멀티 소스 데이터 분류와 분석을 위한 이머징 패턴의 적용 방법

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

Application of emerging patterns for multi-source data classification 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

상호작용하는 구조들을 하나의 클래스로 표현하는 데이터 마이닝 툴로서 이머징 패턴(EP)이 최근에 제안되었다. 기존의 클러스터링 알고리즘과 패턴 마이닝 알고리즘은 고차원의 유전자 발현 데이터 혹은 같은 변수들(e.g. genes)을 가지고 실험한 멀티 소스 데이터 분석을 다루기에 부적절하고, 실험 결과를 이해하는 데에 어려움이 있다. 그러나 EP는 분류 트리의 형태로 표현 가능하기 때문에, 다양한 형식의 데이터를 분류하는 패턴들을 빠르고 간단하게 구성하여 데이터 분석이 가능하도록 돕는다. 본 논문에서는 멀티 소스 바이오 데이터에서 분류 절차의 작업을 향상시키기기 위하여 EP를 사용하는 간단한 스킴을 제안한다.

1. 서 론

마이크로어레이 실험은 생물학적 형태를 분류하는 데에 혁신적인 기술 발전을 가져왔지만, 병과 관련된 유전자나 유전자 네트워크를 구성하는 것과 같은 복잡한 생물학적 작업 수행과 다양한 정보를 갖는 데이터 셋들을 구분하기 위해서는 좀 더 강력하고 효과적인 분석 전략을 개발하는 것이 필요하다.
유전자 발현 프로파일의 형태는 소수의 환자 n(=observations)와 다수의 유전자 p(=variables)를 가지는 전형적인 데이터 상태로 예를 들 수 있다. 그리고 이것을 소위 유전자 발현 분석에서는 'small n large p paradigm'이라고 한다[1][2].
본 논문에서는 멀티 소스 데이터 분류를 하기 위해서 EP를 이용한 새로운 규칙 기반(rule-based) 방법을 제안한다. 먼저, 각 데이터 셋에서 관측치(observation)의 매개 중심성(betweenness centrality)[9]이 높은 순서대로 유전자 셋을 추출한다. 즉, 데이터 셋마다 관측치의 매개 중심성 값을 계산하여 그 값이 큰 순서대로 변수들을 클러스터링하면 관측치의 개수 n 만큼의 클러스터가 만들어진다. 본 논문에서는 관측치마다 매개 중심성을 계산하여 데이터 셋마다 관측치를 가장 잘 반영하는 변수들을 추출하였다. 이는 전체 변수(e.g. genes)에서 관측치를 가장 잘 설명할 수 있는 유전자 서브셋을 만든 것으로, 서브셋에 클러스터링 된 변수들은 어떠한 서브셋에도 중복되지 않도록 구성된다. 그리고 멀티 소스 데이터를 분류하기 위해서 각 데이터 셋에서 관측치를 기반으로 간결한 EP를 만들어 데이터 셋을 분류하는 데에 적용하였다. 일반적으로 분류 방법을 적용할 때는 마이크로어레이 데이터를 많이 이용하지만, 단지 소수의 접근방법들만이 조사하고자 하는 유전자들 사이에 상호작용을 고려하였다. 하지만 본 논문에서는 모든 변수들을 분석할 때 적용할 수 있는 새로운 방법을 제안하고 또한, 실험 데이터로써 같은 변수를 가지고 실험한 값인 멀티 소스 데이터를 적용하였다. 따라서 비록 멀티 소스데이터는 다른 타입의 데이터가 통합되어진 것이지만, EP는 다른 타입의 데이터 셋을 분류하는 데에 유용하다는 것을 보인다.
논문의 구성은 다음과 같다. 2장에서 멀티 소스 데이터와 생물정보학에서 분류방법의 적용문제 그리고 EP 분석에 대해 리뷰하고, 3장에서는 본 논문에서 제안하는 멀티 소스 데이터 셋에서 EP를 추출하여 적용하는 방법을 보인다. 그리고 3장에서 제안한 방법을 적용한 실험 결과를 4장에서 설명하고, 마지막으로 5장에서는 실험적 결과와 결론 그리고 앞으로의 계획에 대해 설명한다.

