클러스터링과 유전자 알고리즘을 이용한 대규모 S-system 모델의 효율적 학습 방법*

정성원[O] 이광형[1] 이도헌[2]
KAIST 전자전산학과 전산학 전공[O]
KAIST 바이오시스템학과, AITrc[1]
KAIST 바이오시스템학과[2]
{swjung[O], khlee[1], dhlee[2]}@biosoft.kaist.ac.kr

An Efficient Method for Learning Large S-System Model using Clustering and Genetic Algorithm

Sungwon Jung[0] Kwang H. Lee[1] Doheon Lee[2]
Division of Computer Science, Department of Electrical Engineering & Computer Science, KAIST[0]
Department of BioSystems, AITrc, KAIST[1]
Department of BioSystems, KAIST[2]

요 약

S-System 모델은 동적 시스템을 기술하기 위한 여러 모델 중의 하나로서, 높은 표현력으로 인해 다양한 분야에서 사용되어져 오고 있다. 하지만 S-System 모델이 갖고 있는 많은 매개변수는 목표 시스템을 모델링하는 데에 있어 고려해야 할 탐색 공간의 넓이를 크게 증가시키는 단점을 갖고 있으며 그로 인해 고려될 수 있는 변수의 수는 극히 적은 수로 제한되어져 왔다. 일반적인 S-System 모델의 경우, n개의 변수로 이루어진 시스템을 모델링하는 데 결정되어져야 할 매개변수의 수는 $O(n^2)$이다. 본 논문에서는 시스템 내의 변수들을 서로간의 연관 정도에 따라 클러스터링하고, 클러스터 사이의 동적 모델링을 통해 고려하는 매개변수의 수를 $O(kn)$ $(k \le n)$으로 줄이는 방법을 제안한다. 매개변수 값의 탐색을 위해 유전자 알고리즘을 사용하며, 제안된 방법이 기존의 방법으로는 학습할 수 없었던 규모의 S-System 모델을 학습할 수 있는 가능성을 지님을 보인다.

1. 서 론

바이오정보학의 대두 이래, 생체 내의 기작을 시스템 모델로 기술하고자 하는 연구가 많이 이루어져 왔다. 특히 인간 게놈 프로젝트 이래, 유전자 사이의 기능적 관계를 밝히고자 하는 연구는 활발히 이루어지고 있는 분야 중 하나이다. 그러한 유전자 사이의 관계 분석을 위해, 유전자 발현 정보가 많이 활용되어져 오고 있다.

유전자 발현 정보를 이용하여 유전자 조절 네트워크를 모델링하고자 하는 연구는 정적 시스템 모델링과 동적 시스템 모델링으로 나뉠 수 있다. 그 중 동적 시스템 모델링은 시간 변화에 따른 유전자 사이의 조절 기작을 기술할 수 있는 방법으로 활용되어지고 있다. 그러나 동적 시스템 모델은 정적 시스템 모델에 비해 내재된 매개변수의 수가 많은 경우가 일반적이며, 이는 네트워크 구조의 복잡도 증가와 함께 대상 네트워크를 학습하기 힘들게 하는 주요 원인 중의 하나가 된다. 이러한 이유로 S-System 모델과 같은 동적 시스템 모델이 활용될 수 있는 대상 시스템의 규모는 불과 수 개의 변수로 기술되는 작은 규모의 시스템들로 제한되어져 왔다.

n개의 유전자 $X_1, X_2, \ldots, X_n$ 으로 이루어진 유전자 조절 네트워크를 S-System 모델을 이용해 기술하는 경우, 유전자 X_i 발현 양의 단위 시간당 변화량은 다음의 식으로 표현된다.

