Learning from a Stream of Nonstationary Data for Multiobjective Evolutionary Computation

Jianyong Sun

School of Mathematics and Statistics
Xi’an Jiaotong University
jy.sun@xjtu.edu.cn
Table of Contents

1 Background
   • Learning-based Multiobjective Evolutionary Algorithms

2 Computational Cost on Learning

3 How to Reduce Cost

4 Learning from Non-stationary and Dependent Data

5 Experiment Results

6 Conclusion
Multiobjective Optimization Problems (MOPs)

\[
\begin{align*}
\min & \quad F(x) = (f_1(x), \cdots, f_m(x))^T \\
\text{s.t.} & \quad x = (x_1, \cdots, x_n)^T \in \Omega
\end{align*}
\]

where

- \( \Omega = \prod_{i=1}^{n} [a_i, b_i] \subseteq \mathbb{R}^n \) defines the decision (search) space;
- \( a_i \) and \( b_i \) are the lower and upper boundaries of variable \( x_i \), respectively;
- \( x = (x_1, \cdots, x_n)^T \) is a vector of decision variable;
- \( F : \Omega \rightarrow \mathbb{R}^m: m \) objective functions \( f_i(x), i = 1, \ldots, m \)
Concepts

- **Pareto set (PS):** the set of optimal solutions
- **Pareto front (PF):** the image of PS
- **Goal:** Diversity and Convergence
Multiobjective Evolutionary Algorithms (MOEAs)

- Approximate the PS (PF) in a single run
- Environment selection and New Solution Creation
- Existing MOEAs
  - Selection based on Pareto dominance (need to consider diversity) [Deb et al., NSGA-II, 2002]
  - Selection based on performance metrics (hypervolume, Hausdorff distance, etc.) [Bader, et al., HypE, 2011]
  - Decomposition based: multiple single optimization problems, or multiple MOPs. Weight vectors to determine diversity, convergence measured by weighted objective values [Zhang and Li, MOEA/D, 2007]
  - Learning-based MOEAs
Learning-based Multiobjective Evolutionary Algorithms

- Focus on learning problem structure (pattern recognition) to aid new solution creation
- To create new solution,
  - choose a way to model *the data* → modelling
  - choose a method to make the model best fit the data → learning
  - create new solutions → variation
Learning-based MOEAs

- unsupervised learning
  - to approximate the manifold structure of MOPs
    - local PCA [Zhang, Zhou and Jin, RM-MEDA, 2008], clustering [Wang, Xiang and Cai, 2012], local learning [Li et al. 2011] and others
  - to estimate the probability distribution
    - hierarchical Bayesian optimization [Okebe et al. 2004], learning automata [Dai et al. 2016], and others
  - to estimate covariance and step-size: CMA-ES [Igel, Hansen and Roth, 2007], adaptive variance scaling [Bosman et al. 2007]
- supervised learning: model the PF to assist the search for the PS
  - model PS and PF jointly [Zhou, Zhang and Jin, 2009, Karshenas et al. 2014]
  - inverse modelling [Cheng, Jin et al. 2015]
Table of Contents

1. Background
2. Computational Cost on Learning
3. How to Reduce Cost
4. Learning from Non-stationary and Dependent Data
5. Experiment Results
6. Conclusion
the data

- at each generation, solutions in the present population
the data

- at each generation, solutions in the present population

the model

- specified in advance, e.g. Mixture of Gaussians, factor analysis, manifolds, etc.
the data

- at each generation, solutions in the present population

the model

- specified in advance, e.g. Mixture of Gaussians, factor analysis, manifolds, etc.

the learning

- Assumption: the data are i.i.d.
- iterative-convergent algorithms: K-means, local PCA, manifold learning, etc.
the data
- at each generation, solutions in the present population

the model
- specified in advance, e.g. Mixture of Gaussians, factor analysis, manifolds, etc.

the learning
- Assumption: the data are i.i.d.
- iterative-convergent algorithms: K-means, local PCA, manifold learning, etc.

the variation
- sampling
- alert: exploration and exploitation
Computational Cost

- High cost on learning
  - MoG: each iteration $O(NKn^2)$ where $n$ is the dimension
  - local PCA: $k$-PCA, $O(n^3 + N^3)$ where $N$ is the population size
- Is it necessary to pay such cost on learning?
  - focus on optimising original problems, yet to deal with the learning problems?
  - learned model can only approximate the theoretical solution structure gradually
  - extensive learning in early stages is NOT necessary
Table of Contents

