산술 연산자 기반 유전자 프로그래밍을 이용한 효과적인 암 분류

홍진혁[0] 조성배
연세대학교 컴퓨터과학과
hjinh@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Effective Cancer Classification Using Genetic Programming based on Arithmetic Operators

Jin-Hyuk Hong[0] Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

최근 생물정보 기술이 암 진단의 새로운 방법으로 관심을 모으고 있다. 다양한 기계학습 기법을 적용하여 우수한 결과를 얻고 있지만, 의학 분야에서는 정확률이 높은 분류기의 획득과 동시에 획득된 분류규칙을 분석하고 이해할 수 있어야 한다. 생물정보 기술에서 많이 사용되는 유전발현 데이터는 데이터 내에 수천 내지 수만의 변수가 존재하여 직접 이들 사이의 복잡한 관계를 표현하고 이해하는 것은 매우 어렵다. 본 논문에서는 이러한 어려움을 극복하기 위해 유전발현 데이터에서 분류에 유용한 특징들을 추출하고 유전자 프로그래밍으로 추출된 특징들을 이용한 암 분류규칙을 생성한다. 림프종 유전발현 데이터에 대하여 실험해본 결과, 90% 수준의 인식 성능을 보였고, 또한 모든 샘플을 완벽하게 분류하는 산술 분류규칙을 발견하였다.

1. 서론

암에 대한 정확한 판단과 분류는 의학 분야에 있어서 매우 중요한 문제인 동시에 매우 어려운 문제이다[1,2]. 정확한 암의 분류는 그에 대한 적절한 치료법과 약품 사용을 가능하게 하여 질병을 치료하고 환자의 생명을 구하는 중요한 일이다. 수세기에 걸쳐 다양한 암 분류 기법이 개발되었지만 대부분 전통적인 형태적 징후 분석에 기반하고 있다. 이들은 진료기반의 방법이어서 사람의 실수나 잘못된 해석 등이 발생할 수 있으며, 다른 종류의 암임에도 불구하고 유사한 징후가 나타나는 경우가 있기 때문에 많은 오분류를 초래하기도 한다. 이러한 한계를 극복하기 위해서 최근에는 사람의 유전자 정보를 이용한 분류기법이 연구되고 있으며 우수한 결과가 보고되고 있다[1,2,3].

사람의 유전자 정보는 최근 주목받는 DNA microarray 기술로부터 수집되며, 이들 유전발현 정보는 생명체에 관한 대량의 유전정보를 포함한다[2]. 많은 경우 유전발현 정보는 다른 종류의 암을 분류하는 데 유용한 정보를 제공한다. 하지만 유전발현 정보의 원시형태는 단순한 숫자들의 나열이기 때문에, 직접적으로 의미를 해석하거나 암을 분류하는 규칙을 발견하기가 매우 힘들다. 따라서 이것을 효과적으로 분석하기 위해 수년전부터 많은 방법이 연구되고 있다[2,3].

다양한 인공지능 기술이 암을 분류하기 위해 적용되어 우수한 분류 성능을 보이고 있다. 하지만 이들 대부분은 사람이 직접 해석하기 매우 어렵고, 많은 변수를 고려해야 하는 문제에서는 우수한 성능을 얻기 힘들다. 유전발현 정보를 이용한 암 분류는 그 특징의 수가 수천 개에 이르기 때문에 쉽게 우수한 분류기를 생성하기 어려우며, 또 사람이 분석 가능한 분류 규칙이 발견되지 않으면 신뢰하기 어렵다[4]. 따라서 본 논문에서는 유전자 프로그래밍을 이용하여 고차원의 유전 발현 정보로부터 우수한 성능을 획득하고 사람이 이해할 수 있는 분류 규칙을 생성하는 분류 시스템을 제안한다.

