AutoPlait: Automatic Mining of Co-evolving Time Sequences

Yasuko Matsubara
Kumamoto University
yasuko@cs.kumamoto-u.ac.jp

Yasushi Sakurai
Kumamoto University
yasushi@cs.kumamoto-u.ac.jp

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Motivation - Given: Co-evolving time-series
e.g., Mocap (leg/arm sensors) - “chicken dance”

Challenges
(1) Unknown # of patterns (2) Different durations
Q. Can we summarize it automatically??

Goal: find patterns that agree with human intuition
AutoPlait: “fully-automatic” mining algorithm

Importance of “fully-automatic”
No magic numbers! … because,
Manual - sensitive to the parameter tuning
- it takes a very long time (hours, days, …)
Automatic - no expert tuning required
Big data mining: we cannot afford human intervention!!

Problem formulation - Key concepts
• Bundle (given) : d co-evolving sequences, \( X = \{x_1, \ldots, x_d\} \)
• Segment : convert \( X \) \( \rightarrow m \) segments, \( S \)
• Regime : segment groups, \( \Theta \)
• Segment-membership: assignment, \( F \)

Proposed method: AutoPlait
Main idea (1): MLCM: multi-level chain model

Main idea (2): Model description cost

AutoPlait at work - (a) Model analysis (WebClick)

Conclusions - AutoPlait has following advantages:
• Effective & Sense-making: it provides reasonable regimes
• Fully-automatic: it needs no magic numbers
• Scalable: it scales linearly with the duration \( n \)

CutPointSearch (inner-most loop) Given: \( X \), regimes \( \Theta = \{\theta_1, \theta_2, \Delta\} \)
Find: cut-points of segs: \( \{s_1, s_2\} = \arg \max P(X | s_1, s_2, \Theta) \)
DP algorithm to compute likelihood: \( P(X | \Theta) \)

RegimeSplit (inner-loop algorithm) Given: \( X \),
Find (1) two segment sets: \( s_1, s_2 \) (2) regimes: \( \Theta = \{\theta_1, \theta_2, \Delta\} \)
Two-phase iterative approach
[P1] split segments (CPS), [P2] update model parameters

Experiments - (a) Sense-making (Mocap)
(NO user defined parameters)

(b) Segmentation/Clustering accuracy (c) Scalability

Q&A

AutoPlait (outer-loop algorithm)
Split regimes \( r=2, 3, \ldots \) as long as cost keeps decreasing
- Find appropriate # of regimes

Q. Any distinct patterns?
Q. How many? - what kind?

\[
\theta = (\theta_1, \theta_2, \ldots, \theta_r) \quad \text{r HMMs across-regime (regimes) transition prob.}
\]

Single HMM parameters

\[
\text{Cost}_{\Theta}(X; C) = \sum \text{Cost}(X; m, r, S, \Theta, T)
\]

Segment membership \( F \)

\[
\text{Cost}_{\Theta}(X; C) = \log|m| + \log^2|d| + \log^2|m| + \log|r| + m|\log r|
\]

The 2014 ACM SIGMOD/PODS Conference
June 22-27, 2014, Snowbird, Utah, USA

ACM SIGMOD/PODS 2014