基于遗传算法的动态 Bayesian 网结构学习的研究

王 飞[1,2],刘大有[3],卢奕南[3],虞强源[3]
(1. 复旦大学计算机科学与工程系,上海 200433;2. 复旦大学智能信息处理开放实验室,上海 200433;
3. 吉林大学计算机科学与技术学院,吉林长春 130023)

摘 要: 动态 Bayesian 网是复杂随机过程的图形表示形式,从数据中学习建造动态 Bayesian 网是目前的研究热点问题. 本文针对该问题提出了一种遗传算法. 文中设计了结合数学期望的适应度函数,该函数利用进化过程中的最好动态 Bayesian 网把不完备数据转换成完备数据,使动态 Bayesian 网的学习分解为两个 Bayesian 网(初始网和转换网)的学习,简化了学习的复杂度. 此外,文中给出了网络结构的编码方案,设计了相应的遗传算子. 模拟实验结果表明,该算法能有效地从不完备数据序列中学习动态 Bayesian 网,并且实验结果说明了隐藏变量的作用和遗传控制参数对结果模型的影响.

关键词: 动态 Bayesian 网; 不完备数据; 数学期望; 遗传算法
中图分类号: TP301.6 **文献标识码**: A **文章编号**: 0372-2112 (2003) 05-0698-05

Research on Learning Dynamic Bayesian Networks by Genetic Algorithms

WANG Fei[1,2], LIU Da-you[3], LU Yi-nan[3], YU Qiang-yuan[3]
(1. *Dept. of Computer Science & Engineering, Fudan University, Shanghai* 200433, *China*;
2. *Lab of Intelligent Information Processing, Fudan University, Shanghai* 200433, *China*;
3. *College of Computer Science and Technology, Jilin University, Changchun* 130023, *China*)

Abstract: Dynamic Bayesian networks are a representation for complex stochastic processes. How to learn structure of Dynamic Bayesian networks from data is a hot problem of research. An evolutionary algorithm is proposed. Fitness function based on expectation is presented to convert incomplete data to complete data utilizing current best dynamic Bayesian network of evolutionary process. Thus dynamic Bayesian networks can be learned by using two Bayesian networks, prior network and transition network, to reduce the computational complexity. Encoding is given, and genetic operators are designed which provides guarantee of convergence. Experimental results not only show this algorithm can be effectively used to learn Dynamic Bayesian networks structure from incomplete data sequences, but also illustrate the role of hidden variables and the influence of genetic control parameters on learned model.

Key words: dynamic bayesian networks; incomplete data; genetic algorithms; mathematic expectation

1 引言

动态 Bayesian 网(DBN, Dynamic Bayesian networks)表示随时间变化的复杂随机过程,可以应用于语音识别[1]、高速公路监控、股票市场指数反应、生物进化过程、病人健康状况监控[2,3]等很多方面. 通常,几乎没有专家能够给出动态随机过程的模型,从数据中学习是一种可行的建模方法.

当训练数据完备时,由于评价动态 Bayesian 网"优劣"的适应度函数可以分解为评价关于初始网的分值和关于转换网的分值两部分,因此动态 Bayesian 网的学习可以分解为两个 Bayesian 网(初始网和转换网)的学习[2,3]. 实际中常常不能完全观察到所要建模的随机过程中的所有属性,即训练数据不完备,这时评分函数不能分解,因此从不完备数据中学习动态 Bayesian 网比从完备数据中学习困难得多. 算法[2,3]扩展学习 Bayesian 网的 EM 算法[4]以学习动态 Bayesian 网,EM 算法是确定性搜索算法,由于可能的动态 Bayesian 网结构组成的搜索空间非常巨大且具有多个局部极值,因此扩展的 EM 算法不可避免地陷入局部极值.