2. 관련 연구

본 장에서는 관련 연구로서 멀티 소스 데이터, 생물정보학에서 분류 알고리즘의 적용, 그리고 멀티 소스 데이터 분류를 위한 EP에 대하여 설명한다.

2.1 멀티 소스 데이터

다양한 형식의 데이터들은 연구자들이 예측할 수 없었던 것을 예측할 수 있는 기회를 제공해 주고 대부분은 인터넷에서 자유롭게 이용이 가능하다. 여러 가지 다른 형식의 데이터를 통합하여 분석하는 목적은 유전자 분류, 클러스터링, 유전자 조절 네트워크 등에서 다양하고 독립된 특성들을 이용하여 더 정확하고 풍부한 상호 관계를 밝히고자 하기 위한 것이다. 바이오 데이터의 특성 가운데 하나는 같은 변수들을 가지고 다른 실험 조건하에서 여러 가지 다른 실험을 하여 다양한 형식의 멀티 소스 데이터를 만들 수 있다는 것이다. 이러한 멀티 소스 데이터는 생물학적 관련성에 보다 많은 통찰력을 제공하기 위하여 다른 정보 소스의 상보성을 이용하는 것이 도움이 된다. 그리고 여러 가지 타입의 데이터 정보를 함께 이용하는 것이 더욱 중요한 정보를 이끌어 낼 수 있다는 것을 보인다.

2.2 생물정보학에서 분류 알고리즘의 적용

분류 문제의 목적은 데이터가 어떤 클래스에 속하는지를 예측하기 위한 효과적인 모델을 만드는 것이다. 현재까지 바이오 데이터 분석에서 가장 많이 적용되는 분석 방법은 클러스터링이었다. 하지만 클러스터링 방법은 생물학적 특성(attribute)은 무시하기 때문에 유전자 발현 데이터와 같은 경우에는 관측치의 특징을 살펴보거나 관측치 사이에 어떠한 규칙이 있는가를 살펴보는 것에는 제약점이 있다. 교사 학습 방법 중에서 결정 트리 방법은 클러스터링 방법과는 다르게 명확한 타겟이 있을 경우 즉, 생물학적 형식에 가장 기여하는 유전자들은 무엇인가와 같은 문제를 정의할 수 있다. 현재의 분류 방법을 적용하기 위한 연구로는 질병의 유무를 판단하는 것을 타겟으로 보고 질병 유전자인지 아닌지를 찾는 교사 학습 방법 연구가 진행되고 있다[4].

2.3 이머징 패턴 분석

EP는 [2]에서 처음으로 소개되었는데 성장률(growth rate)의 적절한 분류 기준을 적용하여 각 데이터 셋 D_1과 D_2의 분명한 변화와 차이를 보이는 변수들의 조합으로 서포트(support)를 증가시키는 아이템 셋을 EP로 정의하였다. 즉, EP는 두개의 파티션으로 데이터 셋을 명확하게 구분해주는 패턴을 말하는 것으로, 이러한 패턴들은 하나의 데이터 셋에서 다른 파티션들 사이에 명확한 차별점을 가진다. 일반적으로 연관(association) 분석에서 자주 발생하는 패턴과는 달리 EP는 높은 식별력으로 분류 문제에 적용되어 더욱 유용하다고 증명되어 있다. EP는 또한 데이터 셋에서 특성들을 모아 놓은 것이기 때문에 이해하기 쉽고, 생물정보학의 적용 문제에서도 매우 중요하다로 할 수 있다. 그러나 고차원의 유전자 발현 데이터에서는 만들어진 EP의 수가 매우 많다. 따라서, 본 논문에서 많은 수의 변수들의 특성과 적은 수의 관측치의 특성을 기반으로 분석하는 것이 더 효율적인, EP를 이용하는 분류 방법을 제안한다.

