다중 개체 집단의 공진화적 학습에 의한 바이오 데이터의 패턴 마이닝

김수진[01,2] 정제균[1,2] 장병탁[1,2,3]
서울대학교 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터[2]
서울대학교 컴퓨터공학부[3]
{sjkim[0], jgjoung, btzhang}@bi.snu.ac.kr

Pattern Mining of Biological Data by Co-evolutionary Learning with Multi-populations

Soo-Jin Kim[01,2] Je-Gun Joung[1,2] Byoung-Tak Zhang[1,2,3]
Graduate Program in Bioinformatics, Seoul National University[1]
Center for Bioinformation Technology, Seoul National University[2]
School of Computer Science and Engineering, Seoul National University [3]

요 약

현재 각 분야에서 다양한 실험 데이터가 산출되면서 이종(heterogeneous) 데이터 간의 상관관계 분석에 대한 중요성이 더욱 부각되고 있다. 특히, 대규모 실험에 의해 급속하게 증가하고 있는 대량의 바이오 데이터에서 이런 문제를 해결하기 위한 새로운 데이터 마이닝 방법이 요구된다. 본 논문은 특성이 다른 두 데이터 셋에서 서로 상관관계가 있는 부분 패턴을 파악할 수 있는 새로운 알고리즘을 제안한다. 제안한 알고리즘은 다중 개체 집단을 유지하면서 상호간 공진화하는 확률적 진화컴퓨팅 방법에 기반하고, 전체의 탐색 포인트들을 분해하여 최적해를 찾는 점에서 장점을 가지고 있다. 실험 결과, 본 논문에서는 효모 유전자에 대한 발현 데이터와 모티프 데이터의 이종 데이터에 적용해 보았으며, 이러한 데이터에 있어서 주요 상관관계가 있는 패턴들을 추출한 결과를 제시한다.

1. 서 론

최근, 다양한 분야에서 각종 실험 데이터가 대량 산출되면서 이종(heterogeneous) 데이터 간의 상관관계 분석을 위한 여러 방법들이 많은 관심을 받고 있다. 특히, 마이크로어레이(microarray) 데이터를 비롯해 급속하게 데이터가 증가하고 있는 생물학 분야에서도 특성이 다른 두 데이터 간의 분석의 중요성이 대두되고 있다. 이는 생체 내에서 발생하는 생물학 메커니즘이 매우 복잡하고 다양한 과정에 의해 발생되며, 또한 각 메커니즘들은 서로 영향을 주고받으며 전체적으로 조절되므로, 위와 같은 접근은 의미가 있다고 할 수 있다.

이전부터, 다양한 타입의 데이터(예를 들어, 문서와 단어, 저자)를 동시에 클러스터링(clustering)하는 비교사학습(unsupervised learning) 방법에 대한 몇몇 연구들이 진행되어 왔다. Dhillon[1]의 정보 이론 목적 함수(information-theoretic objective function)를 확장하여 데이터 간 여러 쌍의 상호작용(pairwise interactions)을 고려한 다중 클러스터링에 대한 연구[2], k-평균 군집화(k-means clustering)를 확장한 모델로 각 다른 데이터 타입을 통합하여 학습하는 클러스터링 알고리즘 연구[3] 등, 특성이 다른 데이터 간의 상관관계를 분석하는 여러 방법이 존재한다.

본 논문은 특성이 다른 두 바이오 데이터 셋에서 서로 상관관계가 있는 부분 패턴 분석이 가능한 새로운 데이터 마이닝 알고리즘 (Co-evolutionary Learning with Multi-populations: CLM)을 제안한다. 제안한 알고리즘은 다중 개체 집단을 유지하면서 상호 간 공진화하는 확률적 진화 컴퓨팅 방법에 기반하여, 전체의 탐색 포인터들을 분해하여 최적해를 찾는다.

본 논문에서는 CLM을 효모(*S. cerevisiae*)의 유전자 발현(gene expression) 데이터와 모티프(motif) 데이터의 이종 데이터에 적용해 보았다. 실험 결과, 이 데이터에 대하여 CLM은 효모의 유전자 발현 양상과 모티프 간에 높은 상관관계를 보이는 패턴들을 추출할 수 있음을 알 수 있었다.

2. 공진화적 학습에 의한 패턴 마이닝

N개의 유전자와 L개의 모티프로 구성된 $N \times L$ 모티프 데이터와 N개의 유전자와 M개의 컨디션으로 구성된 $N \times M$ 유전자 발현 데이터 간의 상관관계 분석 개요는 그림 1에서 대략적으로 설명하고 있다. 두 데이터 간 상관관계를 나타내는 최적 부분 패턴 매트릭스는 다중 개체 집단(multi-populations)의 공진화적(co-evolutionary) 학습을 이용하여 모티프 데이터에서 계산된 부분 매트릭스

점수($score_m$)와 유전자 발현 데이터에서 계산된 부분 매트릭스 점수($score_e$)의 합을 최소로 하는 적합함수(fitness function)에 의해 찾아진다. 집단의 각 개체(individual)는 확률 벡터의 샘플링에 의하여 업데이트 된다.

