림프종 암의 정확한 분류를 위한 산술연산자 분류규칙의 결합

홍진혁° 조성배
연세대학교 컴퓨터과학과
hjinh@sclab.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Ensemble of Classification Rules with Arithmetic Operators for the Accurate Classification of Lymphoma Cancer

Jin-Hyuk Hong° Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

앙상블은 다수의 분류기를 효과적으로 결합하여 분류의 성능을 향상시키는 대표적인 기술이다. 효과적인 앙상블을 위해서는 다양한 특성을 지닌 분류기를 확보하여야 한다. 기존의 앙상블은 개별 분류기의 결과를 바탕으로 분류기 사이의 의존성이나 유사성을 평가하여 분류기 결합을 시도하였다. 따라서 분류기 사이의 유사도의 정확한 측정에 한계를 지니고 있다. 본 연구에서는 이를 극복하기 위해서 다수의 산술연산자 기반 분류규칙을 유전자 프로그래밍을 이용하여 획득하고, 실제 표현형의 유사성을 측정한 후 이를 바탕으로 분류기를 결합한다. 생물정보학에서 많이 사용되는 유전자 데이터 중 하나인 림포마 암 데이터에 제안하는 방법을 적용하여 97% 수준의 높은 분류 성능과 해석가능한 분류규칙을 획득하였다.

1. 서론

다양한 분류기들로부터 얻어진 결과를 결합하여 분류를 수행하는 앙상블은 기계 학습에서 분류 성능을 향상시키기 위한 주요한 기술 중 하나이다[1]. 학습 데이터에 대해 최적의 성능을 보이는 분류기 하나를 이용하는 것 보다 유의한 다수의 분류기를 결합하여 분류를 수행하는 것이 많은 연구에서 더 우수한 성능을 보였다[2]. 앙상블을 이용하여 최대의 분류 효과를 획득하기 위해서 결합할 분류기들의 선택이 매우 중요하다. 동일한 학습 데이터에 대해 다른 패턴이나 성향을 내포한 분류기를 결합할 때에 다양한 패턴과 성향을 동시에 고려한 분류를 수행할 수 있다. 다양한 성향의 분류기를 얻기 위해서 보통 이종의 분류기를 학습시키거나 학습 데이터를 다양하게 구성하여 동종의 분류기를 학습시키는 방법이 많이 사용되며 Bagging이나 Boosting 등의 기술이 활발히 연구되고 있다[1]. 신경망 등을 분류기로 이용하는 대부분의 앙상블은 분류기를 직접 해석하기 어렵기 때문에 학습된 분류기의 분류 결과를 토대로 결합 구조를 설계한다. 따라서 분류기들 사이에 포함된 이종의 성향이 분류 결과가 유사하다는 이유로 무시되기도 한다.

본 논문에서는 이러한 한계를 극복하기 위해서 해석이 용이한 산술연산자 기반 분류규칙을 설계하고 유전자 프로그래밍을 통해 다수의 분류규칙을 획득한다. 분류규칙 사이의 유사성을 직접 평가한 후에 결합 구조를 구축한다. 분류 정확성과 해석성이 동시에 요구되는 암 분류에 제안하는 방법을 적용하여 그 유용성을 확인한다.

2. 배경

2.1 DNA Microarray를 이용한 암 분류

암에 대한 정확한 판단과 분류는 의학 분야에 있어서 매우 중요한 문제인 동시에 매우 어려운 문제이다[3]. 전통적인 형태적 징후 분석에 기반한 진료 방법은 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[3,4]. 사람의 유전자 정보는 최근 주목받는 DNA microarray 기술로부터 수집되며, 이들 유전발현 정보는 생명체에 관한 대량의 유전정보를 포함한다.