2. 배경

2.1 DNA Microarray

생물체는 기본적으로 수천 개의 유전자와 RNA 및 단백질이 복잡하게 결합되어 다양한 기능을 한다. 전통적인 분자생물학은 단일 유전자를 기반으로 분석되었기 때문에 매우 제한적이었다. 최근 개발된 DNA microarray 기술은 기존 기술의 한계를 극복하고 초미세 단위로 유전 정보를 획득하고 하나의 칩상에서 전체 염색체의 발현양상을 관찰하도록 한다. 따라서 보다 복잡한 생물체의 현상을 관찰하고 분석할 수 있게 되었다[1,2,3]. DNA microarray는 용액이 투과되지 않는 딱딱한 지지체 위에 고밀도 cDNA를 고정시켜 수천 개 이상의 DNA나 단백질을 일정간격으로 배열하여 붙이고 분석대상 물질과 결합시켜 그 양상을 분석하는 칩이다. 배열 상의 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 각각 다른 형광물질을 동일한 양으로 합성시킨다. 이것을 레이저 형광 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 얻게 되는데, Cy5/Cy3의 비율에 밑이 2인 로그를 취한 값을 그 셀의 발현정보 값으로 얻는다.

$$gene\ expression = log_2 \frac{Int(Cy5)}{Int(Cy3)}$$

2.2 지식발견

지식발견(Knowledge discovery)은 데이터로부터 자동적으로 지식을 추출하는 작업을 말한다. 추출되는 지식은 정확하고, 사용자가 이해할 수 있으며, 흥미로운 것이어야 한다[4]. 지식발견이 적용되는 대표적인 문제로는 분류, 의존성 모델링, 군집화와 연관성 규칙 발견 등이 있으며, 유전자 알고리즘, 유전자 프로그래밍, 결정 트리 등의 방법이 많이 사용된다[5].

의료분야에서는 다량의 데이터로부터 유용한 지식을 발견하는 기술의 필요성이 증가하고 있다. 방대한 양의 데이터는 의학 전문가가 수작업으로 분석하기에는 거의 불가능하며, 아직 알려지지 않은 유용한 관계는 분석에 의해 쉽게 발견되지 않는다. 이런 한계를 극복하기 위해 데이터로부터 유용한 정보를 분석하는 데이터마이닝이라 불리는 지식발견 기술이 의료정보 분석에 사용되고 있다[4,5].

신경망 등의 기계학습 기법은 학습된 분류기로부터 사람이 이해할만한 분류규칙을 추출하거나 해석하기가 매우 어려우며, 특히 의학 분야에서는 학습된 분류기가 매우 높은 정확률을 가진다하더라도 쉽게 신뢰하지 못한다. 기계학습 기법으로 얻어진 규칙이 전문가에게 해석이 가능하고 그 의미가 유효하다고 판명이 되어야 한다. 따라서 결정트리나 진화연산을 이용한 규칙생성 등의 방법이 보다 적합하다[6].

2.3 유전자 프로그래밍

유전자 프로그래밍은 사용자가 명시적으로 프로그래밍을 하

는 것이 아니라 컴퓨터로 하여금 주어진 문제를 해결하는 프로그램을 자동적으로 짜도록 하기 위해 고안된 기술이다. 프로그램을 함수와 변수로 짜여진 일종의 구조체로 간주하고 미리 정의된 문법에 어긋나지 않도록 이들을 구성한다. 일반적으로 root가 하나인 트리의 형태로 프로그램을 구성하며, 이것이 유전자 프로그래밍의 개체 표현형이 된다[7]. 그림 1은 대표적인 유전자 프로그래밍의 표현형을 보여준다.

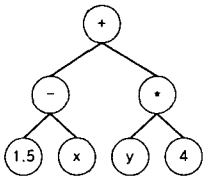


그림 1. 유전자 프로그래밍의 표현형

유전자 프로그래밍은 전통적인 유전자 알고리즘의 확장으로, 집단의 개체를 프로그램으로 정의하였다. 근본적인 동작과 특성은 유전자 알고리즘과 유사하지만, 개체의 표현형이 다르기 때문에 몇몇 차이점이 있다. 유전자 프로그래밍의 해 영역은 함수와 변수의 조합으로 발생할 수 있는 모든 가능한 프로그램이기 때문에 매우 광범위하다. 함수는 산술연산, 논리연산 및 사용자정의연산 등 매우 다양하며, 이들 중 문제에 따라 적절히 선택하여 사용한다. 최근에는 최적화문제나 어셈블리 코드의 진화, 진화하드웨어, 캐릭터행동진화 등의 문제에 많이 도입되고 있다[7].

