Change detection in non-stationary time series based on genetic algorithm

M. Gorji

Spring 2015
Overview

- Motivation
- Problem Definition
- Current Approaches
- Proposed Method
- Simulation
- Future Works
**Motivation**

Example of time-varying systems can be found in science and engineering:

- **Stock market:** Variation of price (volatility) in stock market are changing.
- **Control Systems:** The pressure of a boiler-turbine unit changes by switching among operating points.
- **Climate System:** Due to “climate change”, climatic variables switch among various “regimes”.
- **Biological systems:** Activity of genes in living organism changes in periods of organism’s life.

- Modeling behavior of time-varying systems is an important problem in various fields.
- Time-varying systems can be modeled by analyzing the generated non-stationary time series.
Non-stationary time series

- A non-stationary time series has statistical parameters dependent on time.
- These time series have several clusters, while each cluster has constant statistical parameters.
- A time invariant model is not adequate to represent system behavior.

\[
P(X = x) = f(x, \alpha_1) \quad f(x, \alpha_2) \quad f(x, \alpha_1) \quad f(x, \alpha_3) \quad f(x, \alpha_2)
\]

OR

\[
x_{t+1} = f(x_t, t, \alpha_1) \quad f(x_t, t, \alpha_2) \quad f(x_t, t, \alpha_1) \quad f(x_t, t, \alpha_3) \quad f(x_t, t, \alpha_2)
\]

- The goal of time series analysis is to find model of each cluster and switch times.
Let \( x(t) \in \mathbb{R}^n \) be a multidimensional non-stationary time series over \( t = \{1, \ldots, T\} \) with \( C \) clusters. The model of the time series in each cluster could be a function of time or a probability density function:

\[
x(t) = f(x(t-1), \ldots, x(t-p), u(t-1), \ldots, u(t-q), t, \theta_c)
\]

- Here, \( f \) is the model of the time series, \( \theta_c \) is the set of parameters in the \( c \)th cluster where \( c \in \{1, \ldots, C\} \) and \( u(t) \) is the time series of input.
- Also, \( p \) and \( q \) are the order of the lagged inputs and the lagged outputs respectively.

- For this case, the distance function is defined by Euclidean distance.

\[
d(x(t), \theta_c) \triangleq \|x(t) - f(x(t-1), \ldots, x(t-p), u(t-1), \ldots, u(t-q), t, \theta_c)\|^2
\]

Examples: \( w(t) \) is the white noise.

\[
x(t) = \theta_c + w(t)
\]

\[
d(x(t), \theta_c) \triangleq \|x(t) - \theta_c\|^2
\]
Let $x(t) \in \mathbb{R}^n$ be a multidimensional non-stationary time series over $t=\{1,\ldots,T\}$ with $C$ clusters. The model of the time series in each cluster could be a function of time or a probability density function:

$$P(X = x(t)) = f(x(t)|u(t), \theta_c)$$

Here, $f$ is the model of the time series, $\theta_c$ is the set of parameters in the $c$th cluster where $c \in \{1,\ldots,C\}$ and $u(t)$ is the time series of input.

For this case, the distance function is defined by negative log-likelihood function:

$$d(x(t), \theta_c) \triangleq \ell(f(x(t)|u(t), t, \theta_c))$$

Examples:

$x_t \sim \mathcal{N}(\rho, \sigma^2)$

$$f_{x_t}(x_t) = \frac{1}{\sigma_c \sqrt{2\pi}} \exp\left(-\frac{(x_t - \rho_c)^2}{2\sigma_c^2}\right)$$

$$\ell(x_1, \ldots, x_T | \rho_c, \sigma_c) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_c^2) - \frac{1}{2\sigma_c^2} (x_t - \rho_c)$$
Assume that the observed time series $x_t$ has $C$ clusters. During $c'$th cluster, the piecewise constant parameters are $\theta_c$. The non-convex optimizations can be defined as:

$$
\sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot [x_t - f(x_t, t, \theta_c)]^2 \xrightarrow{\mu(t, \alpha)} \min
$$

Euclidean distance between time series $x_t$ and the piecewise constant models with parameters $\theta_c$

$$
\sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot \ell(\theta_c | x_t) \xrightarrow{\mu(t, \alpha)} \min
$$

Negative log-likelihood of piecewise constant models with parameters $\theta_c$, given the time series $x_t$

The $\mu_c(t)$ is cluster membership function, it shows that to what degree, data in time $t$ belongs to cluster $c$.