DNA microarray 기술은 기존 기술의 한계를 극복하고 초미세 단위로 유전 정보를 획득하고 하나의 칩 상에서 전체 염색체의 발현양상을 관찰하도록 한다. 이는 보다 복잡한 생물체 현상의 관찰과 분석을 가능하게 하였다[3,4]. 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 읽어들이면 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 산술연산자 기반 유전자 프로그래밍

유전자 프로그래밍은 사용자가 명시적으로 프로그래밍을 하지 않고 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[5]. 산술연산자 기반 분류규칙은 유전자 프로그래밍의 개체 표현형에서 사용되는 노드에 기본적인 산술연산자(+,-,×,/)를 적용한 것이다. 트리의 평가를 통해 얻어진 최종 계산값을 이용하여 분류를 수행하며, 높은 정확률과 해석성을 제공한다[6]. 표 1은 개체의 평가에 사용된 각 산술연산자의 유전자에 대한 의미를 표현한 것이다. 분류규칙은 아래와 같이 적용하였다. 그림 1에서처럼, *eval()* 함수는 한 개체의 부류에 대한 근사값을 계산하는 것으로 그 값이 양수일 때 class1, 음수인 경우에는 class2로 분류한다.

표 1. 유전자에 대한 산술연산자 의미

산술연산자	내용
+	class1에 대한 양성영향/class2에 대한 음성영향
-	class2에 대한 양성영향/class1에 대한 음성영향

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

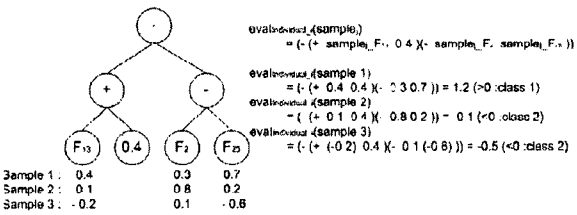

그림 1. 산술연산자 기반 유전자 프로그래밍 분류규칙

2.3 앙상블

일반적으로 하나의 분류기나 시스템이 항상 다른 것들보다 우수하지는 않으며, 여러 방법의 결합을 통해서 최종 분류기의 성능을 향상시킨다는 것이 앙상블의 기본 개념이다. 일반적으로 앙상블의 효과는 개별 분류기의 에러들 사이의 독립성에 의존적이기 때문에 개별 분류기의 성능과 다양성이 매우 중요하다[7]. 가장 기본적인 앙상블은 학습 데이터에 약간의 변화를 주어 다양한 분류기를 생성하는 것으로 Bagging과 Boosting이 그 대표적인 앙상블 기법이다.

Bagging은 Berimen에 의해 소개되었으며, 원본 학습 데이터를 분류기 수만큼 반복선택하여 학습 데이터 집합을 만든 후, 각각의 학습 데이터로 개별 분류기를 학습한다. 이렇게 학습된 분류기의 결과를 합산하여 최종 분류결과를 얻는다. AdaBoost는 Freund와 Schapire에 의해 소개된 Boosting의 대표적인 방법으로, 단계적으로 분류기를 적용하여 분류를 수행한다. 각 단계의 분류기의 에러값에 근거하여 다음 분류기의 학습을 위한 학습데이터를 구성한다. 최소값, 최대값, 평균, 중간값 및 투표 방법 등이 이들 개별 분류기의 분류결과를 합산하는 방법으로 많이 이용된다. 대부분의 앙상블 기법은 학습 데이터를 재구성하여 다양한 분류기를 생성하려고 하였으며, 개별 분류기의 분류결과를 이용하여 분류기들 사이의 유사도를 측정하였다. 분류기간의 정확한 유사도의 측정에 한계를 가지고 있다.

3. 분류규칙 발견 시스템

본 논문에서는 그림 2에서와 같이 고차원의 유전자를 가지는 유전자 발현 데이터에서 분류에 유용한 유전자들을 선택하고 유전자 프로그래밍을 이용하여 복수의 산술연산자 기반 분류규칙을 생성한다. 이들 다중 분류규칙을 규칙 사이의 유사도에 따라 결합하여 최종 분류 결과를 획득한다.

