바이오 데이터 분류화를 위한 BNP 내장 생태계 모방 알고리즘에 대한 연구

최옥주, 맹보연, 이민수
이화여자대학교 컴퓨터학과
pensica@ewhaian.net, mngby@ewhaian.net, mlee@ewha.ac.kr

A Study on Bio-inspired algorithm included BNP for Classification of Bio data

Ok-Ju Choi, Boyeon Meang, Minsoo Lee
Dept of Computer Science and Engineering, Ewha Womans University

요 약

다방면적인 과학기술의 발달은 우리에게 대량의 데이터와 또한 새로운 영역으로의 접근 가능성을 열어주었다. 유전자 정보와 같은 대량의 정보를 다루는 시대가 열리면서 바이오 데이터를 분석하여 새로운 연관성과 정보를 찾아내는 바이오인포매틱스가 고부가가치 창출을 위한 학문으로 특히 부각되고 있다. 본 논문에서는 이러한 연구의 일환으로 보다 효율적인 바이오 데이터 분석을 위해 BNP에 내장된 생태계 모방 알고리즘의 특성을 연구하고, 이를 분류화에 접목시킨 방법에 대해 논하고자 한다.

1. 서론[1)]

저장 공간, 통신 단말기와 같은 하드웨어와 네트워크기술의 발달은 매우 거대한 데이터 환경을 사람들에게 제공했다. 기존의 데이터들끼리의 연관성에 대한 질문도 가능해졌고, 유전자 정보와 같은 새로운 영역에서의 대량 데이터를 접할 수도 있게 되었다. 그러나 대량의 데이터는 오히려 사용자가 그 안의 정보를 파악하기 힘들어 의미성을 잃어버리고, 이처럼 정보를 해석할 수 없는 데이터는 그 양이 얼마든지, 그 안의 정보가 얼마나 중요한지의 여부에 관계없이 사용자로부터 버려질 수 밖에 없다. 또한 설혹 분석을 통한 의미 파악이 가능하다고 해도, 그 속도가 현저하게 느린 탓에 다양한 분석 옵션을 적용하기 힘들다. 때문에 이러한 대량의 데이터 환경에서 신속하게 의미를 지닌 정보를 찾아내 그 관계성을 파악하고 정확하게 분석하는 작업의 필요성이 매우 커졌다. 이러한 환경 속에서 가장 크게 부각된 영역 중 하나는 바이오 데이터를 다루는 바이오인포매틱스 영역이다.

바이오인포매틱스는 원시 바이오 데이터로부터 유전자나 단백질과 같은 바이오 객체들의 각각의 기능과 유기적으로 관련이 되는 총체적인 기능을 밝혀내는 것을 목적으로 하는데,[1] 이를 통해 가공된 바이오 정보는 신약개발에서의 시행착오와 소모비용, 소모시간 등을 줄이고, 의료진단에서도 유전자 레벨의 새로운 진단법을 출현하게 하는 등, 획기적인 진단 및 치료 수준을 높일 수 있을 거라 생각된다. 이를 위해 대용량 바이오 정보간의 관련성을 파악하기 위한 알고리즘과 다양한 바이오 정보에 대한 분석과 해석 그리고 다양한 유형의 바이오 정보들에 대한 분석 작업을 도와주는 총체적인 툴 개발이 요구된다.

이러한 연구의 일환으로, 보다 효율적인 바이오 데이터 분석을 위한 BNP 내장 생태계 모방 알고리즘의 특성을 연구하고, 이를 데이터마이닝의 분류화에 접목시킨 방법을 제시한다. 그리고 마지막으로 이러한 방법을 실제로 응용하여 구현한 DNA 칩 분석 시스템에서의 실험을 통해 그 성능을 논하고자 한다.

2. Bio Network Processor

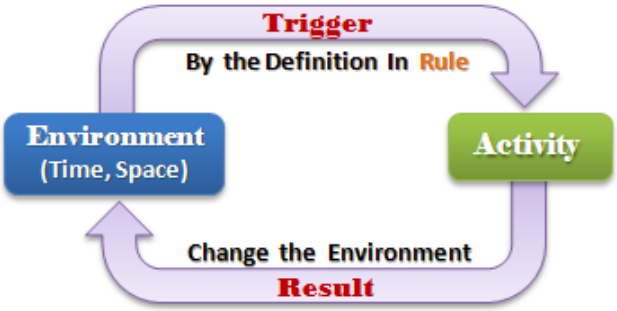
BNP(Bio Network Processor)는 생존형 하드웨어를 내장하여 실제 생태계의 특성인 생존성 및 적응성을 제공해주는데, 이를 위해서 Particle Swarm Optimization(이하 PSO), ANT 알고리즘과 같은 생태계 모방 알고리즘을 내장하여 사용한다.

