针对时间序列多步预测的聚类隐马尔科夫模型

章登义,欧阳黜霏,吴文李
(武汉大学计算机学院,湖北武汉 430072)

摘 要: 时间序列的预测在现今社会各个领域中有着广泛的应用.本文针对时间序列趋势预测中的多步预测问题,提出了基于聚类的隐马尔科夫模型,利用隐马尔科夫模型中的隐状态来表示产生时间序列数据时的系统内部状态,实现对多步时间序列的预测.针对时间序列聚类中的距离计算问题,提出结合时间序列时间性和相似性的聚类算法,并给出了迭代精化基于聚类的隐马尔科夫模型的方法.实验表明,本文提出的方法在时间序列多步预测中精度较高.

关键词: 时间序列; 多步预测; 隐马尔科夫模型; 聚类

中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2014)12-2359-06

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.12.004

Cluster-Based Hidden Markov Model in Time Series Multi-Step Prediction

ZHANG Deng-yi, OUYANG Chu-fei, WU Wen-li
(*School of Computer*, *Wuhan University*, *Wuhan*, *Hubei* 430072, *China*)

Abstract: The study of time series prediction is pervasive in various fields. We propose a cluster-based hidden Markov model to approach the multi-step prediction problem in time series. As multi-step time series prediction problem is not fully addressed from a system angle, we utilize the hidden state of hidden Markov model to represent the inner state of a time series production system. We also promote a cluster algorithm combining the temporal and similarity criteria to address the distance calculating issue in time series clustering. This non-trivial criterion proves effective in multi-step time series prediction. Through a non-parameter approximate method we estimate the inner hidden state distributes from every single state. And we also prove the correctness of an iteratively refinement of the cluster-based hidden Markov model (HMM). Experimental results on authentic data indicate the effectiveness and accuracy of this approach.

Key words: time series; multi-step prediction; hidden Markov model (HMM); cluster

1 引言

随着大数据时代的到来,数据在各个行业和领域中日趋成为改变商业模式的研究重点,其中时间序列数据由于在金融、能源、气象等领域的集中产生成为研究热点.然而由于时间序列数据的复杂性和随机性,当前的时间序列建模方法[1~3]对于单步单个值的时间序列可以做出准确预测,但对于进行多步的时间序列预测方法不多.

已知的时间序列建模方法主要有以下几类:一类是以差分自回归移动平均模型(Auto Regressive Integrated Moving Average model, ARIMA)和线性回归模型(Linear Regression model, LR)为代表,通常假设预测中的时间序列趋势只与其最邻近的时间序列数据相关,针对时间序列数据的局部特征建模,因此不能准确的反应时间序列内部的生成机制.一类是以离散傅里叶变换(Discrete Fourier Transform, DFT)[4]和阶段线性表示(Piecewise Linear Representation, PLR)[5]为代表的基于整体时间序列数据建模的方法[6,7].主要用于精确表示原始时间序列数据,以进行快速模式匹配和模式发现,此类方法将时间序列对象看做离散对象,对时序间的时间性关联研究较少.还有一类是以发现时序数据中的频繁模式(motif)[8]为基础,但是此类方法挖掘出无意义的频繁模式,不利于解释时间序列系统的生产机制.

针对于此,本文提出了一种基于聚类的隐马尔科夫模型(Cluster-based Hidden Markov Model, CHMM),该模型假设时间序列系统是由一系列隐状态构成,而系统的运行本质就是在不同的隐状态间转换,首先运用 PLR 对

收稿日期:2014-03-19;修回日期:2014-06-19;责任编辑:覃怀银
基金项目:国家自然科学基金(No.60903035, No.41001296);国家高技术研究发展计划(863 计划)课题(No.2013AA12A301)

时间序列分段,然后基于时序性和相似性对分段进行聚类,将每个分类作为一个隐状态从而得到状态转移概率矩阵,依此建立关于分段的初始隐马尔科夫模型.基于这个初始模型对隐状态内的线段分布进行估计,同时利用迭代的方法不断精化初始模型直至得到最终预测模型.