3. 멀티 소스 데이터에 이머징 패턴 적용 방법

본 장에서는 이 논문에서 적용한 실험 데이터와 실험 방법에 대해서 설명한다.

3.1 데이터

본 논문에서는 제안하는 방법을 적용하기 위한 멀티 소스 데이터로 두 가지 형식의 게놈 데이터를 사용하였다. 첫 번째는 DNA 마이크로어레이 실험 데이터로 각 데이터 값은 두 가지 다른 실험 조건하에서 특정 유전자의 발현 정도를 logarithm ratio로 계산하여 나타내었다. 데이터는 2,465개의 yeast 유전자를 가지고 79개 샘플에서의 시간 변화에 따른 유전자의 발현 정도로 구성되어 있다[3]. 그리고 2,465개의 yeast 유전자 각각에 대하여 24개 계통학적 프로파일[6]로 특성화시킨 데이터를 적용하였다. 이 데이터 셋은 관심 유전자가 대응하는 지놈과 얼마나 유사성을 가지는지를 값으로 가진다. 본 논문에 포함된 프로파일은 전체지놈과 BLAST version 2.0[7]에 의한 lowest E-value의 negative logarithm 값이다. 프로파일은 24 전체 지놈을 이용하여 구성하였으며, 학습에 앞서 유전자 발현 데이터와 계통학적 프로파일 벡터는 평균은 0, 편차는 1을 가지도록 맞춰졌다.

3.2 관측치에 기반한 매개 중심 값 적용

바이오 데이터의 특징 중 하나가 변수의 수(rows)에 비하여 비교적 적은 수의 관측치(columns)를 갖는다는 것이다. 그리고 바이오데이터의 특성은 데이터들간에 상호작용이 의존적인 경우가 많기 때문에 어떠한 변수를 하나 제거하면 결과의 차이가 많이 날 수 있다. 따라서 본 논문에서는 각 데이터 셋마다 전체 변수들을 고려하여 데이터 셋 마다의 특징을 EP로 표현하고 그 결과를 멀티 소스 데이터를 분류하는 데에 적용하고자 한다. 또한 EP는 쉽게 이해하고 표현 가능하기 때문에 다른 타입의 데이터 셋들의 특징을 살펴보는 데에도 유용함을 보이고자 한다.

다음은 한 개의 데이터 셋을 가지고 본 논문에서 제안하는 방법으로 EP를 생성하는 방법을 순서대로 설명한다.

(1) 우선 바이오 데이터의 셋은 변수의 수에 비하여 관측치의 수가 적기 때문에 관측치 값을 기준으로 관측치를 가장 잘 반영하는 유전자들을 클러스터링 한다. 이때 사회 네트워크의 매개 중심 방법을 적용하여 관측치에 밀접하게 관련 있는 변수들을 추출한다. 네트워크에서 노드는 사람이나 그룹을 말하는 것으로 가장 활동적인 사람은 많은 다른 사람들과 연결 관계를 갖는다는 것을 의미한다. 매개 중심 방법은 다른 노드들과의 사이에서 'bridge' 역할인 노드를 찾는 방법으로 전체 데이터 네트워크가 이 노드에 의해서 밀접하게 연관되어 있다는 것을 말한다. 실험에서는 관측치마다 매개 중심 값을 계산하여 가장 높은 값을 가지는 관측치를 찾고 영향력을 미치는 유전자들을 추출한다.

(2) 앞의 실험에서 가장 높은 매개 중심 값을 가지는 관측치와 관측치에 가장 밀접한 관련성을 가진다고 판단되는 유전자들을 제외한 나머지 관측치와 유전자들을 가지고 다시 매개 중심 값을 계산한다. 그리고 가장 높은 매개 중심 값을 가지는 관측치를 찾고 밀접한 영향력을 가지는 유전자들을 클러스터링한다.

