멀티 소스 바이오 데이터 통합과 분석을 위한 새로운 접근 방법

윤혜성[O]* 이상호* 김주한**
이화여자대학교 컴퓨터학과*, 서울대학교 의과대학 생명의료정보학**
{comet[O], shlee}@ewha.ac.kr, juhan@snu.ac.kr

A New Approach for Multi-Source Bio-data Integration and Analysis

Hye-Sung Yoon[O]* Sang-Ho Lee* Ju Han Kim**
Dept. of Computer Science and Engineering, Ewha Womans University*, Seoul National University Biomedical Informatics(SNUBI), Seoul National University College of Medicine**

요 약

네트워크가 보편화되면서 어떠한 정보의 교환도 시간과 장소에 상관없이 가능하게 되었다. 자체 실험실에서 실험한 값을 포함하여 분산된 다양한 소스로부터 많은 실험값의 정보를 통합하는 즉, 멀티 소스 데이터에 대한 통합 규칙을 만들 수 있다면 다양하고 유용한 정보를 얻을 수 있을 것이다. 또한 통합된 규칙을 통해서 새로운 안목으로 실험을 진행할 수도 있으며, 미처 생각하지 못했던 관련 지식을 습득할 수도 있을 것이다. 본 논문에서는 이러한 분산된 데이터를 통합하여 멀티 소스 데이터들 간의 통합 규칙을 만들고 이의 분석 기반이 되도록 하는 방법에 대해 소개한다.

1. 서 론

생명현상에 대한 연구가 활발해짐에 따라 그 실험 결과의 분석을 위하여 전산학, 통계학, 수학 등의 다양한 접근방법이 적용되고 있다. 또한 계속적으로 발전하는 컴퓨터의 계산능력과 기억 용량, 인터넷의 보급, 자동화되는 비즈니스, 과학적 처리와 같은 기술 발전으로 인하여 실험에 의한 데이터의 크기도 빠르게 증가하고 있다. 이에 따라서 데이터의 분석 방법은 가정(assumption)을 세우고 많은 시간과 노력으로 실험을 하여 분석했던 과거의 가정 중심 방법으로부터, 실험 기술과 분석 방법의 발전으로 인하여 많은 시간과 노력을 줄일 수 있게 되면서 대용량의 데이터를 처리할 수 있는 데이터 중심 방법으로 옮겨 가고 있다. 이러한 실험 데이터들의 상당수는 지리학적으로 다양한 사이트에 분산되어 있다. 예전에는 한 실험실에서 데이터를 수집하고 하나의 시스템과 패스웨이에 따라 가정을 테스트하였던 반면에, 새로운 패러다임은 실험 결과를 여러 실험실의 실험 결과들을 모으고 공유하여 genome-wide의 실험적 결과를 결합시킬 것을 필요로 하고 있다[1]. 예를 들어, n개의 요소를 갖는 계통 데이터의 프로파일은 서열의 비교 분석을 위하여 n개의 완전한 지놈 서열이 필요하다. 이것은 반드시 한 실험실에서 만들어지지 않을 수 있다. 따라서 데이터 중심 모델은 대용량의 이종(heterogeneous) 데이터 셋을 처리할 수 있는 좀더 세련된 계산 기술이 필요하다[2].

본 논문에서는 데이터 중심의 흐름에 알맞은 분석을 위하여 여러 실험실에 저장되어 있는 데이터를 서로 모으고 공유하여 내가 가지고 있는 데이터와 관심 있는 데이터 분석으로 통합 규칙을 살펴봄으로써 보다 넓은 안목으로 문제에 접근하고 많은 정보를 이끌어낼 수 있는 방법론을 제시하고자 한다. 논문의 구성은 다음과 같다. 2절에서는 데이터 통합과 분산된 데이터를 처리하는 방법인 분산 데이터 마이닝에 대해서 살펴보고, 3절에서는 실험한 데이터와 제안 방법에 대해 소개한다. 그리고 4절에서는 3절의 방법에 대해서 실험한 결과에 대해 설명하고, 5절에서 결론을 맺는다.

