When Big meets Complex: Learning and Mining in Large-scale Time Series Data

Yan Liu

Email: yanliu.cs@usc.edu
Computer Science Department
Viterbi School of Engineering
University of Southern California

Joint Work with Taha Bahadori, Hongfei Li, Dian Gong, Gerard Medioni
Time Series Data are Everywhere

One important task in “From Data to Knowledge”:

Discovery of Temporal Dependence Relationships
Output: temporal causal graph of climate forcing agents
Output: social influence networks
Major Challenges

*high-dimensional* time series data with complex relations
Basic idea: Graphical modeling using the notions of Granger causality and methods of variable selection

Granger Causality: Cause happens prior to its effects [Granger 1969, 1980]. A time series $y$ is the Granger Cause of another time series $x$ if the past values of $y$ are helpful in predicting the future values of $x$ given its own past.

Practically, we perform the following two autoregressions:

1. $x_t = \sum_{l=1}^{L} a_l x_{t-l}$
2. $x_t = \sum_{l=1}^{L} a'_l x_{t-l} + \sum_{l=1}^{L} b'_l y_{t-l}$

If Eq. (2) is a significantly better model than Eq. (1) (by statistical significance test), we determine that time series $y$ Granger causes time series $x$. 

Yan Liu (USC)
Granger Graphical Models

**Lasso-Granger** [Arnold et al, KDD 2007]: Given $P$ time series $x^{(1)}, \ldots, x^{(P)}$ of length $T$, we can determine the Granger relationships of $x^{(i)}$ by performing the penalized auto-regression as follows:

$$
\min_{\{a_i\}} \sum_{t=L+1}^{T} \left\| x_t^{(i)} - \sum_{j=1}^{P} a_{i,j} x_{t,Lagged}^{(j)} \right\|^2 + \lambda \| a_i \|_1,
$$

where $x_{t,Lagged}^{(j)} = \left[ x_{t-L}^{(j)}, \ldots, x_{t-1}^{(j)} \right]$.

**Major advantages**
- Variable selection can be efficiently achieved for high-dimensional time series
- Consistency analysis [Arnold et al, KDD 2007; Bahadori and Liu, 2012]
  
  Lasso-Granger: $P[\text{Error}] = o(c'L \exp(-T^v))$ for some $0 \leq v < 1$.
  
  Significant test: $P[\text{Error}] = o(c' \sqrt{T-L} \exp(-c^2(T-L)/2))$

  Learning is possible even when the dimension $P$ is significantly larger than $T$!
Challenges in Practical Applications

Our solutions via Granger Graphical Models

- Non-stationary time series [KDD 2009]
- Natural grouping of time series [KDD 2009]
- Spatial time series [KDD 2009]
- Nonlinear time series [AAAI 2010]
- Relational time series [ICML 2010]
- Irregular time series [SDM 2012]
- Extreme-value time series (Complexity) [ICML 2012]
- Fast learning algorithms for time series (Scalability) [ICDM 2012, ICML 2013 submission]

Successful application in computational biology, climate analysis, social media analysis, business analytics and so on.
**Complexity:** Extreme-value Time Series

**Extreme value:** extremely (deviations from the median) high (or low) values of time series observations, such as heatwave, flood, or social media buzz (extremely high frequency of mentions of related words)

**Challenges:** (1) Heavy tail distributions; (2) Data scarcity
Suppose $X_1, X_2, \ldots, X_n$ are iid distributed. Let $M_n = \max\{X_1, X_2, \ldots, X_n\}$. Under certain conditions, the cdf of $\frac{M_n - bn}{an}$ has the following form:

$$G(x|\mu, \sigma, \xi) = \exp \left\{-1 + \xi \left(\frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}$$

where

- $\mu$ is the location parameter (the mode of the distribution)
- $\sigma$ is the scale parameter (proportional to the variance)
- $\xi$ is the shape parameter (the behavior of the tail)
Our Solution: Sparse-GEV Model

**Basic idea** Instead of directly modeling the temporal dependence among extreme value time series (which is nonlinear and non-Gaussian), we can model the dependence via the hidden location parameters (which determine the model of the distribution)

\[
p(\{x^i_t\}, \{\mu^i_t\}|\beta, \sigma, c) = \prod_{i=1}^{P} \prod_{t=L+1}^{T} p(x^i_t|\mu^i_t, \sigma^i) p(\mu^i_t|\{\mu^j_{t-l}\}, \beta, c),
\]