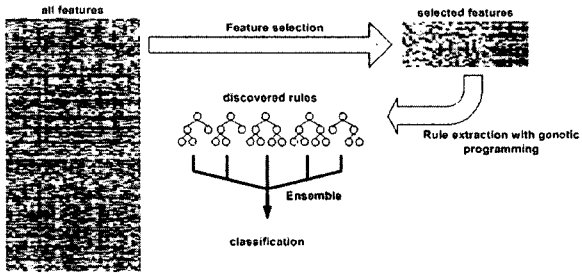

그림 2. 유전자 프로그래밍을 이용한 다중 분류규칙 생성

3.1 특징선택

유전자 선택은 학습속도를 향상시키고, 잡음을 줄이는 효과가 있으며, 특징의 중요성을 측정하는 기준으로 순위를 매겨 선택하는 순위-기반의 방법과 분류기와 연계된 학습데이터 자체의 특성을 이용한 방법으로 구분된다. 본 논문에서는 피어슨 상관계수, 유클리드 거리와 신호 대 잡음비 특징선택 방법을 이용하였다. 이들 방법은 앞서 수행된 산술연산자 기반 분류규칙 발견 연구에서 우수한 성능을 보인바 있다 [6].

일반적으로 선택되는 특징의 개수가 너무 적으면 정보를 제대로 표현하지 못하는 문제가 있고, 너무 많으면 노이즈가 많이 포함된다는 문제가 있다. 따라서 적절한 특징 개수의 선택이 중요한데, 유전자 발현 데이터의 경우 20~70개 정도를 사용하였을 경우 큰 차이없이 비슷한 성능이 나왔다 [4]. 따라서 본 논문에서는 각각 30개의 유전자들을 선택하여 분류규칙 발견에 사용한다.

3.2 분류규칙 추출

본 논문에서는 각 실험에서 10개의 산술연산자 기반 분류규칙을 추출한다. 각 특징선택 방법에 의해 뽑힌 30개의 특징값과 기본적인 산술연산자(+,-)를 이용하여 트리를 만들어 유전자 프로그래밍 개체의 표현형으로 이용하였다. 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 계산되며, 가중치 w_1과 w_2는 각각 0.9와 0.1로 설정하였다.

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

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

w_1 : *weight for training rate* $\quad$ w_2 : *weight for simplicity*

3.3 규칙결합

데이터로부터 분류기나 규칙을 생성할 경우 데이터에 의존적인 해를 종종 발견하기도 한다. 이는 분류기의 일반화 능력을 떨어뜨리기 때문에 분류 성능을 저하시킨다. 특히 생물정보 데이터와 같이 학습이나 진화를 위한 샘플이 적을 경우, 이런 현상은 빈번히 발생한다. 본 논문에서는 보다 안정적이고 높은 정확률을 얻기 위해 해석가능한 다수의 분류규칙을 획득한 후, 이들 사이의 유사도를 측정하고 차별성이 높은 5개의 분류규칙을 선택한다. 이들의 분류결과를 투표 결합으로 합산하여 최종 분류를 수행한다.

분류규칙은 그림 3과 같이 각 규칙사이의 유사도 순위에 따라서 총 5개의 규칙이 선택된다. 두 규칙 사이의 유사도는 트리의 각 노드의 타입과 값을 비교하여 계산한다.

```
R: A set of extracted rules {r1, r2, ..., r10}
S: A set of selected rules {s1, s2, ..., s5}
  For i=1 to 10 {
   For j=i+1 to 10 {
     sij = calculate_similarity(ri,rj);
   }}

  SR = sorting(sij);
  sij = SR(first);
  While k≤5 {
   if(i∉S)
     include i into S
     k++;
   if(j∉S)
     include j into S
     k++;
   sij = SR(next);
  }
```

그림 3. 분류규칙 선택 알고리즘

4. 실험 및 결과

4.1 실험환경

실험 데이터로는 웹상에 공개되어 있는 유전발현 데이터인 럼프종 데이터를 사용하였다[8]. 럼프종 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 10회 반복하였고, 이들의 평균을 최종 결과로 사용하였다.

표 2. 실험 파라메터

Paramter	Setting
Population size	100
Maximum number of generations	10000
Selection probability	0.8
Crossover probability	0.7
Mutation probability	0.3
Permutation probability	0.1
Maximum depth of a tree	3
Elitism	*Yes*

4.2 결과분석

표 3과 그림 4는 림포마 암 분류에 대해 10회 반복 실험한 경우의 평균 학습율과 인식율을 보여준다. 두 방법 모두 학습 데이터를 정확히 분류하는 규칙을 획득하였고 이들 분류 규칙을 결합하여 97%의 인식률을 얻었다. 이는 단일 분류규칙을 사용하였을 경우보다 높은 수치로써, 보다 정확한 분류를 수행하는 것을 확인하였다.