1. Background
2. Computational Cost on Learning
3. How to Reduce Cost
4. Learning from Non-stationary and Dependent Data
5. Experiment Results
6. Conclusion
- re-use previously learned model if data does not change much
- use less iterations at each generation for the learning
- consider the evolutionary procedure as a whole
  - data fed into the learning algorithm sequentially: a stream of data
  - single-pass online learning algorithm
Table of Contents

1. Background
2. Computational Cost on Learning
3. How to Reduce Cost
4. Learning from Non-stationary and Dependent Data
5. Experiment Results
6. Conclusion
Data are dependent

- solutions in adjacent generations at early stage, are not identically distributed
- solutions along the evolution are not independent since later solutions are created based on previous solutions.
- data are not i.i.d.
Data are dependent
- solutions in adjacent generations at early stage, are not identically distributed
- solutions along the evolution are not independent since later solutions are created based on previous solutions.
- data are not i.i.d.

Data are non-stationary and scale-variant
- data is temporal and changes dynamically
- data is scale-variant → underlying structure is scale-variant
  - short-scale, data is pseudo-stationary.
  - long-scale, data is converging.
How to learn from dependent and non-stationary data?

- the data: each non-dominated solution along the evolutionary procedure
- the model: finite mixtures of Gaussian
- the learning: modify online agglomerative clustering
- the variation

Model

- Theory: mixtures of Gaussian are dense in the set of probability distributions with respect to weak topology.
- Practice: other models, such as manifold, factor analysis, are also possible.
Online Agglomerative Clustering – AddC

- **Initialization**: each data is considered as a cluster
- **cluster assignment**: each new data is assigned to the cluster that is the closest to it → minimise the within cluster variance
- **new cluster formation**: each new data is considered to be a new cluster if there is less than $K_{\text{max}}$ clusters → account for the temporal changes in the distribution of the data
- **cluster merging**: if there are more than $K_{\text{max}}$ clusters, merge two redundant clusters which has the closest distance between them → maximize the distance between clusters
Figure: A typical run of AddC on a set of non-stationary data with $K_{\text{max}} = 10$. 
Modifying AddC

- Combine the learning within environmental selection (hypervolume metric based)
- For each new data $y$, use fast non-dominated sorting approach to partition $A \cup y$ into $L$ fronts
- Decide if $y$ can be kept in $A$ by calculating its contribution to the hypervolume
- If $y$ is kept, **(do learning)**
  - remove the worst solution $x^*$: delete it from its cluster $C^k$, and update the centroid $z^k$ and the number of data in its cluster:
    \[
    c^k \leftarrow c^k - 1, \quad z^k \leftarrow z^k - \frac{x^* - z^k}{c^k}.
    \]
  - take $y$ as a new cluster if the current number of clusters is less than $K_{\text{max}}$; else merge two clusters that are closest.
Variation

- Sampling around each cluster centroids encourages exploitation
- To balance exploitation and exploration
  - mating control probability $\beta$, DE and polynomial mutation
  - with probability $\beta$, solutions within a cluster are used as reference solutions $\rightarrow$ exploitation
  - with probability $1 - \beta$, reference solutions are selected from a mating pool constructed by selecting a solution from each cluster $\rightarrow$ exploration
OCEA: Online Clustering based Evolutionary Algorithm

River OCEA, Romania
Table of Contents

1. Background
2. Computational Cost on Learning
3. How to Reduce Cost
4. Learning from Non-stationary and Dependent Data
5. Experiment Results
6. Conclusion
Experiment Results

Test Instances and metrics
- GLT1-6 and WFG1-9
- PF: discontinuous, degenerate, multimodal
- IGD and HV

Compared Algorithms
MOEA/D-DE, TMOEA/D, RM-MEDA, NSGA-II with efficient non-dominated sort, SMS-MOEAs
Figure: Evolution of the statistics of IGD metric values obtained by MOEA/D-DE, TMOEA/D, RM-MEDA, NSGA-II, SMS-EMOA and OCEA on GLT1-6
Figure: Final approximated fronts obtained by OCEA.
Hypothesis Test on Non-stationarity and Dependence

- the evolution search process is governed by a stochastic process
- from the first to the last generation, consider the trace of the population as a multivariate time series, i.e. it is a realisation of the stochastic process
- Augmented Dickey-Fuller test (ADF) can be used to test the stationarity of the time series.
  - It can only be used for univariate process.
  - No multivariate hypothesis test available.
- Autocorrelation function (ACF) can be used to test the dependency of the time series.
### ADF Hypothesis Test

**Table:** The p-values obtained by the ADF hypothesis test on the evolution procedure for the GLT test suite at 5% significance level. Entity value bigger than 0.05 suggests that the corresponding univariate process is non-stationary.