2.1 생태계 모방 알고리즘

생태계의 환경에 대한 생존성 및 적응성은 '스티그머지(Stigmergy)시스템'으로 설명할 수 있다. 스티그머지란 1950년대 Grasse가 흰 개미가 그들의 집을 재건하는 활동에 대한 연구에서 최초로 제시한 개념으로, 생태계에서 자연적으로 발생하는 조정 작용을 설명한 것이다.

1)이 논문은 정부(교육과학기술부)의 재원으로 한국과학재단 지원을 받아 수행된 연구임(No. R01-2008-000-20029-0).

생태계내 사회는 일정한 규칙을 가지고, 규칙에서 정의한 바에 따라 현재의 시·공간적 환경에 부합되는 조정 활동을 수행한다. 수행의 결과에 따라 환경이 변하면 계속해서 또 다른 활동이 유도된다.[2] 이는 아래의 그림과 같다.



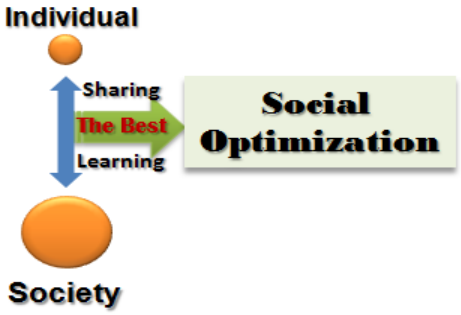
(그림 1) 조정 작용에 대한 메카니즘

이 때, 조정에 대한 규칙과 그에 따른 활동은 흰 일개미(Worker) 한 마리같은, 사회를 이루는 하위 개체에 전적으로 의존된 게 아니라 흰 개미의 사회공동체(the Nest)와 같은 상위 개체에서도 영향을 받는데, 이러한 특성은 실제 생태계에서 단순히 공동으로 양육하는 습성을 가진 집단에서부터 개미와 같은 고등 레벨의 사회에까지 나타나며, 실제로 이들의 규칙과 활동은 생태계에서의 생존을 최우선하여 주어진 환경에의 적응을 목표로 정의된다.

생태계 모방 알고리즘이란 이러한 생태계의 생존·적응적 특성에 따른 실제 자연 생태계 동·식물의 행동패턴 및 습성을 알고리즘으로 구현한 것으로, 그 안에 나타나는 객체간의 상호관계와 영향력을 잘 반영하고, 그에 따른 최적의 값을 구할 수 있다. 이러한 생태계 모방 알고리즘으로는 개미가 먹이를 찾을 때 최단 경로로 가는 방법을 모방한 개미(ANT) 알고리즘, 유전자가 진화해가는 과정을 모방한 유전자(Generic) 알고리즘, 뇌의 뉴런들의 행동을 모방한 뉴럴 네트워크 그리고 새·벌과 같이 하나의 무리(Swarm)를 이루는 동물의 습성을 모방한 PSO 알고리즘 등을 예로 들 수 있다.

2.1.1 Particle Swarm Optimization

PSO 알고리즘은 벌, 철새 등과 같은 무리의 집단적 지성(Swarm Intelligence)에 기반한 알고리즘으로, 집단을 이루는 하나의 구성원과 그 집단 간의 지식체계를 분석하여, 개인 레벨의 입자(Particle)와 사회 레벨의 무리(Global)간의 직·간접적 경험 정보의 공유, 공유를 통한 학습과 비교 판단 그리고 이를 통한 최적값 도출과 같은 사회적 행동을 진화적 계산(Evolutionary Computation) 방법을 사용하여 잘 보여주고 있다. 이를 간단히 표현하면 아래와 같다.



(그림 2) PSO에서 레벨에 따른 두 개체의 관계

2.3 분류화 알고리즘의 구현

분류화 알고리즘은 그 목적과 어떤 데이터를 다루느냐에 따라서 성능이 크게 차이가 난다. 따라서 바이오 데이터의 효율적인 분석을 위해서 가장 적합한 알고리즘을 찾아야 하는데, 실제 생태계 상의 바이오 데이터에 대해서, 생태계 특성에 기반한 생태계 모방 알고리즘이 최적이라 말할 수 있다.

여기에서 분류화에 적용한 생태계 모방 알고리즘인 PSO 알고리즘은 아래와 같다.

```
For each particle
    initialize particle
END

Do
    For each particle
    Calculate fitness value
    If the fitness value is better than the best fitness value [pBest] in history
            set current value as the new pBest
End
    Choose the particle with the best fitness value of all the particles as the gBest
            For each particle
            Calculate particle velocity according equation
            Update particle position according equation
            END
While maximum iteration or minimum error criteria is not attained
```

(그림 3) PSO 알고리즘의 Pseudo Code

PSO 알고리즘에서 입자와 무리는 pbest(Particle Best - 입자 최적값), gbest(Gloabl Best - 무리 최적값)을 가지고 최적화 과정을 수행하는데, 모든 과정을 수행한 뒤 찾아낸 최고의 최적값(Final Global Best)은, 실제 바이오 데이터에서 도출해낸 분류화를 위한 최적의 규칙이 된다.