2 相关工作

文献[9]中提出的可变长度的隐马尔科夫模型(Variable Length Hidden Markov Model,VLHMM),该模型弥补了一阶马尔科夫模型的固定状态数问题,能在最小的背景集下通过改变隐状态序列长度来建立高阶马尔科夫模型,然而此类模型的观测对象仍然是离散的时序点,易受噪声影响.文献[10]中提出的卡尔曼滤波(Kalman Filter,KF)算法结合了时间序列分析法,在传统的时间序列分析基础上,通过更新建模数据,将时间序列分析法与次优近似估计算法相结合,实现了动态递推建模方法.然而 KF 算法没有考虑时间序列观测对象的时间关联性.文献[11]提出一种利用高斯混合模型作为连续 HMM 状态中指定区域的预测专家.该方法首次引入了贝叶斯方法指导隐状态的预测,然而该方法没有解决隐状态分布估计的问题,隐状态的估计直接影响模式选择的精确性,另外估计中随机性较大也是影响精确性的因素之一,本文提出的方法将综合解决观测对象的离散性问题和隐状态的分布估计问题.

3 基于聚类的隐马尔科夫模型

研究首先利用 PLR 算法将观测序列分段,对所有分段按时序性关联和相似性关联聚类,从而得到初始的状态矩阵;基于 HDP-HMM(Hierarchical Dirichlet Process-Hidden Markov Model)模型对隐状态内的线段分布进行推测,利用初始状态矩阵初始化超参;最后利用动态规划思想迭代产生优化模型.

3.1 隐马尔科夫模型

隐马尔科夫模型是一个基于马尔科夫过程的随机模型,它包括一些列有穷状态集,以及与状态对应的观测序列,状态之间的转移是由状态转移概率矩阵来表示,由于实际观测到的是观测序列而不是状态本身,因此把从观测序列得到隐状态的过程叫做隐马尔科夫过程.

为了描述 HMM,需要定义以下符号或变量:

$S=\{1,2,\cdots,K\}$,S 表示状态集,K 表示状态的编号.

$\boldsymbol{A}=\{a_{ij}\}$,$1\leqslant i,j\leqslant K$,$\{a_{ij}\}$表示状态转移概率矩阵 $\boldsymbol{A}$ 元素的集合,a_{ij}表示从状态 i 转移到状态 j 的概率.

$\boldsymbol{E}=\{e_i(o)\}$,$1\leqslant i,j\leqslant K$,$\{e_i(o)\}$表示输出概率矩阵 $\boldsymbol{E}$ 元素的集合,o 表示一个连续或离散的观测值,e_i 表示在状态 i 下输出该观测值的概率.

$\Pi=\{\pi_i\}$,$1\leqslant i,j\leqslant K$,$\pi_i$ 表示时序的初始状态为 i 的概率.

因此一个隐马尔科夫模型可以用 $x=\{S,\boldsymbol{A},\boldsymbol{E},\Pi\}$ 来表示,给定一个 HMM 模型 x,观测序列 O 和状态序列 s,其产生概率公式如下:

$$P(O,s\mid x)=\pi_{s_1}e_{s_1}(o_1)\prod_{i=2}^{m}a_{s_{i-1},s_i}e_{s_i}(o_i) \tag{1}$$

针对本文的情况,给定一个时间序列 $T=\{t_1,t_2,\cdots,t_n\}$,如果要得到基于聚类的马尔科夫模型,则首先必须解决以下两个问题:

(1)将观测对象由时序点转变为线段序列,$L=\{l_1,l_2,\cdots,l_m\}$.

(2)以线段序列 L 为观测对象建立隐马尔科夫模型,则产生概率公式改写为:

$$P(L,s^*\mid x)=\pi_{s_1}e_{s_1}(L_1)\prod_{i=2}^{m}a_{s_{i-1},s_i}e_{s_i}(L_i) \tag{2}$$

当 $s^*=\{s_1,s_2,\cdots,s_m\}$是最优状态序列时,产生概率最大,其中 s_i 产生 L_i 的概率为 $e_{s_i}(L_i)$.

3.2 基于相似性关联和时间性关联的聚类

3.2.1 时序分段

如上所述,由于事先没有观测序列的信息,所以我们采用自底向上的 PLR 方法来对时序分段,首先将序列 $T=\{t_1,t_2,\cdots,t_n\}$分为$\lfloor n/2\rfloor$个线段,那么第 i 个线段 L_i 就表示 $t_{2i-1}t_{2i}$,然后迭代的合并相邻线段.每次迭代中合并相似性误差最小的两个相邻线段,直到最小相似性误差不超过一个设定的误差极值 ε_r,显然在没有得到隐状态分布的情况下这种初始的分段还需优化.