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}$$

이러한 형태의 S-system 모델 학습을 위해 매개변수의 탐색에 유전자 알고리즘을 이용하는 방법이 제안되었으며 [1][2], 또한 보다 큰 규모의 시스템 학습과 매개변수 학습의 정확도 향상을 위한 유전자 알고리즘 또한 제안되었다 [3]. 그 외에도 data collocation을 이용한 진화적 최적화를 이용한 방법도 제안되었다 [6]. 그러나 학습할 수 있는 대상 시스템은 여전히 5개 정도의 적은 변수를 갖는 시스템만으로 제한되어지고 있다. 보다 많은 변수들 사이의 S-system 학습을 위한 방법으로는,

* This work was supported by National Research Laboratory Grant (2005-01450) from the Ministry of Science and Technology. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM-SUR program for providing research and computing facilities.

전체 system을 한번에 추적하지 않고 각 변수별로 나누어 추적하는 방법이 제안되었다 [7][8].

본 논문에서는, S-system 모델을 이용한 보다 큰 규모의 유전자 네트워크의 동적 모델링을 위한 기법으로서 클러스터링과 유전자 알고리즘을 사용하는 방법을 제안한다. 제안한 방법은 탐색하는 매개변수의 수를 원래의 $O(n^2)$에서 $O(kn)$ $(k \le n)$으로 줄여 학습하며, 그 결과 기존의 유전자 알고리즘만으로는 학습하기 힘들었던 규모의 대상 네트워크를 학습할 수 있음을 보인다.

2. 제안된 대규모 S-System 모델 학습 방법

본 논문에서 제안하는 방법은, 대규모 S-System 모델 학습을 위해 클러스터링으로 문제의 크기를 제한하고 유전자 알고리즘을 이용해 S-System 모델의 매개변수 값을 탐색한다. 유전자 조절 네트워크의 S-System 모델링을 위해 시계열 발현 데이터가 입력으로 주어지며, 이 때 제안된 방법의 절차는 다음과 같다:

- Step 1 : 유전자 발현 데이터의 계층적 클러스터링
- Step 2 : 최대 $k(\le n)$개의 하위 노드만을 갖도록 클러스터 계층 구조의 재구성
- Step 3 : 계층 구조의 상위 레벨에서 하위 레벨로, 클러스터들 사이의 S-system 학습
- Step 4 : 상위 레벨의 S-system 모델을 반영한 하위 레벨 클러스터 사이의 S-system 학습

유전자 발현 데이터의 계층적 클러스터링은 일반적인 응집형 계층적 클러스터링을 이용하여 수행된다. 한 클러스터가 갖는 하위 노드 클러스터의 수를 $k(\le n)$개로 제한하기 위해 계층 구조를 재구성하는 과정은 기존 계층 구조에서 적절한 레벨을 선택해 그 사이의 레벨을 없애는 방식으로 수행될 수 있다. 한 클러스터가 갖는 하위 노드 클러스터의 수가 k개로 제한되는 경우, 한 클러스터 내부에서의 S-system 모델 학습은 k개의 변수를 갖는 S-system 모델 학습 문제가 된다. 따라서 전체 n개의 변수를 갖는 S-system 모델 학습을 보다 적은 k개의 변수만을 갖는 S-system 모델 학습 여러 번을 통해 수행할 수 있다.

클러스터간 S-system 모델 학습 과정에서는 각 클러스터 내에 존재하는 유전자들의 발현 값의 평균을 구해 대푯값으로 사용한다. 또한 상위 레벨의 S-system 모델을 하위 레벨 클러스터 사이에 반영하기 위해, 상위 레벨 클러스터의 대푯값을 하위 레벨 클러스터에 '추상 노드'로서 삽입한 후 클러스터 내의 S-system 모델을 학습하는 방법을 사용한다.