2.2.1 입자매핑

BNP에서 위 알고리즘에 실제로 적용하는 데이터는 DNA chip에 담긴 각 실험을 분류한 실험의 ID와 각 유전자들의 실험에 대한 발현값이므로, 이 데이터에 대한 입자 매핑(mapping) 및 구현은 아래와 같다.

Gene Number	gene_1	gene_2	...	gene_n	class
Value	expression_ value_1	expression_ value_2	...	expression_ value_n	1 or -1

(그림 4) 입자 매핑

다음과 같이 입자는 하나의 실험에 대해서 유전자의 수만큼 차원을 가지는 벡터로 표현된다. 따라서 입자는 입력받는 데이터의 총 실험수만큼 존재하며, 각 입자의 차원은 총 유전자 수와 같다. 이 때, 각 입자는 아래의 식을 기본으로 하여 자신의 발현값을 변화시킨다.

$$V_{i(t+1)} = \alpha V_{i(t)} + c_1 \bullet \Phi_1(pbest_i(t) - X_{i(t)}) + c_2 \bullet \Phi_2(gbest_i(t) - X_{i(t)})$$

각 입자는 다차원 공간을 대상을 돌아다니고 있는데 $V_{i(t)}$는 이러한 입자i 의 t 번째 반복에서의 속도를 나타내는 벡터다. 입자 I는 또한 위치를 나타내는 벡터 $X_{i(t)}$를 가지고 있는데 이 위치는 아래의 식에 따라 변화된다.

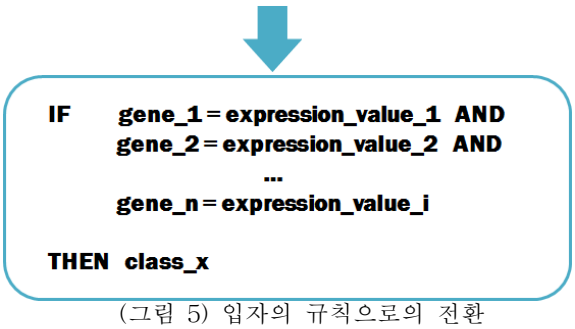
$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)}$$

위의 식에서 상수 c_1, c_2를 변경하여 입자와 무리 중 어느 쪽의 영향력을 크게 할런지 조절할 수 있고, 또한 상수 α를 변경하여 전체 공간의 범위를 조절할 수 있다. 이러한 상수를 통한 환경 설정은 해당 데이터와 시스템 환경에 적합하도록 이루어져야하므로, 각 상수값 변경이 처리 환경에 미치는 정도를 다양한 환경에서 실험하여 해당 환경에서의 최적값을 찾는 작업을 수행할 수도 있다.[3],[4]

2.2.2 규칙 표현

각 입자들은 각기 하나의 규칙을 표현하고 있다. 입자가 가지고 있는 유전자 ID와 발현값은 곧 규칙의 조건이 되는데, 각 입자의 규칙으로의 전환은 아래와 같다.

Gene Number	gene_1	gene_2	...	gene_n	class
Value	expression_ value_1	expression_ value_2	...	expression_ value_n	1 or -1



(그림 5) 입자의 규칙으로의 전환

각 규칙의 정확성은 실제 이 규칙에 부합되는 테스트 데이터의 퍼센트로 계산한다. 즉, 규칙에 따라 테스트 데이터를 분류했을 때, 실제 테스트 데이터의 분류값과의 일치도를 비교하는 것이다. 이러한 비교는 아래의 Confusion Matrix를 따른다.(단, 실제 실험에서는 Class는 1과 -1로 분류된다.)[5]

<표 1> Confusion Matrix

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	f11	f10
	Class=0	f01	f00

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

3. 실험 및 결과

다음은 분류화 작업을 실험한 환경이다.

<표 2> 실험 및 구현 환경

	System	Tool	Language
Implementation Environment	-Win XP(sp2) -Pentirum 4 3.20GHz -2.00G RAM	Visual Studio 6	C++

입력한 바이오 데이터는 유전자 100개와 실험 24개에 대한 발현값과 정확도 평가를 위한 각 실험의 실제 분류값으로 다음에 나타난 바와 같다.