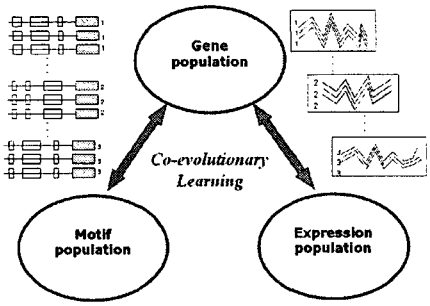

그림 1 공진화적 학습 개요

2-1. 적합함수

적합함수는 모티프 데이터와 유전자 발현 데이터로부터 추출된 각 부분 매트릭스(sub-matrix)에서 구한 모티프 점수와 유전자 발현 점수의 합을 최소로 하는 개체를 선택한다. 각 데이터에서 추출된 부분 매트릭스의 점수에 의해 발생되는 개체를 위한 적합함수는 식 (1)을 따른다. 함수 $g(score_m, score_e)$는 $score_m$과 $score_e$의 합을 의미한다.

$$f(z) = \min g(score_m, score_e) \tag{1}$$

2-2. 모티프 점수

모티프 점수는 각 유전자에서 특정 모티프의 존재의 유무에 따라 결정된다. 따라서 추출된 부분 매트릭스 $I \times J$의 각 요소 e_{ij}의 합을 최대화하여 계산한다. I와 J는 클러스터의 행과 열 인덱스(indices)를 가리킨다. 즉, 유전자와 모티프의 셋을 말하며, 각각은, $|I| \le |N|, |J| \le |L|$이다. 부분 매트릭스에 대한 각 요소의 전체 합은 식 (2)과 같이 구할 수 있다.

$$H_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} h_{ij} = \begin{cases} 1, & e_{ij} = 1 \\ 0, & e_{ij} = 0 \end{cases}, \tag{2}$$

모티프 데이터의 각 엔트리 요소 e_{ij}는 i 번째 유전자에서 특정 j 번째 모티프가 있으면 1, 그렇지 않으면 0으로 표현된다.

2-3. 유전자 발현 점수

유전자 발현 점수는 유전자 발현 데이터로부터 추출된 부분 매트릭스 $I \times K$의 잔차제곱평균(MSR: mean squared residue)을 최소화하여 계산한다. I와 K는 클러스터의 행과 열 인덱스(indices)를 가리킨다. 즉, 유전자와 컨디션의 셋을 말하며, 각각은, $|I| \le |N|, |K| \le |M|$이다. 부분 매트릭스의 전체 잔차제곱평균은 식 (3)에 의하여 구하여진다.

$$R_{ik} = \frac{1}{|I||K|} \sum_{i \in I, k \in K} r_{ik}^2, \tag{3}$$

잔차는 인덱스 셋 I와 K에 의해 결정되며, 식 (4)과 같은 식으로 계산 할 수 있다. 유전자 발현 데이터의 각 엔트리 요소 e_{ik}는 특정 k 번째 컨디션에서의 i 번째 유전자 발현 값을 나타낸다.

$$r_{ik} = e_{ik} - e_{iK} - e_{Ik} + e_{IK}, \tag{4}$$

$$e_{iK} = \frac{\sum_{k \in K} e_{ik}}{|K|}, e_{Ik} = \frac{\sum_{i \in I} e_{ik}}{|I|}, e_{IK} = \frac{\sum_{i \in I, k \in K} e_{ik}}{|I||K|}, \tag{5}$$

e_{iK}는 i번째 행에서 전체 K 열에 대한 평균이고, e_{Ik}는 k번째 열에서 전체 I 행에 대한 평균이며, e_{IK}는 부분 매트릭스에서 모든 엔트리 요소의 평균이다.

3. 데이터 및 실험 설계

본 논문에서는 Pilpel[4]에 의해 추출된 효모의 모티프 정보와 Spellman[5]의 효모의 유전자 발현 마이크로어레이 결과를 이용하였다. 모티프 데이터는 Pilpel의 데이터를 바탕으로 AlignACE를 이용하여 효모 각 유전자가 총 42개의 모티프 중 어떤 모티프를 가지고 있는지를 분석하여 매트릭스를 형성하였다. 또, 유전자 발현 데이터는 효모의 세포주기 상 시간의 흐름에 따라 총 18 시점에서 각 유전자의 발현량으로 매트릭스를 형성하였다.

본 실험에서는 총 6000여개의 효모 유전자들 중 Spellman의 분석 결과에서 세포 주기에 관련되어 있는 약 800여개 유전자에서 발현값이 존재하지 않는 데이터를 제외한 총 551개의 유전자 정보를 이용하여 실험을 수행하였다. 알고리즘의 파라미터 설정은 표 1과 같다.

