Modelling a Penicillin Fermentation Process Using Attention-Based Echo State Networks Optimized by Covariance Matrix Adaption Evolutionary Strategy

Kai Liu, Jie Zhang

School of Engineering, Merz Court, Newcastle University, Newcastle upon Tyne NE1 7RU, UK.
jie.zhang@newcastle.ac.uk

Abstract

Echo state network (ESN) has emerged as an effective alternative to conventional recurrent neural networks due to its simple training process and good modelling ability for solving a variety of problems, especially time-series modelling tasks. To improve modelling capability and to decrease the reservoir topology complexity, a new attention mechanism based ESN optimised by covariance matrix adaption evolutionary strategy (CMA-ES) is proposed in this paper. CMA-ES is a stochastic and derivative-free algorithm for solving non-linear optimization problems. Attention mechanism is incorporated to guide ESN to focus on regions of interest relevant to the modelling task. The proposed optimised ESN with attention mechanism is used to model a fed-batch penicillin fermentation process and the results are better than those from the standard ESN and ESN with attention mechanism.

Keywords: Echo State Network, Attention Mechanism, CMA-ES, Fed-batch Bioprocess

1. Introduction

Fed-batch fermentation processes are widely used in the pharmaceutical industry. In fed-batch fermentation processes, the maximization of yield is often regarded as the main objective, but the features of fed-batch fermentation processes including strong nonlinearity, non-steady-state condition, batch-to-batch variations, and strong time-varying condition make the yield hard to be predicted (Ashoori et al., 2009).

Due to the increasing demand on product quality and safety, optimization of fed-batch fermentation processes is becoming very important. One optimization approach is to use first principle models and stochastic optimisation algorithms such as evolutionary algorithm (EA), differential evolution (DE) and particle swarm optimization (PSO) (bin Mohd Zain et al., 2018). Another optimization approach for fed-batch bioprocesses is to use data-driven models such as artificial neural networks. Yu (2012) presented a Bayesian inference based two-stage support vector machine for soft sensor development in batch bioprocesses. Chen et al. (2004) proposed a cascade recurrent neural network combining with modified genetic algorithm for the modelling and optimisation of fed-batch fermentation processes.

In recent years, recurrent neural networks (RNNs) attracted a mass of attention because of its dynamic temporal nonlinear behaviour and processing arbitrary sequences of
inputs by its internal memory. RNNs are appropriate for modelling complex dynamic processes such as fed-batch fermentation processes. RNNs include but no limit to long short-term memory networks (LSTM), gated recurrent unit (GRU) and reservoir computing (RC). Comparing to LSTM and GRU, the benefit of RC is lower computational training cost and faster convergence with excellent performance. Due to the randomly generated input scaling matrix, the normal ESNs are not able to distinguish the different property of input elements. The concept of “attention mechanism”, which allows models to learn alignments between different modalities, has drawn significant attention in the training of neural networks. Attention mechanisms have been successfully used in speech translation and image caption generation (Luong et al., 2015), but have rarely been used in modelling complex bioprocesses, especially fed-batch fermentation process with a large number controllable and monitoring variables.

In this paper, an input elements scaling method based on the attention mechanism is integrated with ESN which is optimized by covariance matrix adaption evolutionary strategy (CMA-ES). Three global reservoir parameters in ESN are optimized by CMA-ES and they are reservoir size, spectral radius, and leak rate.

2. Methodology

2.1. Echo state networks

An ESN is composed of a reservoir and a linear output layer which maps the reservoir states to the network output. Figure 1 shows the original ESN. The input weights are generated randomly. The internal weights between reservoir neurons can be created with a sparse connection density which means that internal neurons may not be fully connected to each other but connected sparsely. The weights mentioned above will not change during the training process and only the readout output weights need to be learned. The reservoir states of ESN with leak rate are shown in Eq(1) (Lukoševičius and Jaeger, 2009):

\[ \mathbf{x}(t) = (1 - \alpha) \cdot \mathbf{x}(t - 1) + \alpha \cdot f(\mathbf{W}^{in} \cdot \mathbf{u}(t) + \mathbf{W} \cdot \mathbf{x}(t - 1) + \mathbf{W}^{back} \cdot \mathbf{y}(t - 1) \]  

where \( \mathbf{u}(t) \) and \( \mathbf{x}(t) \) are the inputs and the reservoir states at time \( t \) respectively, \( \alpha \) is the leak rate, \( f() \) is the general activation function, and the weights donated by \( \mathbf{W}^{in}, \mathbf{W}^{back}, \) and \( \mathbf{W} \) represent the weights for inputs, feedback, and reservoir respectively. Then \( \mathbf{W} \) needs to be rescaled by a spectral radius (the largest absolute eigenvalue of \( \mathbf{W} \), \( |\theta|_{\text{Max}} \) ) and then multiplied by a spectral radius factor \( \delta \) as shown in Eq(2) to keep its echo state property.