<table>
<thead>
<tr>
<th>test suite</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
<th>$d_7$</th>
<th>$d_8$</th>
<th>$d_9$</th>
<th>$d_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLT1</td>
<td>0.600</td>
<td>0.705</td>
<td>0.854</td>
<td>0.906</td>
<td>0.776</td>
<td>0.864</td>
<td>0.855</td>
<td>0.686</td>
<td>0.129</td>
<td>0.362</td>
</tr>
<tr>
<td>GLT2</td>
<td>0.708</td>
<td>0.152</td>
<td>0.379</td>
<td>0.382</td>
<td>0.447</td>
<td>0.434</td>
<td>0.400</td>
<td>0.141</td>
<td>0.056</td>
<td>0.001</td>
</tr>
<tr>
<td>GLT3</td>
<td>0.004</td>
<td>0.672</td>
<td>0.723</td>
<td>0.766</td>
<td>0.803</td>
<td>0.730</td>
<td>0.654</td>
<td>0.438</td>
<td>0.009</td>
<td>0.425</td>
</tr>
<tr>
<td>GLT4</td>
<td>0.724</td>
<td>0.789</td>
<td>0.865</td>
<td>0.845</td>
<td>0.930</td>
<td>0.864</td>
<td>0.777</td>
<td>0.584</td>
<td>0.066</td>
<td>0.149</td>
</tr>
<tr>
<td>GLT5</td>
<td>0.813</td>
<td>0.755</td>
<td>0.579</td>
<td>0.463</td>
<td>0.115</td>
<td>0.001</td>
<td>0.366</td>
<td>0.570</td>
<td>0.622</td>
<td>0.683</td>
</tr>
<tr>
<td>GLT6</td>
<td>0.870</td>
<td>0.723</td>
<td>0.383</td>
<td>0.223</td>
<td>0.0769</td>
<td>0.051</td>
<td>0.218</td>
<td>0.337</td>
<td>0.488</td>
<td>0.582</td>
</tr>
</tbody>
</table>
**Dependence Test**

Figure: The autocorrelation function for lag 0 through 20 on the mean evolutionary time series obtained by OCEA on GLT1. \( d_1, \ldots, d_{10} \) represent the variable dimension.
### Experiment Results

#### Parameter Sensitivity Analysis

<table>
<thead>
<tr>
<th>$K_{\text{max}}$ Values</th>
<th>GLT1</th>
<th>GLT2</th>
<th>GLT3</th>
<th>GLT4</th>
<th>GLT5</th>
<th>GLT6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{max}}=4$</td>
<td>0.05</td>
<td>0.1</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{\text{max}}=5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{\text{max}}=7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{\text{max}}=10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{\text{max}}=20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta$ Values</th>
<th>GLT1</th>
<th>GLT2</th>
<th>GLT3</th>
<th>GLT4</th>
<th>GLT5</th>
<th>GLT6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta=0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta=0.6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta=0.7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta=0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta=0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$CR$ Values, $F$ Values</th>
<th>GLT1</th>
<th>GLT2</th>
<th>GLT3</th>
<th>GLT4</th>
<th>GLT5</th>
<th>GLT6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CR=1, F=0.1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CR=1, F=1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure:** The mean values and standard deviations of the IGD metric values of approximated fronts obtained by OCEA with different $K_{\text{max}}, \beta, F, CR$ values over 22 independent runs on GLT1-GLT6.
Clustering Effectiveness Analysis

- The evolution procedure couples with the online clustering procedure

**Figure:** The clusters of final approximated sets obtained by OCEA for GLT1-GLT6
Table of Contents

1. Background
2. Computational Cost on Learning
3. How to Reduce Cost
4. Learning from Non-stationary and Dependent Data
5. Experiment Results
6. Conclusion
Conclusion

- incorporate a *online clustering* to address the temporal, non-stationary and dependent nature of the evolutionary search.
- experimental study showed effectiveness.
- comparisons showed the scheme indeed improve the search efficiency (in terms of search speed) and effectiveness (in terms of the quality of the final approximated sets and fronts).
- the non-stationarity and dependence of the evolutionary time series has also been confirmed empirically.