2. 데이터 통합과 분산 데이터 마이닝

이 절에서는 여러 가지 생물학적 데이터를 가지고 있는 이종 사이트에서 분산된 데이터를 마이닝(distributed data mining)하기 위해서 고려해야 하는 점에 대해 설명한다. 통합 데이터를 분석하기에 앞서 우선 여러 사이트에 분산되어 있는 데이터를 통합해야 한다. 모든 데이터들이 공유될 수 있도록 네트워크화 되어 있다는 것과 다양한 실험에 의한 데이터를 통합하는 것이 하나의 데이터 셋을 가지고 실험한 결과보다 더 나은 이해를 이끌어 낼 수 있을 것이라고 가정한다.

2.1 데이터 통합

데이터를 통합하고자 하는 목적은 유전자 분류, 클러스터링, 유전자 조절 네트워크 등에서 다양하고 독립된 특성들을 이용하여 더 정확하고 풍부한 상호관계를 밝히고자 하기 위한 것이다. 예를 들어 단백질-단백질 상호작용 데이터가 있을 때, 이것은 단백질 상호관계망을 실험을 통해 정보를 획득하고, 관계망을 구축할 수 있다. 하지만 단백질 상호작용 데이터와 유전자 발현 데이터를 통합하면 단백질 상호관계망의 발현여부도 살펴볼 수 있게 된다. 따라서 외부 변화에 따른 반응을 설명할 수 있게 되어 또 다른 클러스터들을 형성할 수 있다. 현재까지 바이오 데이터 통합을 위하여 이용되어지고 있는 방법들은 다양한 데이터 셋을 기능적 구조적 클러스터링으로 보고 거리 계산 방법으로 클러스터링을 통합시켜 나가는 방법을 적용하거나[3] 그래프 모델링 방법을 적용하여 각 데이터 소스에서 비슷한 확률계산 값이 나오면 그래프를 통합하고 확장하는 방법 등을 적용하였다[4]. 하지만 무조건 다양한 사이트에서 필요에 따라 데이터를 통합하는 것은 하나의 유사하지 않은 데이터 셋으로 인하여 정확하지 않은 분석결과를 초래하여 신뢰도를 떨어뜨릴 수 있으므로 먼저 이러한 점을 고려하여 통합해야 한다. 따라서 다음은 이종 사이트에서 분산된 데이터를 마이닝하는 방법에 대해 살펴본다.

2.2 분산 데이터 마이닝

분산 데이터 마이닝은 분산된 데이터를 이용하여 데이터의 패턴을 찾는 문제를 다루는 분야이다. 비록 오늘날 대부분의 데이터가 중앙 집중화된 데이터 분석 시스템을 사용하고 있고 여러 기관, 회사, 국가 등의 모든 데이터를 공유할 수 없지만, 어떤 정보에 대해서는 공유가 필요할 때가 있다. 용량이 크고 분산된 데이터 셋을 마이닝하기 위하여 여러 가지 분산 알고리즘이 연구되었다. 하지만 이들의 단점은 여러 가지 알고리즘에 기반하여 데이터를 통합하고 동종(homogeneous) 데이터 셋만을 고려하여 데이터를 통합하거나 데이터 저장 공간과 시간 단축만을 고려하는 연구에 편중되어 있었다. 따라서 데이터 통합은 이루어졌지만 이를 활용하면 정확성은 떨어지는 경우가 많았으며 동종 데이터 셋만을 통합한 경우에는 데이터 크기는 커졌지만 활용에 있어서 다양한 정보는 이끌어내지 못하는 경우가 발생하게 되었다. 따라서 최근에는 유사성에 기반하여 데이터를 수집·통합하여 에러율을 줄이고자 하는 연구가 진행되고 있다[5].

