数仅由数据的对数似然值决定时,动作识别率为66.98 %. 随着β值的增大,动作识别率逐渐提高,而当$\beta>0.015$时,识别率开始下降. 可以看到,对模型做一定的约束是有利于提高模型的准确度,但当模型参数个数编码长度项的比例过大时,同样会影响模型性能. β的作用用于调节

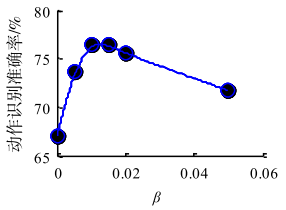

图6　动作识别准确率与参数β的关系

MDL第二项对于模型的影响程度. 而对于β值的如何选取可做如下考虑,由每个高斯分布的参数个数$L=D+D(D+1)/2$可知,式(5)的第二项关于数据特征维度D的计算复杂度为$O(D^2)$,第二项其余参数对此项计算复杂度要小于参数D. 所以在MDL准则中,D的大小是影响第二项模型复杂度大小的主要因素. 另外,一个准确的模型结构是在模型拟合数据的程度和模型复杂度之间做平衡,即在MDL的第一项值和第二项值之间做平衡. 由于第一项是个似然概率对数项,模型一般都会追求一个尚可的似然概率值,第一项值会较为稳定. 此时,为达到平衡效果,不妨将第二项值也视为一个基本的固定值. 对于固定的第二项,β值可以根据仿真实际使用的数据特征维度D的大小来选择. 当D较大时,可以先从较小的数量级来设定β值;当D较小时,可以先从较大的数量级设定β值. 根据识别准确率,确定β值的大概数量级,再进行微调来确定β值. 例如,本文使用的动作数据的特征维度$D=60$,可先将β确定在10^{-2}数量级上,再进行微调确定具体的β值,如图6对β进行微调,确定在人体动作识别仿真中β取值在0.01左右较为合适.

6　结论

HMM作为典型的概率图模型,广泛用于语音信号处理、基因组测序以及动作识别等应用领域. 本文提出遗传非参数MDL-BW算法来优化HMM的结构. 遗传非参数MDL-BW算法通过GA算法随机搜索的特性以及结合自适应重组操作保证了算法在更大的搜索空间中寻求全局参数解,通过非参数变异操作改变模型参数个数,引入MDL准则作为优化准则,帮助寻求合适的模型结构. 相较同类算法,遗传非参数MDL-BW算法不仅能对模型参数值进行更全局地优化,还能对模型的参数个数进行推断,避免了盲目设置参数个数的问题,这也给非参数模型选择方法带来了一定启发. 对模型做约束,MDL只是众多优化准则中的一种. 为了获得更佳性能,今后可以利用其他的非参数模型选择准则[14]作为优化准则,或从贝叶斯观点引入非参数贝叶斯方法[15]对模型做约束,进一步还可结合其他性能更好的生物学优化方法进行改进.

参考文献

[1] 刘建伟, 黎海恩, 等. 概率图模型的表示理论综述[J]. 电子学报, 2016, 44(5): 1219-1226.
LIU Jian-wei, LI Hai-en, et al. A survey on the representation theory of probabilistic graphical models[J]. Acta Electronica Sinica, 2016, 44(5): 1219-1226. (in Chinese)

[2] SUNDARARAJAN P K. Improving the Performance and Understanding of the Expectation Maximization Algorithm: Evolutionary and Visualization Methods[D]. Pittsburgh, PA: Carnegie Mellon University, 2016.

[3] 张西宁, 雷威, 等. 采用自适应基因粒子群算法优化隐马尔科夫模型的方法及应用[J]. 西安交通大学学报, 2018, 52(8): 1-8.
ZHANG Xining, LEI Wei, et al. Adaptive genetic particle swarm algorithm for optimization hidden Markov models with applications[J]. Journal of Xi'an Jiaotong University, 2018, 52(8): 1-8. (in Chinese)

[4] BENMACHICHE A, MAKHLOUF A, BOUHADADA T. Evolutionary learning of HMM with Gaussian mixture densities for automatic speech recognition[C]//Proceedings of the 9th International Conference on Information Systems and Technologies(ICIST). Cairo, Egypt: ACM, 2019: 1-6.

[5] MAKHLOUF A, LAZLI L, BENSAKER B. Evolutionary structure of hidden Markov models for audio-visual Arabic speech recognition[J]. International Journal of Signal and Imaging Systems Engineering, 2016, 9(1): 55-66.

[6] 崔佳旭, 杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10): 3068-3090.
CUI Jia-xu, YANG Bo. Survey on Bayesian optimization methodology and applications[J]. Journal of Software, 2018, 29(10): 3068-3090. (in Chinese)

[7] 杨扬. 基于最小描述长度的大规模图数据结构分析[D]. 合肥: 中国科学技术大学, 2018.
YANG Yang. Structure Analysis of Large Graph Data Based on Minimum Description Length[D]. Hefei: University of Science and Technology of China, 2018. (in Chinese)

[8] 王为凯. 基于GMM-HMM的声学模型训练研究[D]. 广州: 华南理工大学, 2016.
WANG Weikai. Rearch of the GMM-HMM Based Acoustic Models[D]. Guangzhou: South China University of Technology, 2016. (in Chinese)

[9] 李航. 统计学习方法[M]. 北京:清华大学出版社, 2012: 181-183.

[10] AMSALU S B, HOMAIFAR A. Driver behavior modeling near intersections using hidden Markov model based on genetic algorithm[C]//IEEE International Conference on Intelligent Transportation Engineering(ICITE). Singapore: IEEE, 2016: 193-200.

[11] BISHOP C M. Pattern Recognition and Machine Learning [M]. New York: Springer, 2006: 627-629.

[12] LEHRMANN A M, GEHLER P V, NOWOZIN S. Efficient nonlinear markov models for human motion[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Columbus, OH, USA: IEEE, 2014: 1314-1321.

[13] LI Wanqing, ZHANG Zhengyou, LIU Zicheng. Action recognition based on a bag of 3D points[C]//Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops(CVPRW). San Francisco, CA, USA: IEEE, 2010: 9-14.

[14] SOSIAWAN A Y, NOORAENI R, SARI L K. Implementation of using HMM-GA in time series data[J]. Procedia Computer Science, 2021, 179: 713-720.

[15] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: a review for statisticians[J]. Journal of the American Statistical Association, 2017, 112(518): 859-877.

作者简介

徐佳伟 男,1996年12月出生于湖南省株洲市.现为北京信息科技大学硕士研究生.主要研究方向为机器学习.
E-mail: wangyiemail163@163.com

罗 倩 女,1965年12月出生于山西省太原市.现为北京信息科技大学副教授.主要研究方向为信号与信息处理,大数据处理.
E-mail: luoqian@bistu.edu.cn