표 3. 실험결과 (림포마)

특징추출 방법	학습율	인식률	인식률(단일)
코사인 계수	100%	97.0%	95.5%
유클리드 거리	100%	97.0%	95.9%
신호 대 잡음비	100%	97.2%	96.6%

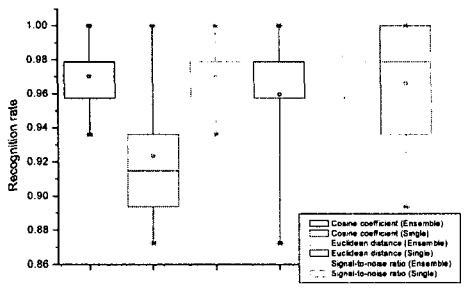

그림 4. 인식률 (림포마)

그림 5는 신호 대 잡음비 특징을 사용한 실험에서 가장 자주 발생한 분류규칙을 보여주며, 이 분류규칙은 모든 샘플을 정확히 분류하였다. 사용된 특징에 대한 설명은 표 4에 기술하였으며, 그 중 F14는 실제 림포마 암과 관련된 유전자임을 확인하였다 [9].

5. 결론

본 논문은 DNA 유전발현 데이터의 효과적인 분석을 위하여 규칙발견 및 표현에 유용하고 생물의 진화과정을 모델로 한 방법인 유전자 프로그래밍을 사용하였다. 특징추출을 수행하여 유용한 특징을 선택하고, 유전자 프로그래밍을 이용하여 선택된 특징들로 다중의 산술구조의 규칙을 생성하였다. 이들 규칙을 표현형의 유사도를 바탕으로 결합하여 최종 분류를 수행함으로 높은 정확률과 안정성을 확보하였다. 또한 진화과정에서 얻어진 다양한 분류규칙으로부터 100%의 분류성능을 가지는 산술구조의 분류규칙을 발견하였으며, 이들 분류규칙이 림포마 암과 관련되어 있음을 확인하였다.

유전자 프로그래밍은 사람이 이해할 수 있는 수준의 분류규칙 생성에 유용하다. 논리 및 산술구조를 복합적으로 사용한 분류규칙은 보다 높은 설명력과 성능을 가질 것으로 예상된다. 또한 다중의 분류규칙을 결합할 때에 다양한 결합 기법을 적용한다면 보다 높은 성능을 얻을 수 있을 것이며, 이들 규칙을 효과적으로 분석하여 분류에 유용한 정보를 추출하는 방법도 매우 유용할 것이다.

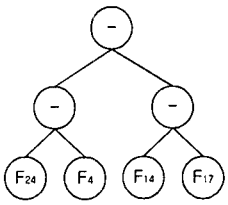

그림 5. S2N 특징 기반 100% 분류규칙

표 4. 그림 6의 규칙에 사용된 유전자 정보

특징 번호	DNA 번호	내용
F24	684	Unknown; Clone=1352715, 14377
F4	1279	*Unknown; Clone=825199, 19288
F14	1914	Lymphotoxin-Beta=Tumor necrosis factor C; Clone=1320296, 13297
F17	680	*Unknown; Clone=1372162, 19541

감사의 글

이 연구는 과학기술부가 지원한 뇌과학 연구 프로그램에 의해 지원되었음.

참고문헌

[1] R. Bryll, et al., "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, no. 6, pp. 1291-1302, 2003.

[2] A. Sharkey and N. Sharkey, "Combining diverse neural nets," *The Knowledge Engineering Review*, vol. 12, no. 3, pp. 231-247, 1997.

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

[4] 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.

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

[6] J.H-Hong and S.B. Cho, "Lymphoma cancer classification using genetic programming with SNR features," *Lecture Notes in Computer Science*, vol. 3003, pp. 78-88, 2004.

[7] A. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 2, no. 3, pp. 75-83, 2003.

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

[9] P. Koni and R. Flavell, "A role for tumor necrosis factor receptor type 1 in gut-associated lymphoid tissue development: genetic evidence of synergism with lymphotoxin β," J. of Experimental Medicine, vol. 187, no. 12, pp. 1977-1983, 1998.