3. 분류규칙 발견 시스템

본 논문에서는 그림 2에서와 같이 고차원의 특징을 가지는 유전자 발현 데이터에서 분류에 유용한 특징들을 특징선택과정에서 추출한 후, 유전자 프로그래밍으로 선택된 특징들만을 사용하는 분류규칙을 생성한다.

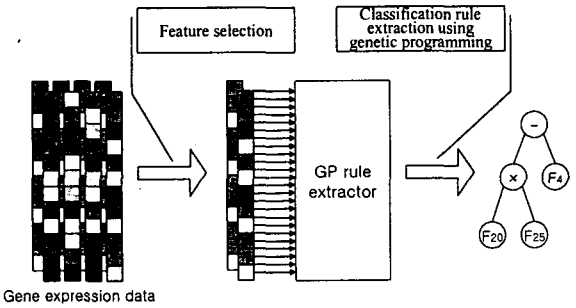


그림 2. 유전자 프로그래밍을 이용한 분류규칙 추출 시스템

3.1 특징추출

유전발현 정보의 유전자들이 모두 특정 질병과 연관되어 있는 것은 아니다. 따라서 특정 질병과 관련된 유전자를 선별하는 작업이 필요하며, 이를 특징선택 혹은 유전자선택이라 한다[3,8]. 특징선택은 학습속도를 향상시키고, 잡음을 줄이는 효과가 있다. 특징선택은 특징의 중요성을 측정하는 기준으로 순위를 매겨 선택하는 순위-기반의 방법과 분류기와 연계된 학습데이터 자체의 특성을 이용한 방법으로 구분되며, 본 논문에서는 유클리드 거리, 정보 이득, 통합 상관 특징선택 방법[8]을 이용하여 성능을 비교하였다.

3.2 분류규칙 인코딩

기존에는 보통 유전자 프로그래밍을 이용하여 IF-THEN 규칙을 발견하였으며, 아래와 같이 AND, OR 등의 논리연산과 <, >, = 등의 크기를 비교하는 연산자가 많이 사용되었다[4,6].

$$Rule1: IF((A1 < 0.5)\ OR\ (A3 > 0.3))\quad THEN\quad class1$$
$$Rule2: IF((A2 = 0.7)\ AND\ (A1 > 0.7))\ THEN\quad class2$$

이러한 규칙은 비교적 해석이 쉬우나, 높은 정확률을 가지기 어려우며 데이터의 변수들 사이의 복잡한 연관성을 표현하는데 한계가 있다. 본 논문에서는 보다 높은 정확률을 가지는 분류규칙을 발견하기 위해 단순한 논리연산이 아닌 산술연산을 이용하여 분류규칙을 구성한다. 특징추출 단계에서 뽑힌 30개의 특징값과 기본적인 산술연산자(+, -, *, /)를 이용하여 트리를 만들어 개체의 표현형으로 이용하였다. 표 1은 개체의 평가에 사용된 각 산술연산자의 유전자에 대한 의미를 표현한 것이다. 분류규칙은 아래와 같이 적용하였다. 그림 3에서처럼, *eval*() 함수는 한 개체의 부류에 대한 근사값을 계산하는 것으로 그 값이 양수일 때 class1, 음수인 경우에는 class2로 분류한다.

$$IF\ \ eval(Individual_i) >= 0\ \ THEN\ \ class1\ \ ELSE\ \ class2$$

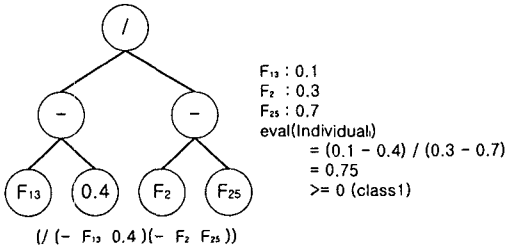


그림 3. 유전자 프로그래밍 표현형 및 분류규칙

표 1. 유전자에 대한 산술연산자 의미

산술연산자	내용
+	class1에 대한 양성영향/class2에 대한 음성영향
-	class2에 대한 양성영향/class1에 대한 음성영향
*	두 자식트리에 대한 product 가중치 연결
/	두 자식트리에 대한 divide 가중치 연결