本文扩展文献[5]中学习 Bayesian 网的 EGA 算法,提出了一种从不完备数据中学习动态 Bayesian 网的遗传算法(简称 EGA-DBN 算法). 概括地说,该算法提出的适应度函数基于数学期望思想,由算法执行过程中的当前最优动态 Bayesian 网把不完备数据转换成完备数据,使得评价动态 Bayesian 网的适应度函数可以分解为关于初始网的适应度和转换网的适应

收稿日期:2001-07-30;修回日期:2002-12-30
基金项目:国家 863 高技术项目(No. 863-306-ZD05-01-2);国家自然科学基金(No. 69883003);教育部高校博士点专项科研基金项目;教育部符号计算与知识工程重点实验室的资助

度两部分,从而能够利用完备数据学习的优点.并且该算法属于随机搜索类算法,避免了收敛到次优模型的问题.

2　问题定义

首先说明文中不同变量的含义.如果没有特殊说明,大写字母 X, Y, Z 表示随机变量;小写字母 x, y, z 表示随机变量的某个取值;大写黑体字母 X, Y, Z 表示随机变量的集合.

假定 $\mathbf{X} = \{X_1, \cdots, X_n\}$ 是时态过程中发生变化(或进化)的属性集,$X_i[t]$ 表示属性 X_i 在 t 时间点对应的随机变量,$\mathbf{X}[t]$ 表示属性集 X 在 t 时间点对应的随机变量集.

一个动态 Bayesian 网由初始网 B_0 和转换网 $B_{\rightarrow}$ 两个 Bayesian 网组成.初始网 B_0 指定时态过程初始状态的概率分布 $p(x[0])$,转换网 $B_{\rightarrow}$ 对所有时间点 0,1, …, t 指定从 $t-1$ 时间点到 t 时间点属性集状态的转换概率 $p(x[t] \mid x[t-1])$.

图 1(a) 给出了一个动态 Bayesian 网的简单例子.

一个动态 Bayesian 网定义了在动态随机过程中无穷变化轨迹上的概率分布.实际,我们一般只在有穷时间间隔 0,1, …, T 上推理,那么可以把一个动态 Bayesian 网展开成在 $\mathbf{X}[0], \mathbf{X}[1], \cdots, \mathbf{X}[T]$ 上的"长"Bayesian 网.图 1(b) 给出了图 1(a) 所示的动态 Bayesian 网展开 3 个时间片的相应 Bayesian 网.

给定动态 Bayesian 网 $B = (B_0, B_{\rightarrow})$,在 $\mathbf{X}[0], \mathbf{X}[1], \cdots, \mathbf{X}[T]$ 上的联合概率分布可以通过初始网和转换网指定的概率分布简化表示:

$$p_B(x[0], \cdots, x[T]) = p_{B_0}(x[0]) \prod_{t=1}^{T} p_{B_{\rightarrow}}(x[t] \mid x[t-1]) \tag{1}$$

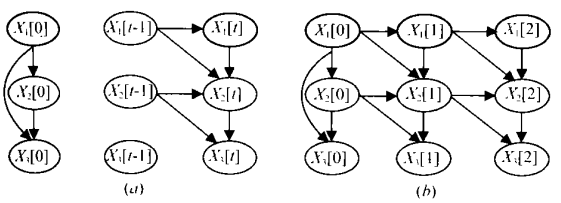

图 1　(a) 定义关于属性 X_1, X_2, X_3 的 DBN 的初始网和转换网,(b) 相应的"展开"Bayesian 网

通常,几乎没有专家能够给出动态随机过程的模型,从数据中学习是一种可行的建模方法.从数据中学习动态 Bayesian 网实际是寻找和训练序列集匹配度最高的动态 Bayesian 网.因为动态 Bayesian 网和 Bayesian 网有着非常密切的联系:一个动态 Bayesian 网由两个 Bayesian 网定义,并且一个动态 Bayesian 网可以展开成一个"长"的 Bayesian 网,所以可以考虑扩展 Bayesian 网的学习算法用于动态 Bayesian 网的学习.但是,动态 Bayesian 网的学习并不等同于把动态 Bayesian 网展开成"长"Bayesian 网的学习,因为展开后的 Bayesian 网中含有重复的网络结构和重复的概率参数,如图 1(b) 所示,没有必要学习重复的内容,增加复杂度.扩展 Bayesian 网的学习算法处理动态 Bayesian 网学习的做法是从训练序列中学习到两个 Bayesian 网结构:初始网 B_0 和转换网 $B_{\rightarrow}$,通常做法同样是引入一个评分函数,评价初始网 B_0 和转换网 $B_{\rightarrow}$ 定义的动态 Bayesian 网反映训练序列的准确度,然后由搜索算法寻找最好的动态 Bayesian 网,即寻找最好的初始网 B_0 和转换网 $B_{\rightarrow}$ 的组合.