Time series belongs to only one of the clusters at each time, Thus:

$$
\mu_c(t) \in \{0, 1\} \quad \sum_{c=1}^{C} \mu_c(t) = 1
$$
Consider the following time series

- The time series has 3 clusters with different mean values and 3 switch points at $t=100$, 150 and 200.
- The model in each cluster is:
  \[ x(t) = \theta_c + w(t) \]
  \[ \theta_c = \{2, 3, 4\} \]
  \[ w(t) \sim N(0, 0.5^2) \]
  \[ c \in \{1, 2, 3\}; \ t \in [1, 300] \]
- The goal is to find cluster parameters $\theta_c$ and cluster membership function $\mu_c(t)$
<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search algorithms</td>
<td>Search for best change-points by brute-force, ...</td>
<td>Not applicable for long time series, High volume of computation</td>
</tr>
<tr>
<td>Parametric or Nonparametric tests</td>
<td>Student’s t-test, Spearman’s ρ, Mann–Kendall’s τ , and etc ...</td>
<td>Assume the data are uncorrelated and independent, It can’t find the change-points when the autoregressive models exist in the time series</td>
</tr>
<tr>
<td>Bayesian Techniques</td>
<td>Change points times and other model parameters are assumed to be random variables and their statistical distribution are estimated</td>
<td>Needs statistical assumptions on the distribution of data in time series</td>
</tr>
<tr>
<td>Markov Models</td>
<td>Assume that data in time series are generated by a Markov process</td>
<td>Needs statistical assumptions on the distribution of data in time series. The Markovian assumption may be restrictive</td>
</tr>
<tr>
<td>Convex relaxation</td>
<td>Convert the optimization to a convex problem</td>
<td>Suffers from local solution, and are not suitable for long time series.</td>
</tr>
</tbody>
</table>
The defined problem is a non-convex mixed-integer optimization with two sets of unknown parameters, the cluster membership function $\mu_c(t)$ and the model parameters $\theta_c$.

It is shown that the optimization problem doesn’t have a unique solution.

Furthermore, small changes in the $x(t)$ might lead to large changes in the estimated $\mu_c(t)$, and thus the solution is not robust.

An assumption is made in literature that the number of switches between the clusters are known and is equal to $W$.

In this case, another constraints can be added to the problem definition.

$$\sum_{c=1}^{C} \sum_{t=1}^{T} |\mu_c(t + 1) - \mu_c(t)| = 2W$$

This constraint should be added to the optimization problem, in order to limit the number of switches between the clusters.

This process make the problem solvable, and is called regularization.
Coordinate-descent

- If the $\mu_c$ is known, the switch points can be found and finding the cluster parameters $\theta_c$ is usually straightforward.
- The optimization can be solved by coordinate-descent algorithm, i.e. iteratively with respect to two set of parameters $\mu_c$ and $\theta_c$, similar to K-means.

**Optimization with respect to $\mu_c$**

$$\sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x(t), \theta_c) \xrightarrow{\mu(t)} \min$$

Subject to

$$\sum_{c=1}^{C} \sum_{t=1}^{T} |\mu_c(t+1) - \mu_c(t)| = 2W$$

$$\sum_{c=1}^{C} \mu_c(t) = 1$$

Constrained optimization, non-convex, difficult

**Optimization with respect to $\theta_c$**

$$\sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x(t), \theta_c) \xrightarrow{\theta_c} \min$$

Unconstrained optimization, easy to solve by statistical methods
Genetic Algorithm

- Generate a population of individuals, each one might be a solution of a non-convex optimization
- Apply crossover and mutation,
- Guided by fitness of individuals, find the best individual.

In general, GA is not suitable for constrained optimization.

Constraint-handling techniques:

- Penalize the solutions that violate the constraints.
- Design special representations and operators, based on the problem.
- Make the search areas close to the boundary of the feasible region.
- etc...

In this work, individuals are generated through crossover/mutation operators such that they always stay in the feasible region.
The Individuals represent $W$ switch times among the $C$ clusters, and the corresponding $W + 1$ indices of the clusters sequence in a non-stationary time series with the length of $T$. The switch times are integer numbers from the set of $\{f, \ldots, T\}$ and the sequence indices indicate the name of the clusters from the set of $\{1, \ldots, C\}$.