3.2.2 分段聚类

在得到初始分段后,我们将对分段进行聚类 $C=\{C_1,C_2,\cdots,C_k\}$.聚类的一个主要问题就是如何定义线段间的相似性标准,使用相似性误差或者最小描述长度作为衡量标准对于隐马尔科夫模型常常忽略了时间序列数据内在的时间联系,我们讨论两种聚类标准如下:

(1)相似性标准,与传统聚类所用标准一样,在本文中则是考虑将拥有类似形状(斜率和长度)的线段归为一类,不同类中的线段形状不同.因此同类线段间的相对误差公式如下,对于类 C_i 有:

$$E(i)=\frac{1}{|C_i|}\sum_{(l_j,\theta_j)\in C_i}\left\{\left(\frac{l_j-\bar{l}_i}{\bar{l}_i}\right)^2+\left(\frac{\theta_j-\bar{\theta}_i}{\bar{\theta}_i}\right)^2\right\} \tag{3}$$

其中$|C_i|$表示 C_i 中的线段数量,$\bar{l}$ 和$\bar{\theta}$ 表示类中的平均线段长度与斜率,$E(i)$越小那么 C_i 中的线段就越相似.

(2)时间性标准,如果 L_i 和 L_j 属于同一个类,那么 L_{i+1}和 L_{j+1}也应该有相同的分布,有较大概率属于同一个类.本文用熵来衡量这种不确定性,聚类 C_i 后跟随的是聚类 C_j 中线段的概率如下:

$$H(i)=\sum_{j=1}^{K}-p(j|i)\log p(j|i) \tag{4}$$

其中 $p(j|i)$表示 C_i 中线段后跟随的是 C_j 中线段的概率,并且 $H(i)$越小表示跟随的确定性越大.

本文采用贪心算法来结合两种标准.我们将时序分段后的线段 L_i 本身看作一个聚类 C_i,然后基于以上两个标准通过迭代方式合并相同的类直至得到最终聚类,基于时间性和相似性的时序聚类算法(Time-Series Clustering Algorithm Base on Temporal and Similarity Criterion):

算法 1　TSCABTSC 算法

```
输入:时间序列线段 L,误差阈值 ε
输出:聚类 C
For i = 1:length(L)
  min_dis = 0;
  For j = 1:length(L)
    If j ! = i
      dis = E(i,j); %式(3)
      If dis < min_dis
        best(i) = j;
        best_dis(i) = dis; min_dis = dis;
      end
    end
  If best_dis(i) < ε
    best_info(i) = H(i,best(i))   %式(4)
  else
    best(i) = 0; best_info(i) = 99;
  end
end
For i = 1:length(best)
  b = find(best_info == min(best_info));
  C = merge(b,best(b));
End
```

3.3　隐状态内分布估计

通过上述算法得到聚类 C 后,将聚类看作 K 个隐状态,则可以基于聚类则可以建立状态转移矩阵 $\boldsymbol{A}$,同时由式(2)变形可得:

$$\begin{aligned}P(L,s^*|x)&=\pi_{s_1}e_{s_1}(L_1)\prod_{i=2}^{m}a_{s_{i-1},s_i}e_{s_i}(L_i)\\&=\prod_{i=1}^{m}e_{s_i}(L_i)\cdot\pi_{s_1}\prod_{i=2}^{m}a_{s_{i-1},s_i}\\&=P_{发射}\cdot P_{转移}\end{aligned} \tag{5}$$

其中 $P_{发射}$表示 L 的在状态序列 s^* 中的发射概率,$P_{转移}$表示状态序列 s^* 的状态转移概率,现在由状态转移矩阵可求得状态转移概率 $P_{转移}$,因此下一步需要估计 L 的发射概率.假设线段 L 的斜率和长度是相互独立的,则对于 $L=(l,\theta)$其发射概率可写为:

$$e_i(L)=p(l|i)p(\theta|i) \tag{6}$$

其中 $p(l|i)$表示状态 i 下产生长度为 l 的线段的概率,$p(\theta|i)$表示状态 i 下产生斜率为 θ 的线段的概率.因此发射概率与线段的分布有关.