输入:训练序列集 D,D 中包含 M_{num} 个观察序列,第 l 个序列的长度是 M_l 且给出了 $x^l[0], x^l[1], \cdots, x^l[M_l]$,其中的每个 $x^l[0], x^l[1], \cdots, x^l[M_l]$ 相当于一个事例.

输出:与训练序列集匹配度最高的动态 Bayesian 网

3　EGA-DBN 算法

遗传算法是模拟生物界自然进化和遗传过程的随机搜索算法,适于解决"评分-搜索"类问题.基于遗传算法的 EGA-DBN 可形式化描述为 $EGA\text{-}DBN = (F, \Delta, \odot, \nabla, \lambda, P_{init}, \tau)$,其中 F 表示适应度函数,$\Delta, \odot, \nabla$ 分别表示选择、交叉、变异算子,λ 表示群体规模,P_{init} 表示初始群体,τ 表示算法终止条件.

3.1　适应度函数

一个动态 Bayesian 网可分解成两个 Bayesian 网,那么评价动态 Bayesian 网的适应度函数可以分解成评价两个 Bayesian 网优劣的适应度函数.

数据完备时,评价 Bayesian 网的 MDL/BIC 函数和 BDe 函数均可分解为关于每个家族局部结构(一个变量和其所有父亲节点)的独立因式,如果局部家族结构发生变化不会对其余局部结构的适应度产生影响,也就是评价函数的计算只要统计出关于每个局部结构的充分统计因子即可[5].

数据不完备时,评价 Bayesian 网优劣的函数不存在可分解性,这就使得:(1) 必须执行推理过程计算待评判的网络结构分值,此推理过程是 NP 难度的;(2) 为给网络结构配置上最佳参数,必须利用最大梯度法[6]或 EM(expectation-maximum) 算法[7]执行非线性的优化过程;(3) 网络结构的局部改动,将影响该网络其余局部结构的评估.

为使得能够从不完备数据中学习,必须降低评价函数的计算复杂度,为此设法把不完备数据转换成完备数据以利用函数的可分解性.我们利用了数学期望的概念,采用遗传进化过程中到目前为止"最好"的网络结构 $B^* = (B_0^*, B_{\rightarrow}^*)$ 作为候选网络结构,在该网络结构中进行推理把不完备数据完备化形成新的期望完备数据,在该期望完备数据中计算事件发生的期望充分统计因子.

式(2) 给出了在训练集 D 和动态 Bayesian 网 $B^* = (B_0^*, B_{\rightarrow}^*)$ 条件下评价动态 Bayesian 网 $B = (B_0, B_{\rightarrow})$ 的适应度函数.其中的 $\sum_i \sum_j \sum_k N_{i,j',k}^0 \log\theta_{i,j',k}^0$ 衡量初始网 B_0 表达训练序列集 D 中各个观察序列初始状态的准确度,$\sum_i \sum_j \sum_k N_{i,j,k}^{\rightarrow} \log\theta_{i,j,k}^{\rightarrow}$ 刻画转换网 $B_{\rightarrow}$ 表达各个观察序列中动态随机过程状态转换的准确度,$\sum^{\log M_{num}/2} \| \pi_{X_i}^0 \| (\| X_i \| - 1)$、$\sum^{\log M/2} \| \pi_{X_i}^{\rightarrow} \| (\| X_i \| - 1)$ 分别是关于初始网 B_0 和转换网 $B_{\rightarrow}$ 结构复杂度的惩罚因子.因为不含有任何独立关系的全链接网络结构反映数据样本的准确度最高,但是通常没有什么意义,并且越简单的结构用于预测和推理的计算复杂

度越低,所以适应度函数中应体现出网络结构的简洁度,引入惩罚因子,结构越复杂,惩罚越"严厉",并且数据量愈多,愈倾向于简单结构.