The length of time series is $T = 250$ and there are six switches among the three clusters. The cluster 1 is activate three times at $\{1, \ldots, 9\}$, $\{40, \ldots, 69\}$, $\{150, \ldots, 200\}$, the cluster 2 is activate three times at $\{10, \ldots, 39\}$, $\{95, \ldots, 149\}$, $\{201, \ldots, 250\}$ the cluster 3 is active only at $\{70, \ldots, 94\}$.

Note: a sequence of $\{1, 2, 2, 1, 3, 2, 1\}$ with consecutive occurrence of index 2 is not a valid representation.
A random number $\alpha \in [0, 1]$ determines the crossover site among the switch times or the sequences.

If $\alpha > 0.5$, the switch times of the crossover child is:

$$\text{Floor}[\lambda \times \text{Switch times of the first parent} + ((1 - \lambda) \times \text{Switch times of the second parent})]$$

where $\lambda$ is a random number in the range of $[0, 1]$.

(a) The crossover is applied on the switch times of the parents to find a new set of switch times.
A random number $\alpha \in [0, 1]$ determines the crossover site among the switch times or the sequences.

If $\alpha < 0.5$, the switch times of the crossover child is:

To prevent generating an invalid crossover child, possible crossover sites are defined. A locus $z \in \{3, ..., W\}$ in the sequence of the clusters is chosen as a crossover site if:

- $\text{Index}_{\text{parent}1}(z) \neq \text{Index}_{\text{parent}2}(z + 1)$
- $\text{Index}_{\text{parent}2}(z) \neq \text{Index}_{\text{parent}1}(z + 1)$

(b) The possible crossover sites are marked with circle. For $z = 6$, the crossover child is shown at the bottom of the figure.
A random number $\beta \in [0, 1]$ determines the crossover site among the switch times or the sequences.

If $\beta < 0.5$, switch times in the individual are mutated.

For a random locus $z$ in the switch times, where $z \in \{1, \ldots, W\}$, the switch time $\tau(z)$ is mutated to another value in the Mutation Range ($MR$):

$$MR = \begin{cases} 
\{2, \ldots, \tau(z+1) - 1\} & z = 1 \\
\{\tau(z-1) + 1, \ldots, \tau(z+1) - 1\} & 2 \leq z \leq W - 1 \\
\{\tau(z-1), \ldots, T - 1\} & z = W 
\end{cases}$$

- The mutation site $z$ can be chosen from the set of $\{1, \ldots, 9\}$. As an example:
  - for $z = 1$, the MR is $\{2, \ldots, 39\}$
  - for $z = 4$, the MR is $\{61, \ldots, 109\}$
  - for $z = 8$, the MR is $\{202, \ldots, 244\}$
  - After choosing the mutation site and finding the mutation range, the value of $z$ will be mutated to another value in the MR.
**Mutation cluster sequences:**

A random number $\beta \in [0, 1]$ determines the crossover site among the switch times or the sequences.

**If $\beta > 0.5$, the cluster sequences in the individual are mutated.**

For the loci $3 \leq z \leq W+1$ in the sequence, the cluster index $s(z)$ is mutated to a new index which is different from $s(z-1)$, $s(z)$, and $s(z+1)$.

- As an example, for $C=4$:
  - for $z = 5$, the mutated can be $\{4\}$
  - for $z = 7$, the mutated can be $\{3\}$
  - for $z = 10$, the mutated can be $\{2,4\}$

- After choosing the mutation site $z=5$, the value of $z$ will be mutated to another value

![Diagram showing mutation process]

(b) Mutation on sequence of the clusters: the mutation site $z = 5$ (marked with circle) is randomly selected. The sequence value is mutated to 4.
A random number $\beta \epsilon [0, 1]$ determines the crossover site among the switch times or the sequences.

Since each individual represents the sets of switch times and the sequence, the dissimilarity concept should be considered for both sets.

**Dissimilarity between two set of switch times** $\tau(1), \ldots, \tau(W)$ and $\tau'(1), \ldots, \tau'(W)$ is defined by:

$$\Delta_1 = \frac{\sum_{i=1}^{W} |\tau(i) - \tau'(i)|}{W \cdot (T - 1)}$$

**Dissimilarity $\Delta_2$** between two cluster sequences is the number of unequal indices in the two sequences divided by $W - 1$. For instance, the two cluster sequences: 

\{1, 2, 1, 3, 2, 1, 4, 2, 1, 3\} and \{1, 2, 3, 2, 4, 1, 2, 3, 2, 4\} 

with $W = 9$, have 7 different indices in the same locus, resulting $\Delta_2 = 7/8$.