우수한 분류규칙을 발견하기 위해 기본적으로 학습 데이터에 대한 분류율을 유전자 프로그래밍의 적합도 함수로 사용한다. 또한 이해하기 쉬운 크기의 분류규칙을 얻기 위해 각 개체의 크기에 대한 평가를 적합도 평가에 추가한다. 일반적으로 동일한 성능을 내는 분류기의 경우보다 간단한 것이 일반화 능력이 뛰어나다고 알려져 있다. 본 논문에서 적합도는 아래의 수식과 같이 계산된다.

$$fitness\ of\ individual_i = \frac{number\ of\ correct\ samples}{number\ of\ total\ train\ data} \times w_1 + simplicity \times w_2$$

$$simplicity = \frac{number\ of\ nodes}{number\ of\ maximum\ nodes}$$

w_1 : *weight for training rate* w_2 : *weight for simplicity*

4. 실험 및 결과

4.1 실험환경

실험 데이터로는 웹상에 공개되어 있는 유전발현 데이터인 림프종 데이터를 사용하였다[9]. 림프종 데이터 (http://llmpp.nih.gov/lymphoma/)는 4026개의 유전자로 구성되어 있으며 총 47개의 샘플이 사용되었다. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-like DLBCL이다. 모든 샘플의 특징값은 정규화하여 사용하였다. 특징 수는 많지만 샘플 수가 매우 적기 때문에, 모든 샘플을 각각 테스트 데이터로 설정하고 그 나머지를 학습 데이터로 이용하고 총 47회의 실험결과를 합산하는, leave-one-out 방법으로 제안하는 방법의 성능을 평가하였다. 결과의 신뢰성을 위해 모든 실험은 10

회 반복하였고, 이들의 평균을 최종 결과로 사용하였다.

실험은 3가지 특징선택 방법으로 상위 30개의 특징을 각각 추출하고 추출된 특징을 이용한 산술트리를 구성하였다. 유전자 프로그래밍의 설정은 표 2와 같다. 개체의 적합도 평가 함수의 가중치 w_1과 w_2는 각각 0.9와 0.1로 설정하였다.

표 2. 실험 파라메터

Paramter	Setting
Population size	100
Maximum number of generations	50000
Selection probability	0.6~0.8
Crossover probability	0.6~0.8
Mutation probability	0.1~0.3
Permutation probability	0.1
Maximum depth of a tree	3
Elitism	*Yes*

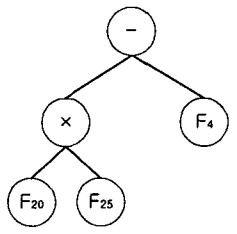
4.2 결과분석

표 3은 10회 반복 실험한 경우의 평균 및 최대 인식률을 보여준다. 통합 상관 특징추출 방법에 의해 추출된 특징들을 이용하여 분류규칙을 만든 경우가 유클리드 거리나 정보이득에 의해 추출된 특징들을 이용한 경우보다 학습에서나 인식에서 높은 성능을 보였으며, 평균적으로 90%의 인식률을 얻어 우수한 분류성능을 확인하였다.

표 3. 실험결과

특징추출 방법	학습율	인식률
유클리드 거리	0.82	0.63
정보이득	0.95	0.77
통합 상관 특징추출	0.99	0.90

그림 4는 반복된 실험에서 가장 자주 발생한 분류규칙을 보여주며, 이 분류규칙은 모든 샘플을 정확히 분류하였다. 사용된 특징에 대한 설명은 표 4에 기술하였고, 표 5는 이 분류규칙을 이용하여 Lymphoma dataset에 적용한 결과로, 각 샘플에 대한 계산값을 보여주며, class1은 양수로, class2는 음수로 정확히 분류되었다.