式(2)适应度函数实质是关于初始网 B_0 的适应度和转换网 $B_{\rightarrow}$ 的适应度之和:

$$Fitness(B^{*}, D) = Fitness(B_0^{*}, D_0) + Fitness(B_{\rightarrow}^{*}, D_{\rightarrow})$$

其中 D_0 表示由 D 的各个观察序列中关于初始状态的事例组成的数据集,$D_{\rightarrow}$ 表示由 D 的各个观察序列中关于动态随机过程状态转换的事例组成的数据集.

$$Fitness(B^{*}, D) = \sum_i \Big(\sum_{j'} \sum_{k'} N^0_{i,j',k'} \log \theta^0_{i,j',k'} - \frac{\log M_{num}}{2} \| \pi^0_{X_i} \| (\| X_i \| - 1) \Big) + \sum_i \Big(\sum_j \sum_k N^{\rightarrow}_{i,j,k} \log \theta^{\rightarrow}_{i,j,k} - \frac{\log M}{2} \| \pi^{\rightarrow}_{X_i} \| (\| X_i \| - 1) \Big)$$

其中 $\pi^0_{X_i}, \pi^{\rightarrow}_{X_i}$ 分别表示 X_i 在初始网 B_0 和转换网 $B_{\rightarrow}$ 中的父亲节点集,$M = \sum_l M_l$.

$$\theta^0_{i,j',k'} = N^0_{i,j',k'} / \sum_{k'} N^0_{i,j',k'}$$

$$N^0_{i,j',k'} = \sum_{l=1}^{M_{num}} p(x_i^{k'}, \pi^{0(j')}_{X_i} | y^l[0], \theta_{B_0^{*}}, B_0^{*})$$

其中 $y^l[0]$ 表示训练集 D 中第 l 个观察序列的第 0 个事例,$x_i^{k'}$ 表示 $val^{k'}(X_i)$,$\pi^{0(j')}_{X_i}$ 表示 $val^{j'}(\pi(X_i[0]))$。

$$p(x_i^{k'}, \pi^{0(j')}_{X_i} | y^l[0], \theta_{B_0^{*}}, B_0^{*}) = \begin{cases} 0, \text{如果在 } y^l[0] \text{ 中 } \pi^0_{X_i} = \pi^{0(j')}_{X_i} \text{ 且 } X_i = x_i^{k'} \\ 1, \text{如果在 } y^l[0] \text{ 中 } \pi^0_{X_i} \neq \pi^{0(j')}_{X_i} \text{ 且 } X_i \neq x_i^{k'} \\ \text{在 Bayesian 网} \langle B_0^{*}, \theta_{B_0^{*}} \rangle \text{中推理计算 } p(x_i^{k'}, \pi_i^{0(j')}), \text{否则} \end{cases}$$

$$\theta^{\rightarrow}_{i,j,k} = N^{\rightarrow}_{i,j,k} / \sum_k N^{\rightarrow}_{i,j,k}$$

$$N^{\rightarrow}_{i,j,k} = \sum_{l=1}^{M_{num}} \sum_{t=1}^{M_l} p(x_i^k, \pi^{\rightarrow(j)}_{X_i} | y^l[t], \theta_{B_{\rightarrow}^{*}}, B_{\rightarrow}^{*})$$

其中 $y^l[t]$ 表示训练集 D 中第 l 个观察序列的第 $t(t \neq 0)$ 个事例,

$\pi^{\rightarrow(j)}_{X_i}$ 表示 $val^{j}(\pi(X_i[t]))$.