对每个聚类中线段形状的分布估计可以看作是隐状态内的序列的分布估计,现有方法通常直接假定为高斯分布等常用分布,但是这类方法受初始参数影响较大,估计精度低,而近年来非参数估计方法[12,13]大量用于对分布的估计,本文将结合隐马尔科夫模型的特点引入非参数贝叶斯模型[14]进行分布估计.

HDP-HMM[15]是一种基于隐马尔科夫模型的非参数贝叶斯估计,其中 stick-breaking 过程的概率图模型如图 1.

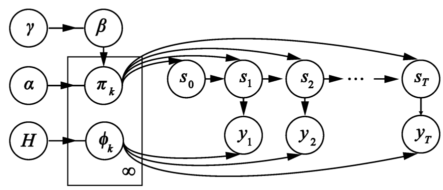


图1　图模型

$$\begin{gathered}\beta\sim\mathrm{GEM}(\gamma),\pi_k|\beta\sim\mathrm{DP}(\alpha,\beta),\phi_k\sim H,\\s_t|s_{t-1}\sim\mathrm{Multinomial}(\pi_{s_{t-1}}),y_t|s_t\sim F(\phi_{s_t})\end{gathered} \tag{7}$$

其中 α,β 和 γ 为用户定义的超参数,H 为基分布函数,观测数据 y_t 服从分布 $F(\phi_{s_t})$.根据式(7),对 HDP-HMM 采样,首先对 β 采样可得

$$\begin{gathered}p(\beta_1,\cdots,\beta_k,\beta_{\bar{k}}|T,K,y_1,\cdots,y_t,\gamma)\propto\\\mathrm{Dir}(m_1,\cdots,m_k,\gamma)\end{gathered}$$

m_k 表示对应每个状态中观测数据出现次数,对状态 s_t 采样可得

$$\begin{gathered}p(s_t|S^*_{-t},y_1,\cdots,y_T,\alpha,\beta,\gamma)\propto\\p(s_t|S^*_{-t},\alpha,\beta)p(y_t|Y_{-t},s_t,S^*_{-t},\gamma)\end{gathered} \tag{8}$$

在式(8)等号右边第二项 $p(y_t|Y_{-t},s_t,S^*_{-t},\gamma)$即为观测数据在隐状态中的分布概率,设分布函数 F 的分布密度函数为$f(\cdot|\theta)$,基分布函数 H 的分布密度函数为 $h(\cdot|\theta)$.将 Concentration 参数等条件省略即:

$$f_k^{-y_{ji}}(y_{ji})=f_k^{-y_{ji}}(y_{ji}|Y_{-ji},\alpha,\beta,\gamma) \tag{9}$$

由于 F 与 H 为共轭分布,因此利用贝叶斯公式将分布参数消去后可得观测数据的条件分布为:

$$f_k^{-y_{ji}}(y_{ji}) = \frac{\int f(y_{ji}|\phi_{s_t}) \prod_{j'i' \neq ji, s_{j'i'} = k} f(y_{j'i'}|\phi_{s_t}) h(\phi_{s_t}) \mathrm{d}\phi_{s_t}}{\int \prod_{j'i' \neq ji, s_{j'i'} = k} f(y_{j'i'}|\phi_{s_t}) h(\phi_{s_t}) \mathrm{d}\phi_{s_t}} \tag{10}$$

由于基分布函数 H 在初始化 CHMM 时已知，同时观测数据后验概率由式(8)可得，因此根据式(10)可以得到隐状态内的分布函数.

3.4 模型优化

通过上述方法我们建立了一个初始的 CHMM 模型，并得到了其隐状态内的分布.然而，本文在 3.2.1 节中的初始线段其分段精度不高，而模型中的隐状态序列又与分段精度相关，为此本文提出一种利用维特比算法来提高分段精度的迭代优化方法.

给定一个隐马尔科夫模型 x 和观测序列 O，通过维特比算法可以获取该模型的最优状态序列 s^*，表示如下：

$$\varphi_t(i) = \max_{s_1,\cdots,s_{t-1}} P(o_1,\cdots,o_t,s_1,\cdots,s_{t-1},s_t = i \mid x) \tag{11}$$

$$= \max_{s_1,\cdots,s_{t-1}} \pi_{s_1} e_{s_1}(o_1) \prod_{j=2}^{t} a_{s_{j-1},s_j} e_{s_j}(o_j) \tag{12}$$