(3) 이러한 방법으로 관측치와 변수의 관계에 따른 클러스터를 형성하기 위해서 매개 중심 값을 반복 계산하는 것을 관측치의 수만큼 되풀이한다.

(4) 마지막으로 어떠한 관측치에도 속하지 않는 변수들을 또 하나의 클러스터로 구성한다.

3.3 데이터 셋마다 이머징 패턴 생성

3.1절의 실험 데이터를 가지고 관측치들의 매개 중심 값을 계산하고, 가장 높은 매개 중심 값을 가지는 관측치에 대하여 설명력 있는 변수들을(=genes) 추출하여 클러스터를 만들었다. 그 결과 첫 번째 실험에 적용한 yeast의 마이크로어레이 데이터에서는 79개 time points로 구성된 관측치를 10개의 관측치 값으로 축약하여 클러스터를 만들었다(10개의 샘플이 time points 79-element로 구성되어 있는 것으로 본 실험에서는 time point가 아닌 샘플 개수로 관측치를 다루었음). 그리고 두 번째 계통학적 프로파일 실험 데이터에서는 관측치에 해당하는 24개 종의 수만큼 클러스터를 만들었다. 본 논문에서는 마이크로어레이 데이터 셋에서 만들어진 EP와 계통학적 프로파일 데이터에서 만들어진 EP를 다음과 같이 표현한다. 마이크로어레이 데이터 셋에서의 EP 표현은, $\exp(X_1)>a_1 \wedge \exp(X_2)<a_2$ 와 같은 형태로, 계통학적 프로파일은 $\text{phylo}(Y_1)>b_1 \wedge \text{phylo}(Y_2)>b_2$ 와 같은 형태로 나타낸다. 여기서 마이크로어레이 데이터의 $\exp(X_i)$와 계통학적 프로파일 데이터의 $\text{phylo}(Y_j)$에서 X_i와 Y_j는 각각 마이크로어레이 데이터 셋에서의 관측치의 발현 정도를 계통학적 프로파일 데이터 셋에서는 관측치의 서열 유사성을 말한다. 그리고 a_i와 b_i는 각각 발현정도와 서열 유사도의 임계 값을 말한다.

4. 실험 결과

본 논문에서는 매개 중심 값 계산을 위해 R 패키지를 이용하였으며, EP를 만들기 위해서는 Weka의 분류 알고리즘 J4.8을 이용하여[8]을 적용하였다. 그림 1은 차례로 마이크로어레이 데이터 셋과 계통학적 프로파일에서의 결과를 보이는데, 각 데이터 셋에서 10개와 24개의 규칙들이 만들어졌다. 마이크로어레이 데이터의 규칙은 10개의 샘플에서 6개의 샘플만이 분류 규칙을 만드는 데에 적용되어 EP가 만들어 졌고, 계통학적 프로파일 데이터에서는 24개 관측치 모두가 분류 규칙에 적용되어 EP가 만들어졌음을 확인할 수 있다. 결과는 다음과 같이 해석할 수 있다. 마이크로어레이 데이터 셋에서 7번째 라인의 EP는 $\exp(dtt)\geq 1.12 \wedge \exp(cold)\leq 0.585$인데, 이것은 전체 마이크로어레이 데이터 셋에서 dtt 관측치의 유전자 발현 값이 1.12 이상이고 cold 값이 0.585 이하인 변수들은 전체 마이크로어레이 데이터 셋에서 관측치인 dtt를 구분할 수 있다. 또한 이와 같은 EP는 마이크로어레이 데이터 셋의 다른 관측치와 구분할 수 있는 구분자로 볼 수 있겠다. 마이크로어레이 데이터의 EP에서 마지막 라인의 alpha는 위의 모든 규칙에 해당되지 않는 유전자들은 alpha 관측치를 설명할 수 있는 유전자들이라는 것을 말한다.