2.1 유전자 알고리즘을 이용한 클러스터간 S-system 모델 학습

클러스터간 S-system 모델 학습을 위하여, 각 클러스터들에 속한 유전자들의 발현 평균값을 해당 클러스터의 대푯값으로 이용한다. 얻어진 유전자 시계열 발현 데이터를 이용하여, S-system 모델의 매개변수 탐색에 다음과 같은 만족도 값을 이용한 유전자 알고리즘을 사용한다 [3].

$$E = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{X'_i(t) - X_i(t)}{X_i(t)} \right)^2 + cnT\{\sum_{i,j} |g_{ij}| + \sum_{i,j,i \neq j} |h_{ij}|\}$$

$$Fitness = \frac{1}{E}$$

n은 S-system 모델 내에 존재하는 변수의 수이며, T는 시계열 데이터 중 평가에 사용되는 값의 샘플링 포인트의 수이다. $X'_i(t)$는 현재 학습된 모델에서의 변수 X_i의 시간 t에서의 발현 값이며, $X_i(t)$는 주어진 학습 데이터에 존재하는 변수 X_i의 시간 t에서의 발현 값이다. E의 우측 변의 첫 번째 텀은 학습된 모델의 시계열 데이터와 원래 시계열 데이터 사이의 차이를 반영하며, 두 번째 텀은 S-system 모델에서의 지수 매개변수가 0에서 멀 수록 에러를 증가시킴으로서 성긴 네트웍 형태를 갖는 S-system 모델을 학습할 수 있는 확률을 높인다. c는 가중치를 나타낸다.

2.2 하위 레벨 클러스터의 S-system 모델 학습

클러스터간 S-system 모델 학습에 의해, 클러스터 C_j가 클러스터 C_i에 영향을 미치는 것으로 판별되었다고 가정하자. 서로 다른 클러스터에 속한 변수들 사이의 모델 연결을 위해 C_i 내부에 있는 변수들 사이에서의 S-system 모델 학습 과정에서, C_j 내부에 있는 변수들을 영향을 미칠 수 있는 변수로 고려한다. 이를 위해 C_i 내부의 변수들 사이의 S-system 모델 학습시 C_j의 대푯값을 반영하는 추상 노드 A_{C_j}를 C_i 내부에 포함시켜 S-system 학습을 수행한다. C_i 내부의 한 변수 X_k의 변화량은 학습된 S-system 모델에서 다음과 같이 추상 노드 A_{C_j}를 포함하여 표현될 수 있다.

$$\frac{dX_k}{dt} = \alpha_k X_1^{g_{k1}} \times \ldots \times A_{C_j}^{g_{kC_j}} - \beta_k X_1^{h_{k1}} \times \ldots \times A_{C_j}^{h_{kC_j}}$$

$C_j = \{X_l, X_{l+1}, \ldots, X_{l+k-1}\}$일 때, 추상 노드 A_{C_j}에 의한 X_k에의 영향은 $A_{C_j}^{g_{kC_j}}$와 $A_{C_j}^{h_{kC_j}}$를 각각 다음 식의 형태와 같이 재구성함으로서 C_j에 속한 변수들에 의한 X_k에의 영향으로 표현될 수 있다.

$$A_{C_j}^{g_{kC_j}} = X_l^{g_l} \times X_{l+1}^{g_{l+1}} \times \ldots \times X_{l+k-1}^{g_{l+k-1}}$$

$$A_{C_j}^{h_{kC_j}} = X_l^{h_l} \times X_{l+1}^{h_{l+1}} \times \ldots \times X_{l+k-1}^{h_{l+k-1}}$$

위 식을 만족시키는 지수 매개변수 g와 h를 각각 찾기 위해, 좌변과 우변의 시계열 유전자 발현 값이 보다 잘 일치하게 하는 매개변수 값을 유전자 알고리즘을 이용하여 탐색한다. 이를 위해 다음 식으로 표현되는 만족도 값을 이용한다.