\[ \mathbf{W} \leftarrow \delta \mathbf{W} / |\theta|_{\text{Max}} \]  

The readout matrix is then obtained by solving a linear regression problem:

\[ \mathbf{X} \cdot \mathbf{W}^{out} = \mathbf{Y} \]  

where \( \mathbf{X} \) is a matrix of hidden states which are updated at discrete time steps using Eq(1) and \( \mathbf{Y} \) is the corresponding target outputs.

Ridge regression has been shown to be an efficient method to calculate the readout matrix (Dutoit et al., 2009). Ridge regression is a shrinkage method that consists of adding a penalty term proportional to the Euclidean norm of the readout matrix:

\[ \mathbf{W}^{out} = \arg \min_{\mathbf{w}} (\|\mathbf{Xw} - \mathbf{Y}\|^2 + \gamma \|\mathbf{w}\|^2) \]  

where \( \gamma \geq 0 \) is the ridge parameter determined on a hold-out validation set. The solution of readout matrix is given as:

\[ \mathbf{W}^{out} = (\mathbf{X}^T \mathbf{X} + \gamma^2 \mathbf{I}_N)^{-1} \mathbf{X}^T \mathbf{Y} \]  

where \( \mathbf{I}_N \) is the identity matrix of size \( N \).
Figure 1. Typical structure of an echo state network

2.2. Covariance matrix adaption evolution strategy

CMA-ES is a well-established evolutionary algorithm for real-valued optimization with many successful applications. The main advantages of CMA-ES lie in its invariance properties, which are achieved by carefully designed variation and selection operators and its efficient adaptation of the mutation distribution. The CMA-ES is invariant against order-preserving transformations of the fitness function value and in particular against rotation and translation of the search space. It has been demonstrated by experiments that the covariance matrix $\mathbf{C}(g)$ is similar to the inverse of the Hessian matrix of the problem at the optimum point. CMA-ES is particularly useful on non-smooth and ill-conditioned problems as it estimates a covariance matrix using an iterative procedure.

In CMA-ES, the population of new search offspring $\mathbf{x} \in \mathbb{R}^n$ is generated by sampling a multivariate normal distribution (Hansen et al., 2003):

$$\mathbf{x}^{(g+1)}_k \sim \mathbf{m}(g) + \sigma(g) \cdot \mathbf{N}(0, \mathbf{C}(g)) \quad \text{for } k = 1, \ldots, l$$

where $\mathbf{x}^{(g+1)}_k$ denotes the $k$th offspring at the $(g+1)$th generation; $\mathbf{m}(g)$ is the mean value of the search distribution at generation $g$; $\mathbf{N}(0, \mathbf{C}(g))$ is a multivariate normal distribution with zero mean and covariance matrix $\mathbf{C}(g)$; and $\sigma(g)$ is the step-size at generation $g$.

After those $l$ individuals have been created, they are evaluated on the objective function which is the mean squared errors (MSE) of the ESN and sorted according to their objective function values.

2.3. Attention mechanism

Attention mechanism in neural networks was proposed to guide networks to focus on regions of interest relevant for a particular modelling task. This prunes the network search space and avoids computing features from irrelevant input data (Luong et al., 2015). In order to make predictions based upon the relative input data, a variable-length alignment vector $\mathbf{a}_t$, whose size equals the input dimension, is derived by comparing the current target state $\mathbf{s}_t$ with input hidden state $\mathbf{h}_{ut}$:

$$\mathbf{a}_t = \text{align} (\mathbf{s}_t, \mathbf{h}_{ut}) = \text{sigmoid} [\text{score} (\mathbf{s}_t, \mathbf{h}_{ut})]$$

(7)

Here, score is obtained as:

$$\text{score} (\mathbf{s}_t, \mathbf{h}_{ut}) = \mathbf{s}_t^T \mathbf{W}_a \mathbf{h}_{ut}$$

(8)

where $\mathbf{W}_a$ is a rescale matrix of $\mathbf{s}_t$ and $\mathbf{h}_{ut}$.

