Finding Semantics in Time Series

肖钦哲

Peng Wang, etc, "*Finding Semantics in Time Series*," SIGMOD, 2011

INTRODUCTION

背景

时序研究

- Prediction
- Classification
- Anomaly Detection
- Clustering
- Segment
- …

背景

时序研究

- Pattern discovery (motif)

从时序中找到有意义的，频繁的模式(子序列)，比如根据相似度度量子序列。

然而，研究序列内部动态的几乎没有。所以这篇论文提出了pattern-based hidden Markov model (pHMM)，来解释时序是如何生成的.

Motivation

将系统对序列的生成是基于一系列的隐含状态，在每个时刻会有一个状态对应，当发生某个事件时触发状态改变。

1. 当前状态生成序列(观察)
2. 序列之间有转移概率

HMM建模

Motivation

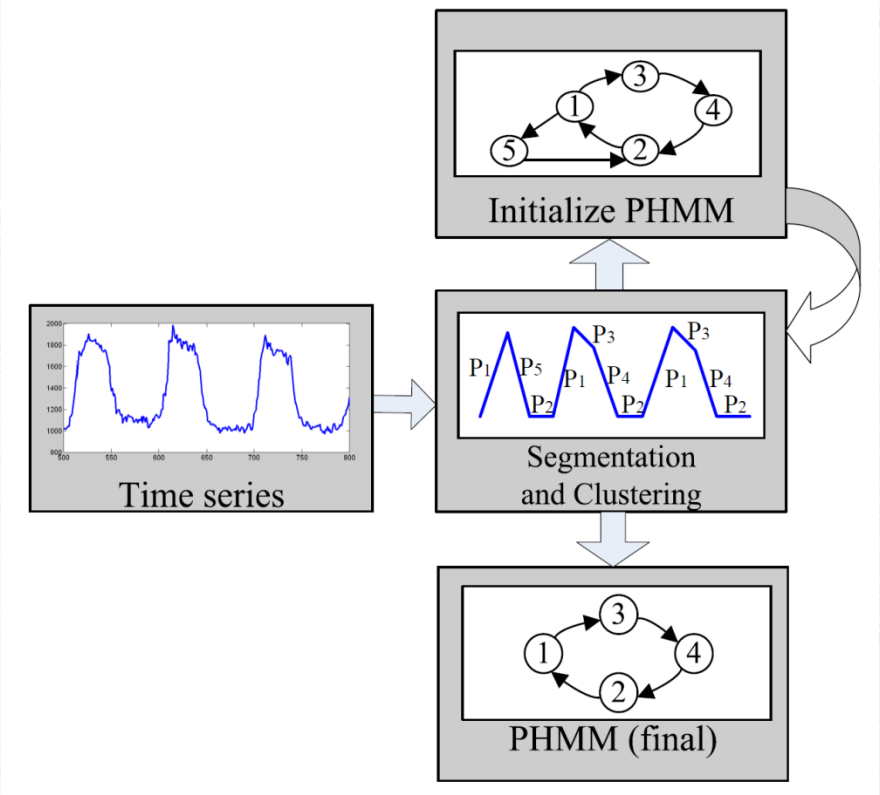
观察序列 —— 状态序列

1. 一个点对应一个状态
2. 多个点(线段/曲线)

对观察序列分段，然后聚类，一个类就是一个状态。

pHMM

Two phase approach



PRELIMINARY AND PROBLEM DEFINITION

Preliminary

HMM：

状态：$S = \{1, 2, \cdots, K\}$

初始状态概率：π

状态转移概率：$A = \{a_{ij}\}$

输出概率：$B = \{b_i(o)\}$

3个问题：

给定参数，求观察序列的概率；给定参数，求观察序列最可能的对应的内部状态序列；给定观察，求参数

Problem

- 将序列X转为线段序列L=($L_1 \cdots L_m$)
- Learn HMM from L，使

$$P(\mathbf{L}, \mathbf{s}^* | \lambda) = \pi_{s_1} b_{s_1}(L_1) \prod_{j=2}^{m} a_{s_{j-1}, s_j} b_{s_j}(L_j)$$