The dissimilarity $\Delta$ between the two individuals is obtained by:

$$\Delta = \Delta_1 + \Delta_2$$
Finding cluster parameters

Each individual represents both the switch times and the sequence of the clusters in the time series. Thus, unique values of \( c(t) \) are derived from each individual for \( c \in \{1, \ldots, C\} \) and \( t \in \{1, \ldots, T\} \).

According to each individual \( \mu_c(t) = 1 \), if \( x(t) \) belongs to the cluster \( c \) at time \( t \). Otherwise \( \mu_c(t) = 0 \).

In the next step, the cluster parameters \( \theta_c \) should be obtained by solving the following unconstrained optimization:

\[
\sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x(t), \theta_c) \rightarrow \min_{\theta_c}
\]

Since the parameters of the \( C \) clusters are independent, this unconstrained optimization is converted to \( C \) separate optimizations. Depending on the mathematical model of the clusters, each optimization is solved either by the least-mean-square or the maximum likelihood:
Example 1: switched autoregressive model

\[ x_t = [a_1^c \ a_2^c \ \ldots \ a_p^c][x_{t-1} \ x_{t-2} \ \ldots \ y_{t-p}]^T + [b_1^c \ b_2^c \ \ldots \ b_q^c][u_{t-1} \ u_{t-2} \ \ldots \ u_{t-q}]^T + \text{noise} \]

\[ \varphi(t) = [x(t-1) \ x(t-2) \ \ldots \ x(t-p), \ u(t-1) \ u(t-2) \ \ldots \ u(t-q)]^T \]

\[ \theta_c = [a_1^c \ a_2^c \ \ldots \ a_p^c \ b_1^c \ b_2^c \ \ldots \ b_q^c] \]

\[ x(t) = \theta_c \cdot \varphi(t)^T + \text{noise} \]

\[ d(x(t), \theta_c) = \|x(t) - \theta_c \cdot \varphi(t)^T\|^2 \]

\[ \sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x(t), \theta_c) \rightarrow \min_{\theta_c} \]

\[ \Rightarrow \theta_c = \left[ \sum_{t=1}^{T} \mu_c(t) \cdot \varphi^T(t) \cdot \varphi(t) \right]^{-1} \left[ \sum_{t=1}^{T} \mu_c(t) \cdot x(t) \cdot \varphi^T(t) \right] \]
Example 2: Switched Gaussian Model

\[ x_t \sim N(\rho_m, \sigma_m^2) \]

\[ f_X(x_t) = \frac{1}{\sigma_c \sqrt{2\pi}} \exp \left( -\frac{(x_t - \rho_c)^2}{2\sigma_c^2} \right) \]

\[ d(x, \rho_c, \sigma_c) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_c^2) - \frac{1}{2\sigma_c^2} (x_t - \rho_c) \]

\[ \sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x, \rho_c, \sigma_c) \underset{\rho_c, \sigma_c}{\text{min}} \]

\[ \rho_c = \left( \frac{\sum_t \mu_c(t) \cdot x(t)}{T} \right) \]

\[ \sigma_c^2 = \left( \frac{\sum_t \mu_c(t) \cdot (x(t) - \rho_c)^2}{T} \right) \]

After finding the parameters of clusters, the value of \( \sum_{c=1}^{C} \sum_{t=1}^{T} \mu_c(t) \cdot d(x, \rho_c, \sigma_c) \) is considered as the fitness of each individual.
The steady-state GA with CD/RW method uses a hybrid replacement scheme where individuals with poor fitness and low contribution to diversity are replaced.

Definition: The contribution of diversity among an individual $i$ and a population $P$ (excluding individual $i$) is defined as minimum dissimilarity among the individual $i$ and all the members of the $P$.

- An initial random population is generated.
- At each iteration, the crossover operator is applied on two randomly selected individuals to generate one crossover child.
- The offspring is generated by mutating the crossover child.
- The individuals in the population with poorer fitness values than the offspring are considered and the individual $i_{\text{min}}$ with the lowest contribution of diversity is found. The contribution of diversity of $i_{\text{min}}$ is compared with the contribution of diversity of the offspring to the population (removing $i_{\text{min}}$ from the population). The offspring replaces $i_{\text{min}}$ if it has a better contribution of diversity.
- If the offspring does not provide more diversity than $i_{\text{min}}$, then the offspring is replaced with the individual with poorest fitness in the population.
The population size of the GA is assumed 50. After applying the proposed method, the best individual is found as {62, 253, 289, 323, 374, 1, 2, 1, 3, 2, 3} including the switch times and the cluster sequence.