3. 멀티 소스 실험 데이터 분석을 위한 알고리즘

본 장에서는 실험한 데이터에 대한 설명과 실험 방법에 따른 순서에 대해 설명한다.

3.1 데이터

수행된 실험은 지놈 데이터의 2가지 종류를 이용한다. 첫 번째 데이터 셋은 2,465개 yeast 유전자에서 79 유전자 발현 벡터 셋으로 구성되어 있는 값이며, 각각의 샘플은 두 가지 다른 실험 조건하에서 특정 유전자의 발현정도를 대수 비율(logarithm ratio)로 표현한 것이다. 두 번째 데이터 셋은 2,465개 yeast 유전자에서 24 요소 벡터 셋으로 계통 프로파일에서 비롯된 값이다. 이 프로파일은 관심 유전자와 완전한 지놈과의 동족체(homolog) 정도를 BLAST 버전 2.0으로 수행한 최저 E-값으로 음의 대수(negative logarithm)로 측정된 값이다. 유전자 칩 발현 데이터는 diauxic shift(DeRisi et al., 1997), mitotic cell division cycle(Spellman et al., 1998), sporulation(Chu et al., 1998) 그리고 temperature와 reducing shocks와 같이 유전자에 대한 여러 실험실의 실험 데이터를 모아놓은 것이다. 이것은 본 논문의 실험 목적인 여러 실험실에서 실험된 분산 데이터를 통합하고 다양한 소스의 데이터 셋과 함께 분석 기법을 적용하여 통합 규칙을 형성하는 목적에 부합되며, 이 데이터 셋은 Stanford 웹 사이트에서 이용 가능하다(http://www-genome.stanford.edu).

다음으로 이종 데이터인 분산되어 있는 다양한 데이터 셋을 통합하여 통합된 데이터 셋에서의 통합 규칙을 발견하는 방법을 설명한다. 본 논문에서는 분산되어 있는 데이터를 통합할 때에 먼저 다양한 종류의 데이터 셋을 보다 완전한 정보의 양을 가지는 데이터 구조 셋 순으로 데이터 셋을 정렬한다. 그림 1에서 하나하나의 회색 셀이 정보를 뜻한다면 P_1보다는 P_2가 정보의 양이 더 많으며 P_2보다는 P_3가 정보의 양이 더 많은 것으로 다양한 종류의 셋을 정보의 양으로 정렬한다.

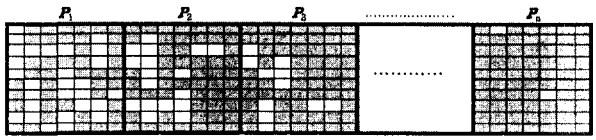

그림 1. 데이터 셋을 정보량의 순서로 정렬. $P_1 < P_2 < \cdots < P_n$은 다양한 소스의 데이터 셋을 정보량의 순으로 정렬함.

3.2 실험 방법

본 논문에서는 이종 데이터 분석을 위하여 사회 네트워크 분석(social network analysis) 방법을 적용하였다. 사회 네트워크 연결망은 구성원들간의 조직 내에 사회적 관계를 파악하고 상호작용을 통해 지식창출 및 공유 활동을 촉진할 수 있는 여건을 마련하는 접근방식이다[6]. 유전자들간의 상호 관계는 어떤 정해진 수학 공식으로 유사한 그룹을 만들어 풀 수 있는 방법이 아니라 여러 가지 특성이나 사회적인 관계 등의 다양한 이유 등을 고려하여 판단해야 할 것이다. 본 논문에서는 이러한 사회 네트워크 방법을 유사한 유전자 그룹을 찾아 규칙을 발견하기 위한 하나의 척도로 사용하였다. 즉, 여러 가지 분산되어 있던 다른 타입의 실험 데이터를 통합한다. 정보를 알고자하는 유전자에 대하여 통합되어진 데이터 셋들에서 보다 많은 정보량을 갖는 완전한 데이터 셋에서 사회 네트워크 분석을 적용하였다. 이것은 생명체내의 각종 생화학 물질 분자들을 네트워크의 노드로 취급하고 세포의 생명을 유지하는데 필요한 반응에 함께 참여하는 분자들 사이의 링크를 통하여 생명현상의 복잡성은 유전자의 수가 아니라 분자들 사이의 네트워크에 의해서 밝혀낼 수 있을 것이라는 근거로 적용하였다.