$(F_{20} \times F_{25}) - F_4$

그림 4. 100% 분류 규칙

표 4. 사용된 특징

특징 번호	DNA 번호	내용
F20	2229	(19539, (Unknown UG Hs.187478 ESTs Clone=1372156))
F25	702	(17614, *Protein tyrosine phosphatase, non-receptor type 12; Clone=289965)
F4	1277	(19274, (Unknown UG Hs.136345 ESTs Clone=746300))

5. 결론

본 논문은 DNA 유전발현 데이터의 효과적인 분석을 위하여 규칙발견 및 표현에 유용하고 생물의 진화과정을 모델로 한 방법인 유전자 프로그래밍을 사용하였다. 특징차원은 크고, 샘플의 수는 매우 적은 유전발현 데이터로부터 유의한 분류규칙을 추출하는 것은 매우 어려운 문제다. 특징추출을 수행하여 유용한 특징을 선택하고, 유전자 프로그래밍을 이용하여 선택된 특징들로 산술구조의 규칙을 생성하였다. 진화과정에서 얻어진 다양한 분류규칙으로부터 100%의 분류성능을 가지는 산술구조의 분류규칙을 발견할 수 있었다.

유전자 프로그래밍은 사람이 이해할 수 있는 수준의 분류규칙 생성에 유용하다. 논리 및 산술구조를 복합적으로 사용한 분류규칙은 보다 높은 설명력과 성능을 가질 것으로 예상된다. 또한 단일 분류규칙만을 사용할 때보다 다양한 분류규칙을 생성하여 이들을 결합한다면 보다 높은 분류 성능을 얻을 수 있을 것이다.

표 5. Lymphoma dataset 분류결과

No	Class	Value	No	Class	Value	No	Class	Value
1	1	0.80805103	17	1	0.60253795	33	1	0.623302
2	1	0.91941047	18	1	0.0014755	34	1	0.789822
3	0	-0.3012227	19	0	-0.8381169	35	0	-0.01361
4	0	-0.4325704	20	0	-0.6163975	36	0	-0.18281
5	1	0.24226597	21	1	0.347796	37	1	0.903315
6	1	0.08384417	22	1	0.489726	38	1	0.683536
7	0	-0.6274667	23	0	-0.29092	39	0	-0.10021
8	0	-0.585635	24	0	-0.74153	40	0	-0.00694
9	1	0.40721924	25	1	0.445419	41	1	0.35351
10	1	0.40769639	26	1	0.597151	42	1	0.230959
11	0	-0.7575564	27	0	-0.39227	43	0	-0.02769
12	0	-0.6155358	28	0	-0.35562	44	0	-0.18113
13	1	0.30338966	29	1	0.110855	45	1	0.065017
14	1	0.31195374	30	1	0.602598	46	0	-0.69592
15	0	-0.453115	31	0	-0.65939	47	0	-0.9116
16	0	-0.287293	32	0	-0.41936			

감사의 글

본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.

참고문헌

[1] A. Ben-Dor, et al., "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.

[2] A. Brazma and J. Vilo, "Gene expression data analysis," *Federation of European Biochemical Societies Letters*, vol. 480, pp. 17-24, 2000.

[3] C. Park and S.-B. Cho, "Genetic Search for Optimal Ensemble of Feature-Classifier Pairs in DNA Gene Expression Profiles," *Int. joint Conf. on neural networks*, pp. 1702-1707, 2003.

[4] K. Tan, et al., "Evolutionary computing for knowledge discovery in medical diagnosis," *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 129-154, 2003.

[5] A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," *Advances in Evolutionary Computation*, pp. 819-845, 2002.

[6] I. Falco, et al., "Discovering interesting classification rules with genetic programming," *Applied Soft Computing*, vol. 1, no. 4, pp. 257-269, 2002.

[7] J. Koza, "Genetic programming," *Encyclopedia of Computer Science and Technology*, vol. 39, pp. 29-43, 1998.

[8] H.-H. Won and S.-B. Cho, "Neural network ensemble with negatively correlated features for cancer classification," *Lecture Notes in Computer Science*, vol. 2714, pp. 1143-1150, 2003.

[9] A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.