The true and estimated parameters of three clusters show the accuracy of modeling:

\[ \hat{\theta}_c = \{0.3011, 0.6983; 0.6600, -0.1019; -0.3183, 0.4969\} \]

\[ \theta_c = \{0.3, 0.7; 0.7, -0.1; -0.3, 0.5\} \]
Simulation: SARX system

The true $x(t)$ and its estimated values based on the identified SARX model.

The minimum fitness of the steady-state GA for SARX model.

The result of simulations on several randomly SARX system with different values of $T$.

<table>
<thead>
<tr>
<th></th>
<th>$T = 100$</th>
<th>$T = 200$</th>
<th>$T = 500$</th>
<th>$T = 1000$</th>
<th>$T = 2000$</th>
<th>$T = 5000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-time</td>
<td>8</td>
<td>32</td>
<td>43</td>
<td>140</td>
<td>346</td>
<td>588</td>
</tr>
<tr>
<td>Frequency of success %</td>
<td>84</td>
<td>80</td>
<td>84</td>
<td>76</td>
<td>84</td>
<td>76</td>
</tr>
</tbody>
</table>
Simulation: switched Gaussian model

A time series with length $T = 5000$, $C = 5$, and $W = 8$ is generated. The Gaussian model of the time series in five clusters are as:

$$N(1, 0.2^2), N(1, 0.5^2), N(2, 0.2^2), N(2, 0.5^2) \text{ and } N(3, 1^2)$$

The switch times between the five clusters are in $t = \{700, 1500, 2000, 2400, 2700, 3500, 4000, 4500\}$.

- The population size of the GA is assumed 50.
- After applying the proposed method, the best switch times are found as: $t = \{701, 1501, 1999, 2401, 2700, 3502, 4000, 4500\}$.
- The estimated Gaussian models for five clusters are:

$$N(0.9946, 0.2022^2) \quad N(0.9970, 0.5017^2) \quad N(1.9947, 0.2068^2) \quad N(2.0121, 0.4706^2) \quad N(2.0299, 1.1012^2)$$
Simulation: switched Gaussian model

The time series $x(t)$ with five clusters and nine switches. The *means* and *means±standard deviations* detected by identification algorithm are shown in red and green lines.

The minimum fitness of the steady-state GA for switched Gaussian model.

The result of simulations on several randomly Gaussian time series with different values of $T$

<table>
<thead>
<tr>
<th></th>
<th>$T = 100$</th>
<th>$T = 200$</th>
<th>$T = 500$</th>
<th>$T = 1000$</th>
<th>$T = 2000$</th>
<th>$T = 5000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run-time</td>
<td>13</td>
<td>21</td>
<td>33</td>
<td>101</td>
<td>141</td>
<td>437</td>
</tr>
<tr>
<td>Frequency of success</td>
<td>84</td>
<td>80</td>
<td>84</td>
<td>76</td>
<td>76</td>
<td>76</td>
</tr>
</tbody>
</table>
Future Works

- Extreme Event: refer to an event whose intensity exceeds expectations, has high impact, and is rare in the historical record: heat waves, droughts, tornadoes, and floods

- As the climate has warmed, some types of extreme weather have become more frequent and severe in recent decades,
  - Heat waves are longer and hotter.
  - Heavy rains and flooding are more frequent.
  - Drought is more intense and more widespread.

Question: Are extreme events changing in frequency and intensity as a result of human influences on climate?

- Statistical Analysis of extremes is a hot topic in climate studies.
Future Works

- Generalized Extreme Value (GEV) model: is used to model the largest or smallest value from a group of data.

\[
f(x; \mu, \sigma, \xi) = c \exp\left(-\left[1 + \xi \frac{x - \mu}{\sigma}\right]^{-\frac{1}{\xi}}\right)
\]

\[
1 + \xi \frac{x - \mu}{\sigma} > 0 \quad \sigma > 0
\]

- Non-stationary GEV:

\[
\mu(U) = \beta_1^T U \\ \xi(U) = \beta_2^T U \\ \sigma(U) = \beta_3^T U
\]

- U is the set of covariates, possibly includes the time, greenhouse gases, etc.

- Finding several clusters in extremes time series, while each cluster has different \( \beta_1 \), \( \beta_2 \) and \( \beta_3 \) shows how probability of extremes have been changed during time affected by human.
References