전제

- base learner L
- partial data set, P_1, P_2,…,P_n
- association rules made by $Joint_i$, A_i
- clusters made by L, C_1, C_2,..C_k
- the number i of iterations to be performed by partial data set
- Joint function of present data set ∪ added data set, ◇

알고리즘

(1) Use L, to create clusters C_k in P_i , for $1 < i < n$
(2) Set $i = n$, $j = 0$, $Joint_j = P_i$
(3) Loop for $i \geq 1$
(4) - $Joint_{j+1} = Joint_j \diamond P_{i-1}$
(5) - Finding association rules A_{j+1} of $Joint_{j+1}$ through C_k made by P_i
(6) - Set $i = i - 1$
(7) - Set $j = j + 1$
(8) CreateOutput($Joint\ A_{j+1}$)

그림 2. 순서화된 멀티 소스 데이터 셋의 분석 알고리즘

그림 2는 본 논문에서의 실험 방법을 알고리즘으로 표현한 것으로 3.1의 데이터 셋으로 실험한 순서와 함께 설명한다.

(1) 분산된 데이터 셋에서 데이터를 통합할 때에 정보량에 따라 데이터 셋의 순서를 정한다(P_1, P_2,…,P_n). 실험 데이터에서는 계통 프로파일은 전체 서열과의 비교를 통한 유사도를 측정한 값이다. 따라서 계통 프로파일의 데이터 값이 발현 데이터의 정보량보다 더 많다고 판단하여 계통 데이터를 P_2로 발현 데이터를 P_1으로 정하였다.

(2) 계통 데이터에서 사회 네트워크 분석의 매개 중심성(betweenness centrality) 방법 L을 적용하여 계통 데이터의 각 24개 종(species)에 관련된 유전자 그룹으로 25개 클러스터(C_1, C_2,..C_{25})를 만들었다(24개 요소 벡터와 데이터 내에서 24개 종과 아무 관련이 없는 유전자 그룹을 하나의 클러스터로 만듦).

(3) 각 25개 클러스터에는($Joint_i = P_i$) 발현 데이터의 값들이 포함되어 있고 이 둘의 데이터 셋을 결합한다($Joint_{i+1} = Joint_i \diamond P_{i-1}$). 이 값들을 상향 조절(up-regulation, ↑), 하향 조절(down-regulation, ↓) 및 변화폭이 크지 않은(unchanged, –)

값으로 나누어 클러스터마다 발현 데이터의 연관 규칙 패턴(A_i)을 살펴보았다. 따라서 25개의 종에 대한 클러스터의 79개 샘플들의 발현 패턴을 찾는다(A_{i+1}).
(4) 따라서, 위의 순서에 의하여 종에 의해 관련된 유전자 그룹인 클러스터를 만들고 클러스터 마다 시간대별로 발현패턴의 연관 규칙(association rule)을 살펴본다.

4. 실험 결과

3.1의 데이터를 가지고 3.2의 실험 순서에 따르면 표 1과 같은 결과를 보인다. 표 1은 25개 각 클러스터마다 DNA 마이크로어레이 발현 패턴을 3가지(↑, ↓, -)로 나눈 것으로 가장 많이 매핑되는 값을 대표로 기록한 것이다. 클러스터 1인 경우에 어떠한 종에도 속하지 않는 유전자 클러스터 값으로 시간 순서에 따른 79개 발현 패턴에서 샘플의 시간대별로 대표 값만을 나열한 것이다. alpha와 두 번째 spo 그리고 cold 샘플에서는 시간 순서에 관계없이 변화폭이 크지 않은 경우가 가장 많았고, elu 샘플에서는 시간대별로 변화폭이 크지 않거나 하향 조절인 경우가 가장 많았으며, cdc나 첫 번째 세 번째 spo 샘플에서는 하향 조절인 경우가 많았다. 또한, heat, dtt 및 diau는 상향 조절인 경우가 가장 많았다와 같이 해석할 수 있다. 표 1은 시간대별로의 변화 패턴을 보인 것이 아니라 대표 값만을 나타내었지만, 각 25개 클러스터에 대한 샘플의 시간대별 발현 변화 패턴도 한눈에 볼 수 있다.