$$p(x_i^k, \pi^{\rightarrow(j)}_{X_i} | y^l[t], \theta_{B_{\rightarrow}^{*}}, B_{\rightarrow}^{*}) = \begin{cases} 0, \text{如果在 } y^l[t] \text{ 中 } \pi^{\rightarrow}_{X_i} = \pi^{\rightarrow(j)}_{X_i} \text{ 且 } X_i = x_i^k \\ 1, \text{如果在 } y^l[t] \text{ 中 } \pi^{\rightarrow}_{X_i} \neq \pi^{\rightarrow(j)}_{X_i} \text{ 且 } X_i \neq x_i^k \\ \text{在 Bayesian 网} \langle B_{\rightarrow}^{*}, \theta_{B_{\rightarrow}^{*}} \rangle \text{中推理计算 } p(x_i^k, \pi^{\rightarrow(j)}_{X_i}), \text{否则} \end{cases} \quad (2)$$

式(2)中的 $B^{*} = (B_0^{*}, B_{\rightarrow}^{*})$ 表示当前从 D 中学习到的最佳动态 Bayesian 网,$N^0_{i,j',k'}$ 表示 $\pi^0_{X_i} = \pi^{0(j)}_{X_i}$ 且 $X_i = x_i^k$ 基于 B_0^{*} 在 D 中出现的期望充分统计因子,$N^{\rightarrow}_{i,j,k}$ 表示 $\pi^{\rightarrow}_{X_i} = \pi^{\rightarrow(j)}_{X_i}$ 且 $X_i = x_i^k$ 基于 $B_{\rightarrow}^{*}$ 在 D 中出现的期望充分统计因子.

3.2 参数编码

因为 ECA-DBN 算法的目的是寻找最好的动态 Bayesian 网,动态 Bayesian 网结构应被编码成染色体. 因为一个动态 Bayesian 网由两部分组成 $B = (B_0, B_{\rightarrow})$,那么相应的染色体也由两部分表示 $C = (C_0, C_{\rightarrow})$.

Bayesian 网结构 $B_0, B_{\rightarrow}$ 可以通过记录每个变量的父亲节点集来表示,因此 $B_0, B_{\rightarrow}$ 可看作由关于每个变量的局部家族结构(即父亲节点集)组合在一起的,我们可以把各个局部家族编码成基因,组成染色体 $C = (C_0, C_{\rightarrow})$.

基于上述思想,我们把 $B_0, B_{\rightarrow}$ 编码成邻接的表链,表链中每一个位置对应一个基因,记录一个变量的父亲节点集. 图 2 给出了图 1 的编码结果.

C_0		$C_{\rightarrow}$	
$X_1[0]$		$X_1[t]$	$X_1[t-1]$
$X_2[0]$	$X_1[0]$	$X_2[t]$	$X_1[t]\ X_1[t-1]\ X_2[t-1]$
$X_3[0]$	$X_2[0]\ X_1[0]$	$X_3[t]$	$X_2[t]\ X_2[t-1]$

图 2 DBN 的编码

图 2 中 $C_0, C_{\rightarrow}$ 的每一行表示一个基因,如转换网 $B_{\rightarrow}$ 中 $X_2[t]$ 的父亲节点集是 $X_1[t]\ X_1[t-1]\ X_2[t-1]$,相应个体 $C_{\rightarrow}$ 中关于 $X_2[t]$ 的基因编码为 $X_1[t]\ X_1[t-1]\ X_2[t-1]$. C_0 中的第 i 行表示属性 X_i 的初始状态 $X_i[0]$ 的局部家族结构. 特别需要说明的是,转换网 $B_{\rightarrow}$ 中属性 X_i 在 $t-1$ 时刻对应的随机变量 $X_i[t-1]$ 没有父亲节点,因此 $C_{\rightarrow}$ 中没有对应于 $X_i[t-1]$ 局部家族结构的基因,第 i 行表示属性 X_i 在 t 时刻对应的随机变量 $X_i[t]$ 的局部家族结构,表示出属性 X_i 在动态随机过程中的变化规律,它不仅可能和 $t-1$ 时刻的属性状态相关,还可能和 t 时刻的属性状态相关.

这种编码的优点之一是染色体中的一个基因对应一个局部家族结构,也对应着评价该染色体的适应度函数分解后的一个独立因式,如果在遗传进化过程中,该染色体只有某个基因发生变化,那么在从完备数据中学习时,只需要重新计算发生变化的基因的局部适应度值,从不完备数据中学习时,如果用于计算期望充分统计因子的候选网络结构没有变化,那么也只需要重新计算发生变化的基因的局部适应度值.