其中 $\delta_t(i)$ 表示在最优状态序列 s^* 下产生观测序列 $o_1,\cdots,o_t$ 的最大概率.式(11)中引入式(1)可得式(12)，假设观测序列开始于 $t=1$ 且初始最大概率为 $\delta_1(i) = \pi_i b_i(o_1)$，若已知 $\delta_{t-1}(i)$，$1 \leqslant j \leqslant K$ 则式(11)可改写为：

$$\varphi_t(i) = \max_j(\varphi_{t-1}(j) a_{ji}) e_i(o_t) \tag{13}$$

当 $t = n$ 时则得到了其当下的最优概率，通过回溯我们可以得到最优状态序列 s^*.然而，本文中的观测对象不是单个的点，而是由连续点组成的线段因此重新定义式(11)如下：

$$\varphi_t(i) = \max_{L_1,\cdots,L_k} \max_{s_1,\cdots,s_{k-1}} P(L_1,\cdots,L_k,s_1,\cdots,s_{k-1},s_k = i \mid x) \tag{14}$$

$L_1,\cdots,L_k$ 为线段序列，其中最后一个线段 L_k 终止于时刻 t，其对应状态为 i 且 k 的值上限为$\lfloor t/2 \rfloor$.已知式(13)中 $\varphi_t(i)$可由 $\varphi_{t-1}(j)$计算得到，现观测对象为一段范围内的点，则式(13)变为：

$$\varphi_t(i) = \max_{d,j}(\delta_{t-d}(j) a_{ji}) e_i(L) \tag{15}$$

线段 L 的长度由 d 决定，L 在$[t-d+1, t]$间的相似性误差小于等于误差阈值 ε_r.因此当维特比算法到达时刻 $t = n$ 时，通过回溯将得到最优分段序列及对应的最优状态序列.

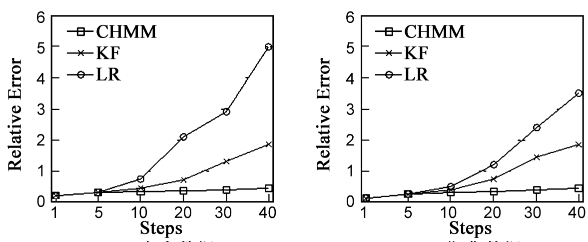
当得到最优状态序列后更新 CHMM 模型并将上诉过程迭代进行，假设在第 k 轮已知 CHMM 模型 x^{k-1}，通过维特比算法得到最优状态序列 s_k 和最优分段序列 L_k，同时根据状态序列和分段序列更新状态转移矩阵 $\boldsymbol{A}$、隐状态内分布 $f_k^{-y_{ji}}(y_{ji})$、线段斜率 θ 和长度 l 则得到更新后 CHMM 模型 x^k.

4 实验

实验基于两个真实数据组成的数据集，一个是美国的电力需求时序数据集，包含了从 1997 年全年的每 15 分钟电力需求变化数据共计 35260 个，该数据集的规律性较强.另一个是期货的日汇率时序数据集，包含了从 1986 年 10 月至 1996 年 8 月的每周期货汇率变化，该数据集的规律性较弱.当训练数据集的比例系数 N 大于等于 0.1 时，产生概率受系数变化的影响明显变小，因此 N 设定为 0.1.

4.1 多步值预测

该节实验将本模型与线性回归模型(Linear Regression model, LR)和卡尔曼滤波作比较.实验在两种数据集中进行，首先随机选取(总数据量 × N)个数据作为训练集，然后依次对当前窗口后的第 1,5,10,20,30,40 步的时间序列数据的值做出预测，比较相对误差结果如图 2.对于 LR 方法，由于依赖当前窗口的数据因此对于间隔大的数据预测性能较差.KF 方法虽然包含了多步预测规则，但是由于缺乏对于观测数据的时间关联性估计，随着步数的积累其准确性也逐渐降低.CHMM 方法在多步值预测中仍能保持较低误差.