最大，s*是最优状态序列。

THE INITIAL PHMM

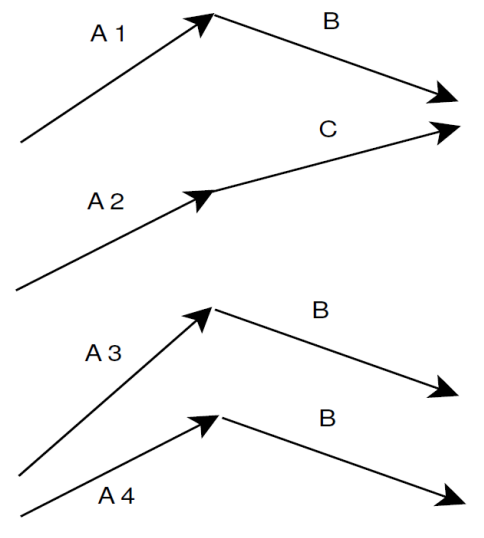
Line Segment

1. 相邻两个点相连
2. 计算每个相邻段合并后的误差，取最小误差的合并
3. 迭代2直到误差大于定义的阈值ε_r时。

Clustering

评估相似性Metrics：

1. Similarity. Shape上的： slope and duration。
2. Temporal. 属于同一类的L_i，L_j 的后一段L_{i+1}，L_{j+1} 应该有相同的分布。



Measure

For relative error of cluster C_i

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

其中l是长度，θ是角度。

For uncertainty of cluster C_i

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

Measure

综合：

$$F = \alpha \cdot R + (1 - \alpha) \cdot I$$

$$R = \sum_{i=1}^{K} |C_i| R(i)$$

$$I = \sum_{i=1}^{K} p(i) I(i)$$

Clustering

Step 1. 对于每个聚类C_i，对应一个T_i，满足

1. $R(T_i \cup C_i) \leq R(C \cup C_i)$ *holds for any* $C \neq T_i$, *where* $T_i \cup C_i$ *is the new cluster generated by merging* T_i *and* C_i, *and* $R(T_i \cup C_i)$ *is its relative error.*

2. $R(T_i \cup C_i) \leq \varepsilon_c$, *where* ε_c *is a user-specified threshold, called the relative error threshold.*