계통학적 프로파일의 결과 해석도 마찬가지로 첫 번째 라인의 EP가 $\text{phylo}(cpneu)>1.1 \wedge \text{phylo}(tpal)\leq 1.09$인데 이것은 계통학적 프로파일 데이터 셋에서 cpneu 관측치를 구분할 수 있는

마이크로어레이 데이터의 이머징 패턴

```
exp(spo)≥2.24 ∧ exp(spom)≥2.19 ∧ exp(cold)≤0.526 ∧ exp(spom)≥2.69: spom
exp(cold)≥0.712 ∧ exp(heat)≤0.662 ∧ exp(spo)≥2.81 ∧ exp(heat)≤0.236: cold
exp(spo)≥3.23 ∧ exp(spo5)≤-1.35 ∧ exp(elu)≤-0.184: spo5
exp(spom)≤-2.21 ∧ exp(elu)≥0.547 ∧ exp(elu)≤1.08 ∧ exp(heat)≤-0.484 ∧ exp(diau)≤-1.12: spo5
exp(spo)≥2.42 ∧ exp(cold)≤0.109 ∧ exp(cdc)≥0.818: spo5
exp(spo5)≥1.18 ∧ exp(spo5)≥3.51: spo5
exp(dtt)≥1.12 ∧ exp(cold)≤0.585: dtt
exp(elu)≥1.05 ∧ exp(spom)≤-2.21 ∧ exp(spo)≤-1.77: elu
exp(alpha)≤0.745 ∧ exp(elu)≥0.555 ∧ exp(elu)≥1.24 ∧ exp(spo5)≤0.174: elu
:alpha

Number of Rules : 10
```

계통학적 프로파일의 이머징 패턴

```
phylo(cpneu)>1.1 ∧ phylo(tpal)≤1.09: cpneu
phylo(tmar)>1.1 ∧ phylo(tpal)≤1.07 ∧ phylo(ctra)≤1.14 ∧ phylo(mpneu)≤1.11: tmar
phylo(aero)>1.1 ∧ phylo(ctra)≤1.06 ∧ phylo(mpneu≤0.87 ∧ phylo(tpal)≤1.09: aero
phylo(aful)>1.1 ∧ phylo(tpal)≤0.999 ∧ phylo(ctra)≤1.24 ∧ phylo(mpneu)≤1.75: aful
phylo(mpneu)>1.07 ∧ phylo(ctra)≤1.18 ∧ phylo(tpal)≤0.94: mpneu
phylo(mthe)>1.1 ∧ phylo(tpa)≤ 0.999: mthe
phylo(tpal)≤1.09 ∧ phylo(rpxx)>1.1 ∧ phylo(ctra)≤1.12: rpxx
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(mtub)>1.1: mtub
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(hpyl99)>1.1: hpyl99
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(bbur)≤1.1 ∧ phylo(dra)≤1.1 ∧ phylo(synecho)>1.1: synecho
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgen)≤1.1 ∧ phylo(bbur)>1.1: bbur
phylo(tpal)≤1.09 ∧ phylo(ctra)≤1.1 ∧ phylo(mgenv)>1.1: mgen
phylo(tpal)>1.09: tpal
phylo(ctra)≤1.1 ∧ phylo(dra)>1.1: drai
phylo(ctra)>1.1: ctrau
phylo(aquae)>1.09: aquae
phylo(bsub)>1.1 ∧ phylo(hpyl)≤1.05: bsub
phylo(hpyl)≤1.11 ∧ phylo(ecoli)≤1.08 ∧ phylo(pabyss1)≤1.03 ∧ phylo(mjan)≤1.08 ∧ phylo(pyro)≤1.1 ∧ phylo(hinf)≤1.09: vora
phylo(pabyss1)>1.1: pabyss1
phylo(hpyl)≤1.11 ∧ phylo(mjan)≤1.09 ∧ phylo(pyro)≤0.981 ∧ phylo(hinf)>1.1: hinf
phylo(hpyl)≤1.06 ∧ phylo(mjan)>1.09: mjan
phylo(hpyl)≤0.996 ∧ phylo(pyro)≤0.981: ecoli
phylo(hpyl)>0.996: hpy
: pyro

Number of Rules  :      24
```