Where

\[
x^i_t|\mu^i_t, \sigma^i \sim \text{Gumbel}(\mu^i_t, \sigma^i) = \exp \left\{ - \exp \left\{ - \frac{x - \mu^i_t}{\sigma^i} \right\} \right\}
\]

\[
\mu^i_t|\{\mu^j_{t-l}\}, \beta, c \sim \text{Gaussian}(c^i + \sum_{\ell=1}^{L} \sum_{j=1}^{P} \beta^i_{j,\ell} \mu^j_{t-l}, \sigma_0)
\]
Experiment Results on the Twitter Datasets

Twitter Meme:

- Tiger Woods
- Haiti
- Scott Brown
- Christmas
- Avatar
- Dubai Burj

Occupy WS:

- O-LA
- O-WS
- O-SF
- O-DC

Prediction Accuracy

<table>
<thead>
<tr>
<th></th>
<th>T-Meme</th>
<th>T-OWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse-GEV</td>
<td>0.0503</td>
<td>0.1190</td>
</tr>
<tr>
<td>Granger</td>
<td>0.0819</td>
<td>0.1410</td>
</tr>
<tr>
<td>TE</td>
<td>0.0972</td>
<td>0.1302</td>
</tr>
<tr>
<td>Copula</td>
<td>0.1009</td>
<td>0.1240</td>
</tr>
</tbody>
</table>
Challenges in Practical Applications

Our solutions via Granger Graphical Models

- Non-stationary time series [KDD 2009]
- Natural grouping of time series [KDD 2009]
- Spatial time series [KDD 2009]
- Nonlinear time series [AAAI 2010]
- Relational time series [ICML 2010]
- Irregular time series [SDM 2012]
- Extreme-value time series (Complexity) [ICML 2012]
- Fast learning algorithms for time series (Scalability) [ICDM 2012, ICML 2013 submission]
Scalability: Fast Retrieval Algorithms

Fast (approximate) nearest neighbor search has been studied for non-time series for more than a decade, e.g. k-d tree and locality sensitive hashing.

New challenges of hashing algorithms for time series data

- The temporal length of time series observations could be different
- Defining a proper similarity metric is challenging
- Efficiently computing the codes for a new query is also difficult.

Dynamic time warping (DTW)

- Traditionally $O(nm)$, and recently significantly improved [Rakthanmanon et al., 2012].
- DTW is not a proper distance metric and the induced kernel is not a positive definite kernel.
Scalability: Fast Searching Algorithms

Our solution: Kernelized Global Alignment Hashing, which naturally integrates kernelized hashing (KH) with global alignment (GA) kernels.

Random Projection: given an input vector \( x \) and a hyperplane defined by \( r \), we let \( h(x) = \text{sign}(v \cdot r) \). That is, \( h(x) = +/−1 \) depending on which side of the hyperplane \( x \) lies.

**Kernelized Locality Sensitive Hashing** [Kulis & Grauman, 2011]

\[
h_{KAH}(\phi_{GA}(X)) = \text{sgn}\left(\sum_{j=1}^{m} \omega_j k_{GA}(X(j), X)\right)
\]

(4)

where \( \omega = [\omega_1, ..., \omega_m]^T \) is given by \( K_b^{-1/2} e_t^m \), and \( e_t^m \) is length-\( m \) select vector containing ones at \( t \) random positions, i.e., select \( t \) time series from total \( m \) basic ones.

**Global Alignment Kernels** [Cuturi, 2011]

\[
k_{GA}(X_{1:L_x}, Y_{1:L_y}) = \sum_{Q \in \mathcal{A}(L_x, L_y)} |Q| \prod_{t=1}^{\left|Q\right|} \kappa_\sigma(x_{q_t(1)}, y_{q_t(2)})
\]

(5)

where \( \kappa_\sigma(\cdot) \) is the pre-defined local kernel for vector input.
Scalability: Fast Searching Algorithms

Computational Complexity: $O(Nml^2d)$

Experiment Results: smart phone data and social behavior data

### Prediction Accuracy

![Prediction Accuracy Graph](image1)

### Searching Accuracy

![Searching Accuracy Graph](image2)
Big Complex Data: Time Series Learning and Mining

One important task in “From Data to Knowledge”:

*Discovery of Temporal Dependence Relationships*
Thank you!

Code is available at: http://www-bcf.usc.edu/~liu32/code.htm