3.3 遗传操作设计

3.3.1 选择 我们采用等级比例法对个体进行选择. 该方法按适应度大小将个体分成不同的等级,每个等级的选择概率不同. 这样,选择概率和个体适应度的等级有关,和适应度的绝对大小无关,避免了超常个体选择概率过大,出现早熟现象. 如果 S_t^j 表示第 t 代第 j 个个体,$rank(F(S_t^j))$ 表示 S_t^j 适应度的等级,那么 S_t^j 被选择的概率 $p_{j,t}$ 为:$p_{j,t} = \frac{rank(F(S_t^j))}{\lambda(\lambda+1)/2}$,其中 λ 表示初始群体规模.

3.3.2 交叉 一个动态 Bayesian 网的编码由初始网和转换网中各个局部结构的对应基因组成,那么改变个体中的基因就改变了相应的局部结构,进一步局部结构的变化导致整体动态 Bayesian 网结构的改变.

交叉算子是将两个网络结构中关于同一个变量的局部结构进行交叉,使得不同网络结构中各自较好的局部结构有机会组合在一起进化出更好的个体.

选择操作选出的个体随机配对,并按概率 p_c 决定是否进行交叉操作. 我们采用多个点分别进行单点交叉的方法[5],也就是在配对的一对个体中随机选择多个变量,针对这多个变

量的局部结构按照概率 p_c 进行交叉操作.

3.3.3 变异

变异操作是随机改变个体的性状,增加群体的多样性以避免局部极值,具体做法是按变异概率 p_m 随机改变某些个体串的某些基因座的基因值. EGA-DBN 算法中基因表示一个变量的父亲节点集,改变一个基因即改变一个变量的父亲节点集,可以使整个网络结构发生变化,从而网络结构可以从一个局部区域跳到另一个局部区域,有可能搜索到全局最优解.

EGA-DBN 算法中个体 S 的基因 $Gene(X_i[t])$ 表示变量 $X_i[t]$ 在 S 中的父亲节点集,为改变一个变量的父亲节点集,定义下面三个基本的变异算子:为 $X_i[t]$ 增加一个父亲节点、删除 $X_i[t]$ 的一个父亲节点、逆转有向弧 $X_j[t] \rightarrow X_i[t]$,但是 $t-1$ 时刻的随机变量 $X_j[t-1]$ 到属性 t 时刻的随机变量 $X_i[t]$ 的有向弧 $X_j[t-1] \rightarrow X_i[t]$ 不能被逆转,只可以采取删除操作.

增加 $AddParent(X_i[0])$:

$$Gene(X_i[0]) = Gene(X_i[0]) + Ran\_sel(X[0] - \{X_i[0]\} - \{X_j[0] \mid X_j[0] \in Gene(X_i[0])\})$$

$AddParent(X_i[t])$:

$$Gene(X_i[t]) = Gene(X_i[t]) + Ran\_sel(X[t] + X[t-1] - \{X_i[t]\} - \{X_j[t'] \mid X_j[t'] \in Gene(X_i[0])\})$$

删除 $SelParent(X_i)$:

$$SelParen(X_i[t]) : Gene(X_i[t]) = Gene(X_i[t]) - Ran\_sel(Gene(X_i[t]))$$

逆转 $Reverse(X_j[t], X_i[t])$:

$$Gene(X_i[t]) = Gene(X_i[t]) - \{X_j[t]\};$$

$$Gene(X_j[t]) = Gene(X_j[t]) + \{X_j[t]\}$$

其中函数 $Ran\_sel(\cdot)$ 表示在集合 $\cdot$ 随机选择一个元素.

3.4 初始群体的设定

初始群体可以随机生成;或由专家依据先验知识给出[5].

3.5 设定控制参数

群体规模、遗传操作概率 p_c、p_m 及算法的结束条件根据算法处理的具体情况设定.

因为适应度的计算复杂度高,所以群体规模不能太大;另一方面,为提高群体中个体的多样性,避免未成熟收敛,群体规模不能太小,一般选在 10 到 100 之间.