그림 1. 마이크로어레이 데이터와 계통학적 프로파일 데이터의 이머징 패턴

구분자가 된다. 그리고 이 구분자에 의하여 만들어지는 클러스터에는 cpneu를 가장 잘 설명할 수 있는 유전자들이 서열의 유사도 값에 따라서 클러스터링 되었다는 것으로 해석 할 수 있다. 이렇게 생성된 EP가 얼마나 두 가지 데이터 셋을 구분하는 데에 정확한 구분자 역할을 수행하는지에 대한 타당성을 검증한 결과 마이크로어레이 데이터에서는 정확도가 85.76%, 계통학적 프로파일 데이터는 97.79%로 정확하게 구분해준다는 것을 발견할 수 있었다. 마이크로어레이 데이터 셋의 정확도가 더 낮은 이유는 실험 할 때부터 79개의 time points를 10개의 관측치로 줄였기 때문이라고 설명할 수 있다.

5. 결론 및 향후 연구

바이오 데이터는 같은 변수들을 가지고 여러 가지 다른 타입의 멀티 소스 데이터 셋을 만들 수 있는데 이러한 데이터 셋들을 분류 할 수 있는 구분자와 데이터 셋의 특징을 쉽게 이해할 수 있는 방법도 필요하게 되었다. 따라서 본 논문에서는 바이오 데이터의 샘플수 보다 유전자의 수가 더 많은 특성을 고려하는 샘플에 기반한 특징적 유전자들을 추출하는 방법을 제안하였다. 이 방법은 데이터 셋 내에서 유전자들의 상호 관련성까지 고려하는 EP를 만드는 것이고 그 결과를 멀티 소스 데이터를 분류하는 데에 적용하였다. 앞으로의 연구 방향은 본 논문에서 제안하는 실험 방법에 몇 가지 고려 사항을 첨가하고자 한다. 먼저, 바이오 데이터에서 멀티 소스 데이터를 분류할 때에 더 나은 분류자를 찾는 문제를 고려할 수 있다. 또한 본 논문에서는 멀티 소스 데이터 분류를 하기 위해서 적용한 데이터 셋이 단지 두 가지 생물학적 데이터 형식이었다. 따라서 멀티 소스 데이터 분류 문제에 EP를 발견하고 제안하는 방법을 확장하기 위해서 더욱 다양한 생물학적 데이터 형식을 적용하는 방향을 연구하고자 한다. 또한 다른 데이터 소스의 이론적, 실험적 증명을 하는 것 역시 중요한 과제이다.

6. 참고 문헌

[1] A. L. Boulesteix, G. Tutz and K. Strimmer, A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*, 19: 2465-2472, 2003.

[2] G. Dong and J. Li, Efficient mining of emerging patterns: discovering trends and differences. *Proceedings of the SIGKDD(5th ACM International Conference on Knowledge Discovery and Data Mining)*, 5: 43-52, 1999.

[3] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8): 4285-4288, 1999.

[4] M. West, J. R. Nevins, R. Spang and H. Zuzan, Bayesian regression analysis in the 'large *p*, small *n*' paradigm with application in DNA microarray studies. *Technical Report* 15, Institute of Statistics and Decision Sciences, Duke University, USA, 2000.

[5] P. Pavlidis, J. Weston, J. Cai and W. N. Grundy, Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9: 401-411, 2002.

[6] S. F. Altschul, T. L. Madden, A. A. Schaffer, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25: 3389-3402, 1997.

[7] T. H. Yu. Larray, C. Fu-lai and C. F. Stephen, Using Emerging Pattern Based Projected Clustering and Gene Exptression Data for Cancer Detection. *Proceedings of the Asia-Pacific Bioinformatics Conference*, 29: 75-87, 2004.

[8] http://www.cs.waikato.ac.nz/~ml/weka/

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