Step 2. 对每个cluster pair (*Ci; Ti*), 计算交叉熵

Step 3. 合并最小交叉熵的一对

重复直到所有R()都大于ε_c

Cluster to HMM

1. 每个类对应一个状态
2. 输出概率：

$$b_i(L) = p_l(l|i)p_s(\theta|i)$$

用高斯分布拟合

3. 根据状态序列得到转移概率和初始概率

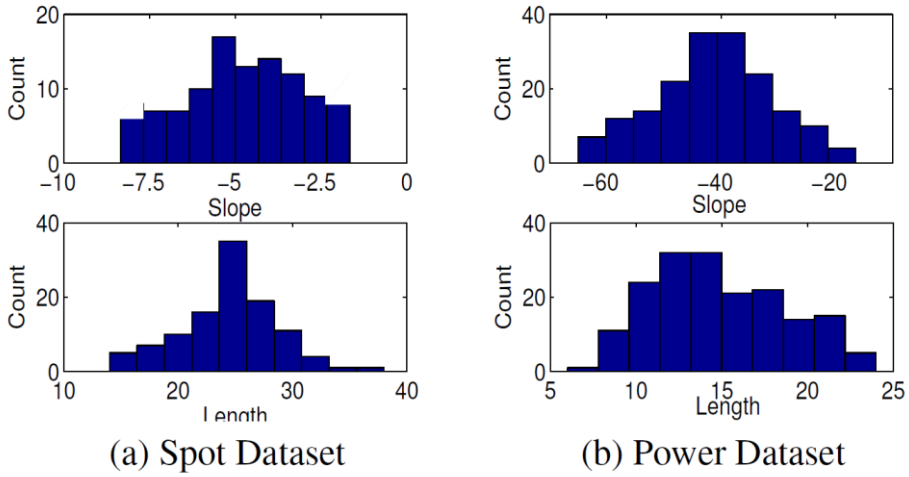


Figure 4: Slope and length distribution

ITERATIVE PHMM REFINEMNT

Why Refinemnt

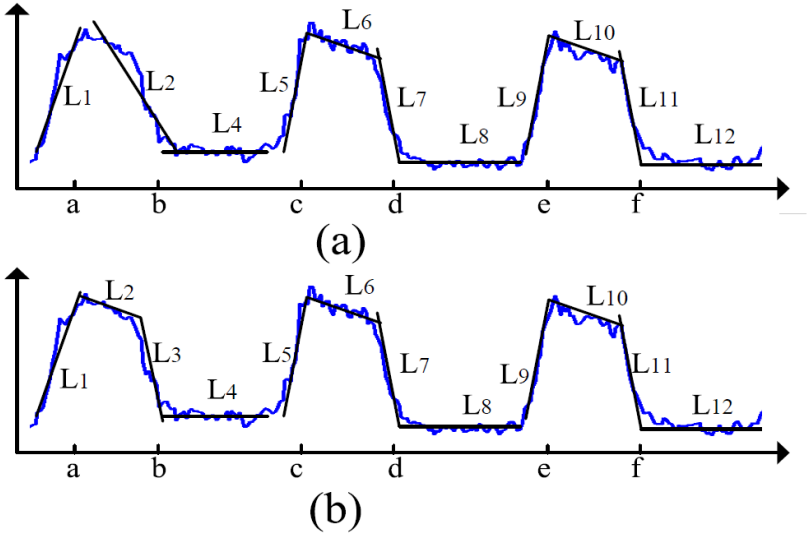


Figure 5: Segmentation with/without Refinement

pHMM based Segment

Traditional Viterbi Algorithm：

$$\begin{aligned}\delta_t(i) &= \max_{s_1,\cdots,s_{t-1}} P(o_1,\cdots,o_t,s_1,\cdots,s_{t-1},s_t=i|\lambda) \\ &= \max_{s_1,\cdots,s_{t-1}} \pi_{s_1} b_{s_1}(o_1) \prod_{j=2}^{t} (a_{s_{j-1},s_j} b_{s_j}(o_j))\end{aligned}$$

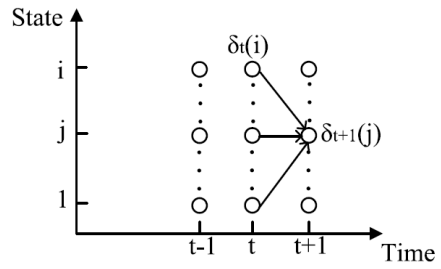
迭代计算：

$$\delta_1(i) = \pi_i b_i(o_1)$$

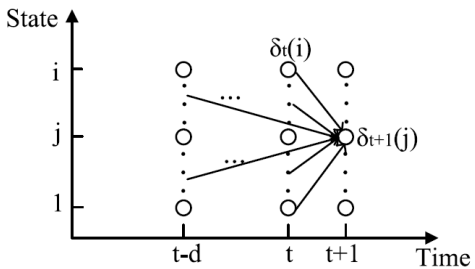
$$\delta_t(i) = \max_j (\delta_{t-1}(j) a_{ji}) b_i(o_t)$$

pHMM based Segment

Extending Viterbi Algorithm：



(a) Traditional



(b) Extended

$$\delta_t(i) = \max_{L_1,\cdots,L_k} \max_{s_1,\cdots,s_{k-1}} P(L_1,\cdots,L_k,s_1,\cdots,s_k = i|\lambda)$$

$$\delta_t(i) = \max_{d,j}(\delta_{t-d}(j)a_{ji})b_i(L)$$

d的限制是线段误差。

Pruning Strategies

分段时可以做些剪枝，降低计算量

1. 计算t时刻的delta时，找到t`满足

$$Err(BestLine(t'-1,t)) > \varepsilon_r \text{ and } Err(BestLine(t',t)) < \varepsilon_r$$

t`之前的就不用计算

2. 略

3. 略

Boundary Point

如果t时刻两边的slope相差大，那么这个点更像边界点。

利用：在clustering时，按点两边的线段被合并的顺序排列，取最后被合并的n个点作为边界点，用于pHMM segment。

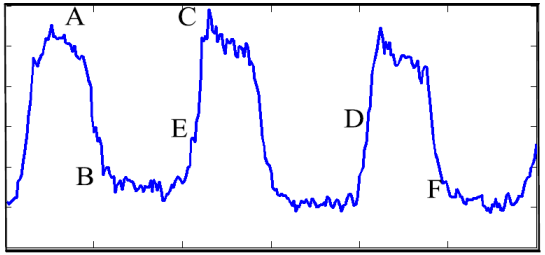


Figure 7: Boundary points

Update pHMM

由Extended Viterbi得到最优状态和分段，即可更新

THEOREM 1. *The production probability of round k is not less than that of round $k-1$, that is:*

$$P(\mathbf{L}^k, \mathbf{s}^k | \lambda^k) \geq P(\mathbf{L}^{k-1}, \mathbf{s}^{k-1} | \lambda^{k-1})$$

Applications

1. Multi-Step Prediction

根据当前状态预测（？）

2. Trend Prediction

当前状态的时间与平均时间对比，得知当前状态持续多久；根据转移概率得到后续状态，从而预测更远的。

3. Correlation Detection

EXPERIMENTS

Prune & N

N	0.05	0.1	0.5	1
No-Prune	96	381	9265	-
Prune 1	41	93	689	1921
Prune 1&2	29	70	425	934

Table 1: Runtime Comparison(s)