<표 3> 입력 바이오 데이터 파일

Gene ID	Experiment ID	Expression Value
297784	N000287	5.11359232185826
297784	N000288	-1.3389279170472
...	...	...
297784	N000310	4.70523273962433
297912	N000287	-0.30576094332653
297912	N000288	2.70243282124646
...	...	...
297912	N000310	-1.06540479678734
297990	N000287	1.82213569753198
297990	N000288	2.70243282124646
...	...	...
297990	N000310	-1.71125530371349

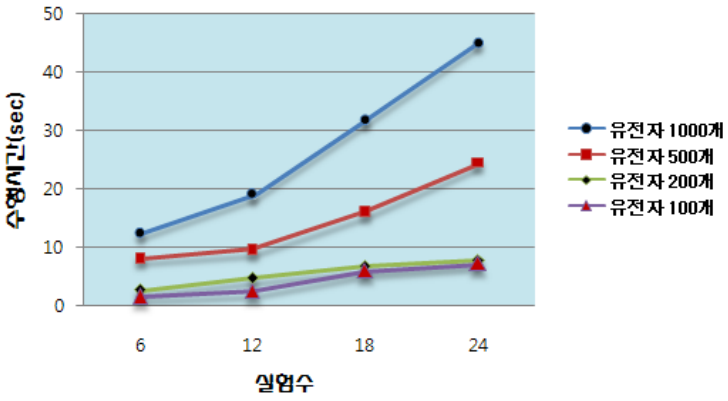
<표 4> 정확도 측정을 위한 실제 실험의 클래스

Experiment ID	Real Class
N000287	-1
N000288	1
N000289	1
:	:
:	:
N000305	1
N000306	-1
N000307	1
N000308	-1
N000309	1
N000310	1

이러한 환경에서 각각 총 유전자 수와 적용 실험 수에 변화를 주며 실험한 결과는 아래 표와 같다. 총 유전자수와 각 유전자에 대한 해당 실험 수 변화에 따른 정확도를 <표 5>에서 나타내며, (그림 6)은 이러한 실험 조건 변화에 따른 수행 속도를 그래프로 보여 주고 있다.

<표 5> 실험 조건에 따른 PSO 분류화 정확도

실험 수 / 유전자 수	6	12	18	24
25개	100	100	100	100
50개	100	100	100	100
75개	100	99.50	99.50	95.83
100개	100	98.18	98..33	98.16



(그림 6) 실험 조건에 따른 PSO 분류화 수행 속도

4. 결론

총 1000개의 유전자에 최대 24가지 실험을 적용한 조건에서 약 45초의 짧은 시간이 소요되지만, 현재 실험에서 적용한 데이터의 양은 (유전자 1000개×실험 24가지)=24000 정도로, 이를 대량의 데이터라고 정의하기엔 부족하다. 때문에 실험의 소요 시간이 적게 걸림에 의의를 두기엔 아직 이르다. 실험 결과에 보다 무게를 실어주기 위해선 보다 거대한 바이오 데이터에 대한 실험이 행해지고, 그에 대한 세밀한 분석 작업이 뒤따라야 한다. 또한 본 실험에서 적용한 데이터는 실제 암 환자의 데이터인데, 실제 모든 바이오 데이터가 이와 동일한 형태는 아니며, 동일한 형태의 데이터인 경우에도 그 자료 구조가 본 실험과 다를 수 있다. 때문에 진정한 바이오 데이터 분석을 위해서는 다양한 형태와 포맷의 바이오 데이터를 지원할 수 있으며, 이러한 기능에 신뢰성이 뒷받침 되어야한다. 또한 나아가 의료진단을 위한 시스템을 목표로 하기 위해선 위 실험과 같이 암 이라는 하나의 질병에 한정되지 않고 다양한 질병에 대해서도 동일한 성능을 보장하기 위한 노력이 필요하다.

참고문헌

[1] 박선희 "바이오 지식 생성을 위한 IT 기반 기술", 전자공학회지 2003年 10月 제 30권 제 10호, p 43-p 50

[2] Eric Bonabeau, Artificial Life, Spring 1999, Vol.5, No.2, pages 95-96

[3] Tiago Sousa. "A Particle Swarm based Data Mining Algorithm for classification tasks." Parallel Computing. 2004, Vol.30, issue5-6, pages 767-783,

[4] 이윤경, 윤혜정, 이민수, 윤경오, 최혜연, 김대현, 이근일, 김대영, "Particle Swarm Optimization 알고리즘을 이용한 바이오칩 데이터의 군집화 및 분류화 기법", 한국정보과학회 추계학술대회 2007, Vol.34, No.2, pages 151-154

[5] Tan, Pang-ning/ Steinbach, Michael/ Kumar, Vipin. Data Mining(introduction To). Pearson Addison Wesley, 2006