变异概率 p_m 直接影响到算法的收敛性和最终解的性能. 变异概率大,会使得算法不断得搜索新的解空间,增加模式的多样性. 但较大的变异概率会影响算法的收敛性. 通常变异概率取一个较小的值.

遗传算法中,交叉算子因其全局搜索能力而作为主要算子,变异算子因局部搜索能力而作为辅助算子. 遗传算法通过交叉和变异这一对相互配合又相互竞争的操作而使其具备兼顾全局和局部的均衡搜索能力. 所谓相互配合,是指当群体在进化中陷于搜索空间中某个超平面而仅靠交叉不能摆脱时,通过变异操作可有助于这种摆脱;所谓相互竞争,是指当通过交叉已形成所期望的基因块时,变异操作有可能破坏这些基因块[8]. 本文的实验结果就交叉算子和变异算子的选择进行了一些讨论.

算法终止条件可以选做已经进化了 g_1 代或连续的 g_2 代最佳的网络结构没有变化.

4 实验结果

为检测 EGA-DBN 算法的学习效果,我们进行了实验,具体步骤如下:

(1) 选定一个动态 Bayesian 网 $B = (B_0, B_\rightarrow)$ (结构 + 条件概率分布),按照 S 反映的联合概率分布随机生成训练数据序列 D 和测试数据序列;

(2) 利用 EGA-DBN 算法从 D 中学习到动态 Bayesian 网 $B^* = (B_0^*, B_\rightarrow^*)$;

(3) 衡量 S^* 反映目标概率分布 P 的准确度以评价算法的优劣.

首先,选定的动态 Bayesian 网是描述高速公路上汽车状态的 BAT 网[9],含有 10 个状态变量,10 个观察变量和几个瞬时变量.

依据 BAT 网分别生成含 250、500、1000 个序列的训练数据集,训练数据集中仅包含 10 个观察变量.

对于每一个训练数据集事先分别引入 3 个、4 个、5 个和 6 个隐藏变量记录状态变量和瞬时变量,每种情形在没有任何先验知识的条件下分别利用 EGA-DBN 算法学习 10 次,选择 10 次学习中的最佳网络结构作为每一种情形学习的最终结果.

我们以结果模型反映目标概率分布的准确度评价结果模型的性能. 具体做法是:依据 BAT 网生成含 2000 个序列的测试数据集,其中仅包含 10 个观察变量,计算每一种情形下结果网络关于这个测试集的平均对数损失 (log-loss),即 $\frac{1}{N_{num}} \sum_{i=1}^{N_{num}} \log p(X^i[0], X^i[1], \cdots, X^i[N_i])$.

在实验中,具体的控制参数如下:

群体规模:$\lambda = 20$;