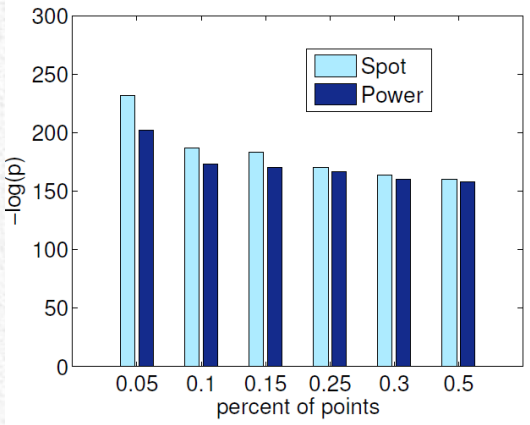


Figure 9: Influence of N

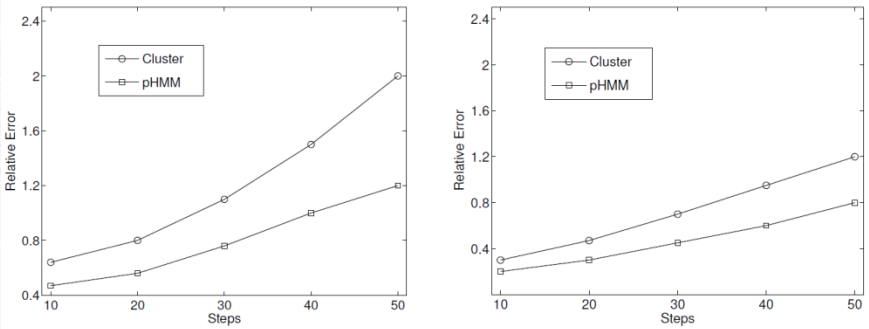
Prune减少时间

小的训练数据即可

Experiments

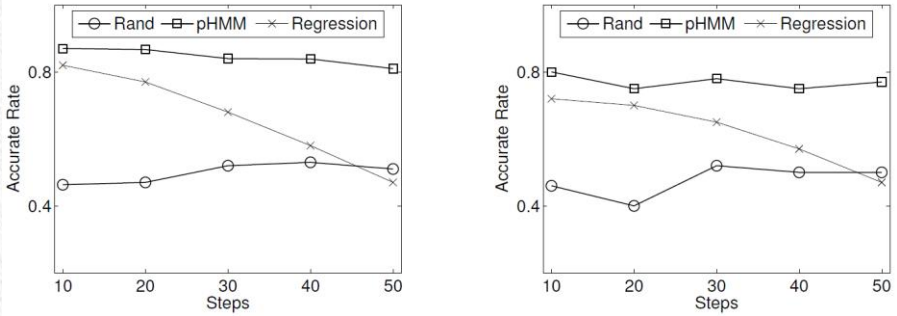
Trend Prediction

Test：随机取50个点，预测后第10，20，…，50个点后长20子串的趋势。



(a) Spot dataset (b) Power dataset

Figure 12: Accuracy of Trend Prediction



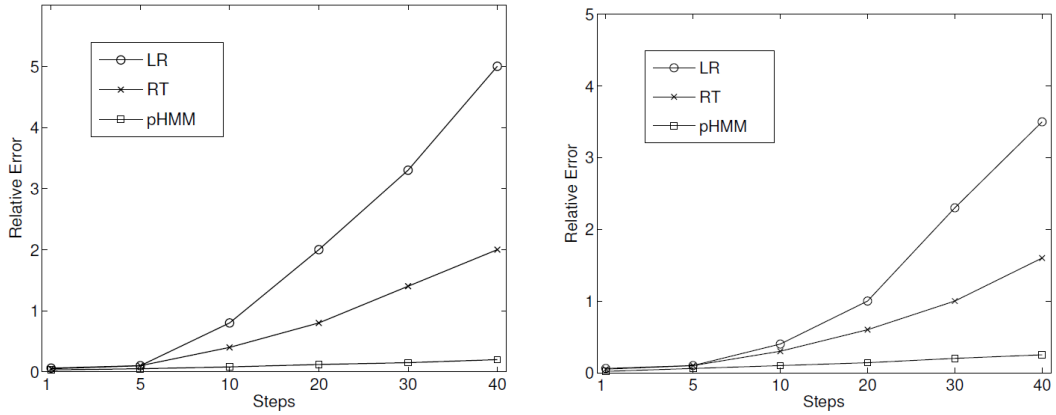
(a) Spot dataset (b) Power dataset

Figure 13: Accuracy of binary trend prediction

Experiments

Multi-step Value Prediction

Test：随机取50个点，预测后第1，5，…，40个点。



(a) Spot dataset (b) Power dataset

Figure 14: Accuracy of Multi-step Value Prediction

Thanks for your Listening