The original input matrix is pruned by $\mathbf{a}_t$ and updated input data $\mathbf{u}_t$ is calculated as:

$$\mathbf{u}_t = \mathbf{a}_t \odot \mathbf{u}_t$$

(9)
where $\odot$ denotes element-wise multiplication.

3. Proposed modelling strategy

This paper proposes using CMA-ES to optimize the structure parameters of ESN with attention mechanism (Atten-ESN) for nonlinear process modelling. Figure 2 shows the flow chart of the proposed algorithm. The procedure can be summarized as follows:

1. Data for model building are divided into three sets: training data, testing data, and unseen validation data, and then they are normalized to have zero mean and unit variance.
2. Prune the input matrix by global attention mechanism.
3. Establish an ESN with random $R$, $\alpha$, and $\delta$ in the range based on sufficient internal units as default. The activation function used here in the hidden layer (reservoir) is $f = \tanh$ and the input weights and reservoir weights are generated randomly.
4. Train the established ESN with training data using ridge regression.
5. Optimize the Atten-ESN by CMA-ES. The MSE on the testing data is used as objective function. The optimization objective is to upgrade the values of $R$, $\alpha$, and $\delta$ to minimize the MSE on the testing data.
6. Test the optimized Atten-ESN (O-Atten-ESN) on the unseen validation data.

Figure 2. Graphical illustration of the proposed approach

4. Experiments

The benchmark industrial penicillin fermentation simulator, IndPenSim (Goldrick et al., 2015), is used to produce simulated process operation data. The simulator code in Matlab R2013b is available to download at www.industrialpenicillinsimulation.com. Three benches of data generated by IndPensim are used in model development. One batch is used as training data, the second batch is used as testing data, and the third batch is used as the unseen validation data. The penicillin concentration is taken as the target output and 30 controllable and monitoring variables are used as model inputs. The fitness function to minimize is the MSE of the ESN on the training data:

\[
MSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}
\]

where $y_i$ and $\hat{y}_i$ are target value and predicted value at sample $i$ respectively, and $N$ is the number of data samples.
5. Results and discussion

In order to investigate the performance of O-Atten-ESN, it is compared with standard ESN, Atten-ESN, and standard LSTM. The comparison between standard ESN and Atten-ESN is to investigate the effect of integrating attention mechanism with ESN. The comparison between Atten-ESN and O-Atten-ESN is to illustrate the effect of optimization using CMA-ES. In all ESNs, the three structural parameters, reservoir size, leak rate, and special radius, are randomly generated in the ranges of \([1-1000]\), \([0-1]\), and \([0-1.5]\) respectively. The initial step size of CMA-ES is 0.1 and the ridge regression parameter is 0.005.

The prediction results on unseen validation batch with different methods are shown in Figure 3 with the MSE of each method shown in the legend. Figure 4 shows the corresponding prediction errors. It can be seen in Figure 3, comparing to the predictions of the standard ESN, predictions of both O-Atten-ESN and Atten-ESN are closer to target values, especially when the slope of the target values is steep. This illustrates that the attention mechanism can take out some irrelevant input signals to reduce their influence. In other words, attention mechanism can increase the model robustness. Figure 4 shows that the errors of O-Atten-ESN are much smaller than those of the other two methods, demonstrating that high dimensional optimized algorithm such as CMA-ES can optimize ESN by searching better structure parameters. Table 1 gives the average MSE values and standard deviations of the three methods which were run 50 times with different random parameters. In summary, the attention mechanism and CMA-ES can improve ESN on predicting penicillin concentration in the fed-batch fermentation process in terms of prediction accuracy and robustness.

![Table 1](image)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average MSE</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-Atten-ESN</td>
<td>0.0781</td>
<td>0.0122</td>
</tr>
<tr>
<td>Atten-ESN</td>
<td>0.9723</td>
<td>0.1740</td>
</tr>
<tr>
<td>Standard ESN</td>
<td>1.6871</td>
<td>0.4116</td>
</tr>
</tbody>
</table>

Figure 3. Predictions of penicillin concentration
Figure 4. Prediction errors of penicillin concentration

6. Conclusions

An ESN with attention mechanism and optimized by CMA-ES is proposed in this paper to model a fed-batch penicillin fermentation process. Based on the preliminary results, the attention mechanism and CMA-ES can improve standard ESN on modelling complex bioprocesses with enhanced accuracy and reliability. It is expected that the proposed method will be effective for modelling other complex bioprocesses.

References