본 논문에서는 보통 어떤 실험값을 얻고자 하는 유전자들의 경우에 가정을 먼저 세우고 각 실험실에서는 실험된 결과 값만을 가지고 가정에 맞는지의 여부에 대해서 결론을 내리던 경우보다 여러 실험실의 실험값을 공유하여 각 실험실에서 내린 결론보다 더 많은 지식을 습득할 수 있지 않을까? 라는 문제를 고려하였다. 이것은 우리가 여러 가지 보다 폭 넓은 관점으로 가정을 세울 수 있고 결과를 추론해 낼 수 있을 것이라는 전제하에 실제로 같은 유전자에 대해서 이종 데이터 타입으로 실험되어진 데이터 셋을 가지고 통합된 규칙 정보를 만들었다. 그 결과 이종 데이터 셋을 가지고 즉, 하나의 정보만이 아닌 데이터 셋들의 관련 정보를 관찰할 수 있었다. 이것은 앞으로 이러한 분석 방법이 우리의 실험 정보가 작아서 간과하고 넘어갈 수 있는 것도 결코 무시하지 않아도 될 뿐만 아니라 그것이 정보로 쓰일 수 있으며, 보유 데이터 셋이 작더라도 관련된 여러 가지 통합된 분석 정보를 가지고 다양하게 활용할 수 있는 기반이 될 수 있을 것이다.

5. 결론 및 향후 연구

본 논문에서는 기존의 클러스터링 방법이 아니라 유전자들 사이에는 어떠한 수리적으로 계산할 수 없는 사회적 네트워크에 의한 규칙이 존재할 것이라는 가정 하에 사회 네트워크 분석 방법을 적용하여 클러스터를 형성하였다. 그리고 클러스터들 내부에 새로운 데이터 소스에서의 연관 규칙 방법을 적용하여 이종 데이터의 통합 규칙을 살펴보았다. 본 논문의 실험 데이터는 2가지 종류의 데이터 타입만을 적용하였다. 그리고 유전자 발현 데이터가 믿을만한 실험 조건하에서 실험된 값인지 그리고 소수의 계통 데이터의 관계를 파악하기위해 규칙을 만드는 것이 타당한 것인지에 관한 의문을 풀기 위한 것이 아니다. 본 논문에서는 다양한 종류의 데이터 셋을 통합하여 다양한 정보를 얻고자한 것으로, 앞으로 여러 가지 새로운 타입의 데이터인 약품 처리에 의한 반응 데이터 혹은 또 다른 조건하에서의 유전자 칩 데이터, 패스웨이 데이터 등을 통합하면 더 다양한 결과를 이끌어 낼 수 있다는 것을 보이기 위함이다. 그리고 새로운 실험을 시작할 때에 기존의 여러 가지 다양한 데이터 타입에서 다양한 통합 규칙을 만들어보고 앞으로의 실험 계획을 결정할 때 보다 넓은 안목으로 가정을 세우고 결론을 이끌어 낼 수 있다는 것을 증명할 계획이다.