$$E = \sum_{t=1}^{T} \left(A_{C_j}^{g_{kC_j}}(t) - \prod_{m=0}^{k-1} X_{l+m}^{g_{l+m}} \right)^2$$

$$Fitness = \frac{1}{E}$$

2.3 추상 노드의 선택적 활용

상위 레벨의 클러스터들 사이에서의 상호 조절 관계를 하위 클러스터 레벨에 전달하기 위해, 조절자 역할의 클러스터를 추

상 노드화 하여 조절되는 클러스터 내부에 대표 변수로 넣는 과정을 앞 섹션에서 언급하였다. 이 때 한 클러스터 내부에 추가되는 추상 노드의 수가 많을수록 S-system 모델 학습의 비용이 증가하게 된다. 이러한 문제를 해결하기 위해, 클러스터 C_i에 대한 조절자 역할의 클러스터 중 높은 조절도 값을 갖는 $p(<k)$개의 클러스터만을 선택하여 추상 노드로 사용한다. 이 때 한 클러스터 C_i에 대한 클러스터 C_j의 조절도 값 d_{ij}는 학습된 클러스터간 S-system 모델에서의 지수 매개변수 값을 이용하여 다음과 같이 정의된다.

$$d_{ij} = |g_{ij}| + |h_{ij}|$$

약한 조절 관계를 고려하지 않기 위해, 미리 정해진 경계값 θ보다 큰 조절도를 갖는 p개의 클러스터만을 추상 노드화할 대상으로 선택한다.

3. 실험 및 결과

제안된 방법의 성능을 평가하기 위해, 10개의 변수를 갖는 S-system 모델을 인공적으로 만들어 학습 대상으로 사용하였다 (Table 1).

$$\frac{dX_1}{dt} = 0.12X_3^{1.4} - 2.5X_2^{2.84}$$
$$\frac{dX_2}{dt} = 0.35X_2^{3} - 4.64X_7^{0.62}$$
$$\frac{dX_3}{dt} = 0.47X_2^{2.7} - 8.44X_3^{2.87}$$
$$\frac{dX_4}{dt} = 2.96X_6^{-0.44} - 14.99X_4^{0.62}$$
$$\frac{dX_5}{dt} = 1.36X_6^{3} - 15X_2^{1.2}X_5^{1.12}$$
$$\frac{dX_6}{dt} = 3.63X_5^{3} - 10.4X_8^{0.98}X_7$$
$$\frac{dX_7}{dt} = 3.23X_2^{2.01}X_4^{3}X_6^{2.46}X_7^{1.29}X_9^{1.35} - 14.51X_7^{2.35}$$
$$\frac{dX_8}{dt} = 0.04X_2^{-2.5} - 13.25X_2^{1.31}X_8^{0.94}$$
$$\frac{dX_9}{dt} = 0.44X_5^{0.65} - 13.67X_9^{1.7}$$
$$\frac{dX_{10}}{dt} = 0.23X_2^{-2.71}X_9^{1.77} - 10.72X_{10}^{1.64}$$

Table 1: 10개 변수를 지니는 벤치마크 S-system 모델

$$\frac{dX_1}{dt} = 0.52X_1^{1.8}X_2^{1.34} - 3.63X_2^{2.05}$$
$$\frac{dX_2}{dt} = 3.04X_2^{0.27} - 4.6X_2^{0.25}$$
$$\frac{dX_3}{dt} = 0.28X_2^{3} - 8.19X_3^{0.98}$$
$$\frac{dX_4}{dt} = 1.87X_6^{-0.5} - 14.67X_4^{0.77}$$
$$\frac{dX_5}{dt} = -12.79X_2^{1.35}X_5^{1.24}$$
$$\frac{dX_6}{dt} = 0.11X_1^{3}X_8^{2.69} - 9.93X_6^{0.99}X_7^{1.12}$$
$$\frac{dX_7}{dt} = 15X_5^{3}X_6^{3} - 13.81X_7^{2.24}$$
$$\frac{dX_8}{dt} = 0.02X_1^{1.48}X_2^{-3}X_5^{-2.02}X_6^{-2.83}X_8^{3} - 15X_2^{2.06}X_8^{1.11}$$
$$\frac{dX_9}{dt} = 0.16X_5^{0.82}X_9^{-0.4} - 13.01X_9^{1.76}$$
$$\frac{dX_{10}}{dt} = 0.02X_2^{-3} - 10.16X_{10}^{1.69}$$

Table 2: 학습된 S-system 모델

벤치마크 S-system 모델로부터 각기 다른 10개의 시계열 데이터를 생성하여 학습 데이터로 이용하였다.