(a) 电力数据 (b) 期货数据

图2 多步值预测

本文采用平均误差(Mean Error, ME)、均值绝对误差(Mean Absolute Deviation, MAD)、平均相对误差(Mean Relative Error, MRE)和均方根误差(Root Mean Square Error, RMSE)，4 个预测精度评价指标对预测结果进行评估.设 $e_l = \theta_{m+l} - \rho_{m+l}$，其中 θ 表示实际线段斜率，ρ 表示预测线段，L 表示预测步长.

$$\mathrm{ME} = \frac{1}{L}\sum_{l=1}^{L} e \tag{16}$$

$$\mathrm{MAD} = \frac{1}{L}\sum_{l=1}^{L} | e_l | \tag{17}$$

$$\mathrm{MSE} = \frac{1}{L}\sum_{l=1}^{L} e_l^2 \tag{18}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{L}\sum_{l=1}^{L} e_l^2} \tag{19}$$

其预测精度评价结果如表 1 所示.

表 1　CHMM、LR 和 KF 方法各个预测步长的预测精度评价

评估指标	ME				MAD				MRE				RMSE			
	1 步	5 步	10 步	20 步	1 步	5 步	10 步	20 步	1 步	5 步	10 步	20 步	1 步	5 步	10 步	20 步
CHMM	3.215	1.961	−1.034	1.002	15.461	20.989	19.002	23.530	0.223	0.416	0.425	0.431	17.538	21.245	22.574	27.316
LR	5.680	3.055	2.736	4.229	25.873	41.818	45.920	56.673	0.249	0.427	0.790	1.523	30.140	42.425	46.286	59.374
KF	4.721	2.321	1.846	3.634	20.805	31.499	36.628	42.351	0.231	0.420	0.492	0.774	23.720	34.751	38.692	45.154

表 1 显示,CHMM 所建模型的各项评价指标,均优于其它两种算法的对应指标.特别是其在多步预测的精度上较后两种算法有较大提升,具有较强的多步预测能力.

4.2　基于 AIC 信息准则的模型对比

为了进一步验证模型的有效性,本文引入了 AIC 信息准则(Akaike Information Criterion).该准则常用来衡量模型的拟合优良性,可以权衡估计模型的复杂度和此模型的拟合数据优良性.一般情况下 AIC 可表示为 $AIC=2k-2\ln L$,其中 k 表示自由参数数量,L 表示似然函数.假设模型的误差服从独立正态分布.则 $AIC=(2k+n\ln SSE)/n$,其中 n 为样本量,SSE 为残差平方和.其中模型的 AIC 值越小表示模型在最少自由参数的情况下能够很好的解释数据.本文对 4.2 节中的三个模型进行 AIC 值计算,其对比结果如表 2.

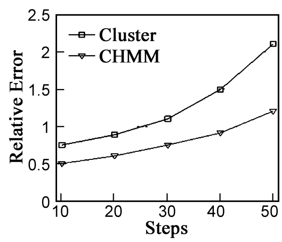
表 2　三种模型不同预测步长的 AIC 值对比

模型 \ 步长	5 步	10 步	20 步
CHMM	38.561	68.335	140.298
LR	39.477	78.696	165.354
KF	41.482	79.112	158.403

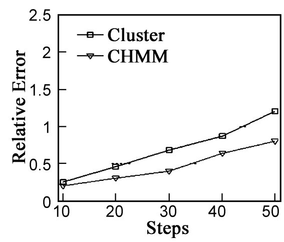
其中 $k_{CHMM}=4$,$k_{LR}=1$,$k_{KF}=3$ 从表 2 可以看出,在引入 AIC 信息准则后,在预测步长较短时,简单的模型比复杂模型更适用于预测场景,这是因为 AIC 准则不仅考虑了模型预测的精度也同时考虑了模型的复杂度.同时,在预测步长变长后 CHMM 模型的 AIC 值明显低于其它两种模型.

4.3　多步趋势预测

首先从两种测试数据集中随机选取连续的(总数据量 × N)个点作为训练集,假设训练后当前点时间为 t,设定一个间隔 $d=\{10,20,30,40,50\}$,对每个间隔 d 我们预测$[t+d+1,t+d+20]$区间长度为 20 的子序列趋势.例如,当前训练集时间点为 50,则预测[61,81],[71,91],…等区间段的趋势.同时以$(\theta-\rho)/\rho$ 表示相对误差,其中 θ 表示实际线段斜率,ρ 表示预测线段.实验比较了精化前和精化后模型的预测相对误差,通过文中式(15)可知,CHMM 模型的迭代过程使得更新后模型精度更高,如图 3 可以看到精化后模型的相对误差在两种数据集上明显降低,其中在 s 数据集中平均相对误差下降 34%,在 p 数据集中平均相对误差下降 18%.同时由于非参数估计方法在每轮更新模型时会重新计算隐状态分布,因此随着预测步数的增加,改进后的预测相对误差增长率低于原方法.