算法结束条件是已经进化了 1500 代或进化的 200

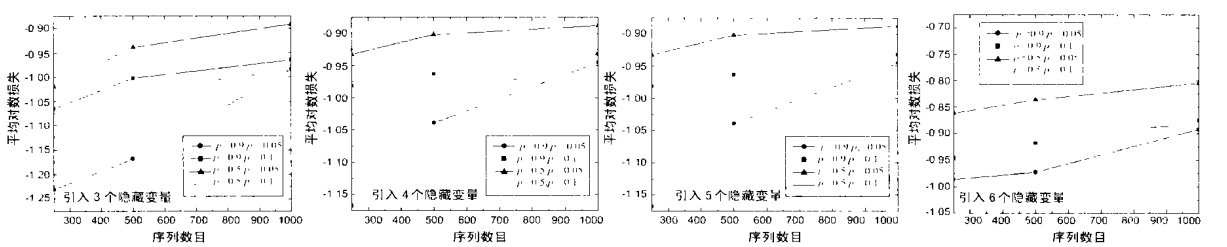

图 3　交叉概率和变异概率取值的比较

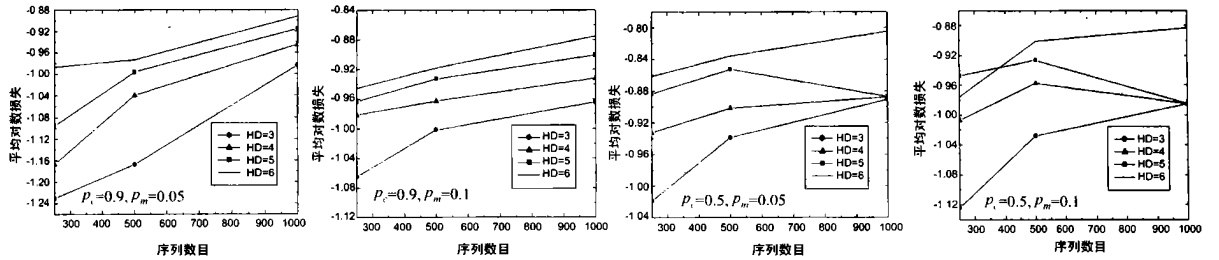

图 4 不同稳藏变量(HD)数目的比较

代最佳网络结构没有变化;

交叉概率和变异概率分别取做: $p_c = 0.9$, $p_m = 0.05$; $p_c = 0.9$, $p_m = 0.1$; $p_c = 0.5$, $p_m = 0.05$; $p_c = 0.5$, $p_m = 0.1$.

整体的实验结果如图 3 和图 4 所示.

从图中可以看出隐藏变量的增加提高了学习结果模型的预测准确度,并且训练数据序列越多,结果模型的预测结果越准确.

该实验说明 EGA-DBN 算法为解决复杂动态随机过程的学习提供了一种可行的方法.

参考文献:

[1] Zweig G, Russell S. Speech recognition with dynamic bayesian networks [A]. Proc. of AAAI-98 [C]. Madison, Wisconsin: AAAI Press, 1998.

[2] N Friedman, K Murkey, S Russell. Learning the structure of dynamic probabilistic networks [A]. Proc. of the 14th Conf on Uncertainty in Artificial Intelligence [C]. Madison, 1998. 139 - 147.

[3] X Boyen, N Friedman, D Koller. Discovering the hidden structure of complex dynamic systems [A]. Proc of the 15th Conf on Uncertainty in Artificial Intelligence [C]. Stockholm, Sewden, 1999.

[4] Friedman N. Learning belief networks in the presence of missing values and hidden variables [A]. Proc of the Fourteenth Inter. Conf. on Machine Learning (ICML) [C]. Madison, 1997. 452 - 459.

[5] 刘大有, 王飞, 等. 基于遗传算法的 Bayesian 网结构学习研究 [J]. 计算机研究和发展, 2001, 38(4): 916 - 922.

[6] John Binder, D Koller, S Russell, K Kanazawa. Adaptive probabilistic networks with hidden variables [J]. Machine Learning, 1997, 29: 213 - 244.

[7] Lauritzen S L. The EM algorithm for graphical association models with missing data [Z]. Computational Statistics and Data Analysis 19: 191 - 201.

[8] 阎平凡, 张长水. 人工神经网络与模拟进化计算 [M]. 北京: 清华大学出版社, 2000 年.

[9] Forbes J, T Huang, K Kanazawa, S Russell. The BATmobile: Towards a Bayesian automated taxi [A]. Proc. of 1995 Intel. Joint Conf. on Artificial Intelligence [C]. Montreal, Canada, 1995.

作者简介:

王 飞 女, 1975 年 10 月生于河南省开封市, 2001 年毕业于吉林大学计算机科学与技术学院, 获博士学位, 目前在复旦大学计算机科学与工程系和复旦大学智能信息处理开放实验室工作, 主要研究方向为机器学习、不确定知识处理、生物信息学等.

刘大有 男, 1942 年 7 月生于河北省乐亭县, 教授、博士导师, 现在吉林大学计算机科学与技术学院和吉林大学符号计算与知识工程教育部重点实验室工作, 主要研究方向为知识工程与 ES, 分布式 AI、多 Agent 系统与移动 Agent, 数据挖掘, 智能计算机辅助教学, 产品数据管理, 空间推理与 GIS 应用, 智能软件等.