표 1. 각 클러스터의 발현 패턴

	Phylo \ micro	alpha	elu	cdc	spo	spo	spo	heat	dtt	cold	diau
Cluster1	nothing (24)	-	-↓	↓	↓	-	↓	↑	↑	-	↑
Cluster2	worm (883)	↓↑	↑	-	↓	↓	↑	↓	↑	↓↑	↓
Cluster3	ecoli (30)	↑	-	-	↓	-	↓	↑	-↑	↓↑	↑
Cluster4	tmar (15)	↓	↓	-	↓	↓-↑	↓-	-	↑	-	-
Cluster5	hpy199 (20)	↓	-	↓	-	↑	-	↑	↓	↑	↓
Cluster6	bbur (36)	↓	↓	↑	-	↓	-↑	↓	-	↑	-
Cluster7	aquae (34)	-	-	↑	-	-	↓-	↓	↓↑	↓	↓
Cluster8	ctra (16)	↓	↓	↓	↑	↑	↑	-	↓	↑	↓
Cluster9	synecho (64)	↓	↓	↑	↓-	↓-↑	↓-↑	↑	↑	↑	↓↑
Cluster10	mjan (45)	↑	↓	↑	↑	↑	↓↑	↓	↓↑	↓-	↓
Cluster11	mgen (42)	↓	↓	↑	↑	↑	-	-↑	↓	↑	-
Cluster12	rpxx (65)	↓	-	↑	-	↓	↓-	-	↓	↓	↓↑
Cluster13	cpneu (33)	↓	↓	↓	↑	↓-↑	↑	-	↓	-	↓
Cluster14	hinf (69)	-↑	-	↓-	↓	↓-↑	↓	-	↑	-↑	↑
Cluster15	bsub (125)	↓	↓	↓	↓-↑	↓	↓	↑	↓	-	↓↑
Cluster16	pyro (41)	↑	↓	-	↓	↓	↓-	-	↑	-	↑
Cluster17	mthe (50)	↓↑	↓	↓	-	↑	↑	↓	↓-	↓	↑
Cluster18	pabyssi (101)	↑	↑	↑	↓↑	-	↓	↓	-↑	↓	-↑
Cluster19	aful (102)	-	↑	-↑	-	-	↓	↓	↑	↑	↓↑
Cluster20	areo (119)	↑	↑	↑	↑	↓	↓↑	↓	↑	↓	↓
Cluster21	mtub (132)	↓	↑	↓	↑	-	-↑	↑	↓	↑	↓
Cluster22	dra (107)	↓	-	↓	↓	↑	↓-	-	↓↑	-	↓
Cluster23	mpneu (72)	↓	↓	↓	-	-	-↑	-↑	↓	↑	↓
Cluster24	tpal (51)	↑	↓↑	↓	↓-↑	-↑	-	↑	-	-↑	↓
Cluster25	hpyl (189)	-	↓	-	↓	-	-↑	-↑	↓-	↑	↓

6. 참고 문헌

[1] Paul Pavlidis, Jason Weston, Jinsong Cai and William Stafford Noble, "*Learning Gene Functional Classifications from Multiple Data Types*," Journal of Computational Biology 9(2): pp.401-411, 2002.

[2] Tao Li, Shenghuo Zhu, Qi Li and Mitsunori Ogihara, "*Gene Functional Classification by Semi-supervised Learning from heterogeneous data.*" In Proceedings of The 18th Annual ACM Symposium on Applied Computing, pp.78-82, 2003.

[3] Madimir, Filkov and Steven Skiena, "*Heterogeneous Data Integration with the Clustering formalism,*" International Workshop on Data Integration in the Life Sciences(DILS), pp.110-123, 2004.

[4] Hartemink, A., Gifford, D., Jaakkola, T., and Young, R, "*Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network,*" In Pacific Symposium on Biocomputing(PSB), pp. 437-449, 2002.

[5] Tao Li, Shenghuo Zhu, Qi Li and Mitsunori Ogihara, "*A New Distributed Data Mining Model Based on Similarity,*" In Proceedings of The 18th Annual ACM Symposium on Applied Computing, pp.432-436, 2003.

[6] Albert-Laszio Barabsi, *Link*, Penguin, USA, 2003.