실험에 사용된 조건은 다음과 같다. 평가를 위한 샘플링 포인트 값 10, 유전자 알고리즘에서의 크로모좀 수 65, 유전자 알고리즘의 최대 세대 수는 35,000, α와 β 값은 0에서 15사이로 제한하였으며, 지수 매개변수인 g와 h의 값은 -3에서 3 사이로 제한하였다. 적합도 값에 사용되는 가중치 c는 0.15로 하였으며, 제안된 방법에서의 k값은 5로 하였다. 조절도 경계값 θ는 0.3, 최대 추상 노드의 수 p는 4로 하였다.

동일한 데이터에 대하여 총 세 번의 실험을 수행하였으며, 계층적 클러스터링에 의해 다음과 같은 클러스터링 결과가 사용되었다.

$C_1 = \{X_2, X_4, X_7\}$, $C_2 = \{X_3\}$, $C_3 = \{X_9, X_{10}\}$,
$C_4 = \{X_1, X_5, X_6, X_8\}$
$C_5 = \{C_2, C_3, C_4\}$
$Root = \{C_1, C_5\}$

세 번의 실험 결과, 0이 아닌 지수 매개변수를 찾아내는 데 있어서의 recall 및 precision 값의 평균은 각각 60.7%와 60.2%였다. Table 2 에서는 세 번의 실험 중 하나의 결과를 보여 준다. Table 1의 벤치마크 S-system 모델 내에 존재하는 모두 28개의 0이 아닌 지수 매개변수 중, 19개가 학습된 S-system 모델에서 0이 아닌 지수로 결정되었으며 이것은 68%의 리콜 비율이다. 또한 학습된 모델에 존재하는 30개의 0이 아닌 지수 매개변수 중 19개가 정확한 0이 아닌 지수 값이므로 63%의 precision값을 보인다. 이 결과의 경우 전체 200개의 지수 매개변수 중 180개가 0인지의 여부가 정확하게 결정되었다. 실험 결과를 통해 제안된 방법이 대규모 동적 시스템 모델링을 위한 효과적인 접근 방법이 될 수 있음을 알 수 있다.

4. 참고문헌

[1] Tominaga, D. and Okamoto, M.. Design of canonical model complex nonlinear dynamics. In Proceedings of the International Conference on Computer Applications in Biotechnology, pages 85-90, 1998.

[2] Tominaga, D., Koga, N. and Okamoto, M.. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. In Proceedings of the Genetic and Evolutionary Computation Conference, pages 251-258, 2000.

[3] Shinichi Kikuchi, Daisuke Tominaga, Masanori Arita, Katsutoshi Takahashi and Masaru Tomita, Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics, 19(5), pages 643-650, 2003.

[4] Voit, E. O., Canonical Nonlinear Modeling, Van Nostrand and Reinhold, 1991.

[5] Zein, A., Kuffner, R., Zimmer, R. and Lengauer, T., Analysis of gene expression data with pathway scores, In Proceedings of the 2000 conference on Intelligent Systems for Molecular Biology (ISMB00), pages 407-417, 2000.

[6] Kuan-Yao Tsai and Feng-Sheng Wang, Evolutionary optimization with data collocation for reverse engineering of biological networks, Bioinformatics, 21(7), pages 1180-1188, 2005.

[7] Shuhei Kimura, Mariko Hatakeyama and Akihiko Konagaya, Inference of S-system Models of Genetic Networks from Noisy Time-series Data, Chem-Bio Informatics Journal, 4(1), pages 1-14, 2004.

[8] Shuhei Kimura, Kaori Ide, Aiko Kashihara, Makoto Kano, Mariko Hatakeyama, Ryoji Masui, Noriko Nakagawa, Shigeyuki Yokoyama, Seiki Kuramitsu and Akihiko Konagaya, Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, Bioinformatics, 21(7), pages 1154-1163, 2005.