파라미터	값
유전자 개체수(Pop_g)	1000
모티프 개체수(Pop_m)	200
유전자 발현 개체수(Pop_e)	400
세대수	200

표 1 파라미터 설정 초기 값

4. 실험결과

본 논문에서는 효모의 모티프 데이터와 유전자 발현 데이터를 이용하여 공진화적 학습에 의해 이종 데이터 간 상관관계를 분석하였다.

적합도는 각 점수의 합이 최소가 되는 시점으로 최적

화된다. 최종 학습된 적합도는 모티프 데이터 0.002, 유전자 발현 데이터 0.0084540이다. 그림 2는 유전자, 모티프, 유전자 발현 각 집단의 200세대동안의 적합도를 나타낸다.

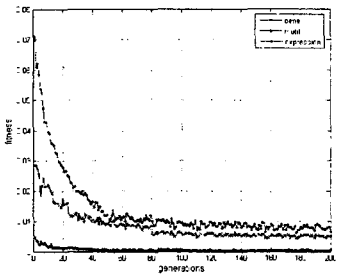

그림 2 세대에 따른 각 개체의 적합도

또한, 표 2는 학습된 결과에 의해서 선택된 유전자이다. 그림 3은 선택된 모티프 (SFF, SFF', ALPHA1') 부분 매트릭스 결과이고, 그림 4는 선택된 유전자의 발현 양상을 보여준다.

유전자 목록
YGR240C, YOR332W, YLR040C, YPR014C, YPR107C, YJR004C, YGL125W, YLR056W, YHR098C, YOR023C

표 2 선택된 10개의 유전자 목록

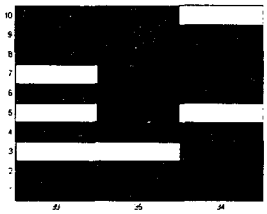

그림 3 모티프 데이터의 부분 매트릭스

선택된 모티프 중 SFF'는 세포 주기 상에서 중요한 역할을 수행하는 전사 인자 FKH1에 대한 결합 영역으로, 효모 세포 주기에 관련되어 있다고 알려진 모티프이다. 그림 3에서 가로축은 선택된 3개의 모티프로 39는 ALPHA1', 35는 SFF', 34는 SFF이고, 세로축은 선택된 유전자를 나타낸다. 또한, 매트릭스에서 진한 부분은 해당 유전자에서 각 모티프가 존재함을 나타내고, 흰 부분은 모티프가 해당 유전자에 존재하지 않음을 의미한다.

그림 4에서 가로축은 효모 세포 주기의 특정 시점이고, 세로축은 유전자 발현량을 나타낸다. 그림에서와 같이 특정 시간대 선택된 10개의 유전자 발현 양상이 비슷함을 알 수 있다.

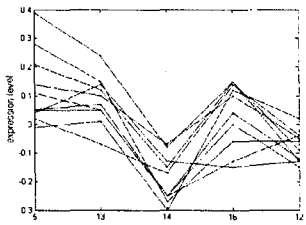

그림 4 선택된 유전자의 발현 양상

이를 통해 유전자 상류 지역(upstream region)에서 특정 전사 조절 인자(TF: transcription factor)의 결합에 따라 특정 시간대의 유전자 발현 양상이 비슷하게 나타나는 부분 패턴을 전체 탐색 포인트를 분해하여 발견 할 수 있음을 보여준다.

5. 결론

본 논문에서는 이종 데이터 간의 상관관계 분석을 위한 알고리즘을 제안한다. 제안한 알고리즘 CLM은 다중 개체 집단을 유지하면서 상호간 공진화하는 확률적 진화 컴퓨팅 방법에 기반하기 때문에 전체의 탐색 포인트들을 분해하여 최적해를 찾는 장점이 있다. CLM을 효모의 모티프 데이터와 유전자 발현 데이터에 적용해 본 결과, 특정 모티프의 존재 유무에 따라 유전자 발현 양상이 비슷하게 나타나는 부분 패턴을 발견할 수 있었으며, 두 데이터 간에 상관관계가 높음을 알 수 있었다.

향후 연구로써, 결과로 선택된 유전자 및 모티프들의 생물학적인 분석을 수행해야 하며, 더 나아가 여러 분야에서 산출되고 있는 다양한 데이터에 적용시켜 볼 수 있을 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실 사업(NRL)에 의하여 지원되었음.

참고문헌

[1] I. S. Dhillon, *et al.*, Information-theoretic co-clustering, *Proceedings of SIGKDD*, pp. 89-98, 2003.
[2] R. Bekkerman, *et al.*, Multi-way distribution clustering via pairwise interactions, *Proceedings of the 22nd International Conference on Machine Learning*, 2005.
[3] C. Yang, *et al.*, Clustering genes using gene expression and text literature data, *Proceedings of IEEE Computational Systems Bioinformatics Conference*, 2005.
[4] Y. Pilpel, *et al.*, Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, Vol. 29, pp. 153-159, 2001.
[5] P. T. Spellman, *et al.*, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol. 9, pp. 3273-3297, 1998.