(*a*) 电力数据



(*b*) 期货数据

图3　多步趋势预测

5　结论

基于聚类的隐马尔科夫模型能够对时间序列的趋势做出有效预测,尤其在时间序列的多步预测精度上高于现有其它方法.其中基于时间性关联和相似性关联的聚类算法克服了以往聚类算法只考虑距离度量的问题,迭代的精化方法使得模型的预测精度更加提高.

参考文献

[1] 李晓光,宋宝燕,张昕.基于滑动多窗口的时间序列流趋势变化检测[J].电子学报,2010,38(2):321－326.
LI Xiao-guang, SONG Bao-yan, ZHANG Xin. Sliding multi-windows based trend change detection on time series stream [J]. Acta Electronica Sinica, 2010, 38(2): 321－326. (in Chinese)

[2] Box G, Jenkins G, Reinsel G. Time Series Analysis: Forecasting and Control[M]. Upper Saddle River: Prentice Hall, 1994. 1－784.

[3] Keogh E J. A decade of progress in indexing and mining large time series databases[A]. Dayal U. International Conference on Very Large Data Bases[C]. San Fransisco: Morgan Kaufmann, 2006. 1268－1269.

[4] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matching in time-series databases[A]. Snodgrass R T.

ACM Conference on Management of Data[C]. New York: ACM, 1994. 419 – 429.

[5] Keogh E, Pazzani M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback[A]. Stolorz P E. ACM Knowledge Discovery and Data Mining[C]. New York: ACM, 1998. 239 – 243.

[6] Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases[A]. Knowledge and Information System[C]. London: Springer, 2001. 263 – 286.

[7] Lin J, Keogh E J, Wei L, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series[A]. Data Mining and Knowledge Discovery[C]. London: Springer, 2007. 107 – 144.

[8] Mueen A, Keogh E. Online discovery and maintenance of time series motifs[A]. Tomkins A. ACM Knowledge Discovery and Data Mining[C]. New York: ACM, 2010. 1089 – 1098.

[9] Wang Y, Zhou L. Mining complex time-series data by learning the temporal structure using Bayesian techniques and Markovian models[A]. International Conference on Data Mining[C]. Washington: IEEE Press, 2006. 1136 – 1140.

[10] Won P, Melek W, Golnaraghi F. A Kalman/particle filter-based position and orientation estimation method using a position sensor/inertial measurement unit hybrid system industrial electronics[J]. IEEE Transactions on Industrial Electronics, 2010, 57(5): 1787 – 1798.

[11] Zhang Y. Prediction of Financial Time Series with Hidden Markov Models[R]. Canada: Simon Fraser University, 2004. 17 – 18.

[12] Griffiths T, Ghahramani L. Infinite latent feature models and the indian buffet process[A]. Platt J C. Neural Information Processing Systems[C]. Vancouver: University of British Columbia, 2005. 672 – 684.

[13] Ferguson T S. A Bayesian analysis of some nonparametric problems[J]. Annals of Statistics, 1973, 1(2): 209 – 230.

[14] 周建英，王飞跃，曾大军. 分层 Dirichlet 过程及其应用综述[J]. 自动化学报，2011，37(4): 389 – 407.
ZHOU Jian-ying, WANG Fei-yue, ZENG Da-jun. Hierarchical Dirichlet processes and their applications: A survey[J]. Acta Automatica Sinica, 2011, 37(4): 389 – 407. (in Chinese)

[15] Jordan M. Hierarchical Bayesian Nonparametric Models with Applications. Bayesian Nonparametrics Principles and Practice[M]. London: Cambridge University Press, 2009. 1 – 47.

作者简介

章登义　男，1965 年 5 月出生于湖北省荆州市. 现为武汉大学计算机学院教授、博士生导师.
E-mail: dyzhangwhu@163.com

欧阳黜霏(通信作者)　男，1988 年 7 月出生于湖北荆州. 现为武汉大学博士研究生，主要研究数据库、数据挖掘.
E-mail: whuxiaoouou@whu.edu.cn

吴文李　女，1986 年 12 月出生于广东湛江. 现为武汉大学博士研究生. 主要研究语义网数据挖掘.
E-mail: huihuigou@whu.edu.cn