기계학습에 의한 압타머칩 데이터 기반 심혈관 질환 단계의 예측

김병희°[1], 김성천[2], 장병탁[1]
[1]서울대학교 컴퓨터공학부 바이오지능 연구실
{bhkim, btzhang}@bi.snu.ac.kr
[2](주) 제노프라
kimgp@cotech.co.kr

Estimation of the steps of cardiovascular disease by machine learning based on aptamers-based biochip data

Byoung-Hee Kim°[1], Sung-Chun Kim[2], and Byoung-Tak Zhang[1]
[1]BI Lab, School of Computer Sci. & Eng., Seoul National University
[2]GenoProt Inc.

요 약

압타머칩은 (주)제노프라에서 개발한 새로운 개념의 바이오칩으로서, 압타머(aptamer)를 이용하여 혈액 중의 특정 단백질군의 상대적인 양의 변화를 측정할 수 있으며, 질병 진단에 바로 응용할 수 있는 도구이다. 본 논문에서는 압타머칩 데이터 분석을 통해 심혈관 질환 환자의 질병 진행 단계를 예측할 수 있음을 보인다. 정상, 안정/불안정성 협심증, 심근경색의 네 단계로 표지된 환자의 혈액 샘플로부터 제작한 (주)제노프라의 3K 압타머칩 데이터를, 일반 DNA 마이크로어레이 분석과 동일한 과정을 거쳐 분류한 결과, 각 단계별 환자샘플이 확연히 구분되는 것을 확인하였다. 분산분석 결과 p-value를 이용하여 자질 선택을 수행하고, 분류 알고리즘으로는 신경망, 결정트리, SVM, 베이지안망을 적용한 결과, 각 알고리즘별로 50대 남성환자 31개의 샘플에 대하여 77~100%의 정확도로 심혈관 질환의 단계를 구분해내었다.

1. 서 론

DNA 마이크로어레이를 통해 수천개의 유전자(gene)의 발현 패턴을 동시에 분석하여 정상인과 환자를 구분하고, 질병, 특히 암의 다양한 형태를 구분해낼 수 있음은 [1-3] 이래 많은 연구 결과가 나와 있다. 최근에는 기존 DNA 마이크로어레이를 통한 유전자 발현 프로파일링 기법을 넘어 다양한 형태의 마이크로어레이가 개발되어 질병 예측 등의 임상분야에의 응용을 확대해가고 있다[4]. 최근 (주) 제노프라에서 개발한 압타머칩(aptamers-based biochip, ABB)은 혈청(serum) 중의 단백질의 양의 변화를 병렬적으로 분석 가능하게 한 마이크로어레이 형태의 바이오칩이다. 현재까지 다양한 질병과 연관된 단백질군이 실험적으로 선정되어 있으며, 압타머칩을 통해 이 단백질군의 혈청 내 발현 패턴을 살펴봄으로써 질병 진단을 할 수 있음이 알려져 있다.

본 논문에서는 심혈관 질환 진단을 위해 제작된 압타머칩에 대해, 기계 학습 기법을 이용한 칩 데이터를 분석을 통해 심혈관 질환의 단계를 구분해낼 수 있음을 보인다.

상관 분석과 클러스터링 기법을 적용하여 일반적인 의학적 구분 기준인 40대와 50대의 질병 진행의 차이가 압타머칩 데이터에서도 나타남을 확인하며, 기계학습 분야의 다양한 분류(classification) 알고리즘을 이용하여 환자가 심혈관 질환의 네 단계 중 어느 단계에 속하는지를 높은 정확도로 구분해낼 수 있음을 보인다.

2. 관련 연구

2.1 압타머칩(ABB)

압타머(aptamer)란 단일염기서열인 DNA 나 RNA로 항원 항체 반응과 같이 타겟 물질에 대한 특별한 친화력과 특이성을 나타내는 생체정보 감지 소재이다. 일반적으로 단일 핵산 가닥이 linear 한 것과 달리 복잡한 3차원적 구조를 가진다. (주)제노프라는 여러조합의 RNA library에서 특이질병에 관여하는 RNA 분자를 찾아내고 분리해 낼 수 있는 시험관증폭선택법(Selection Evolution of Ligands by Exponential Enrichment, SELEX)이라는 기술을 이용하여 표적단백질과 특이적으로 결합하는 새로운 RNA 구조를 만들어 냈다. 이러한 RNA 압타머와 상보적으로 결합하는 capture DNA (anti-aptamer sequence)를 글라스 표면에 고정화하여 마이크로어레이화한 것이 압타머칩이다. 이러한 압타머칩을 통해 대량의 표적단백질의 패턴을 파악할 수 있으며, 2006년 초에 3,000개의 단백질 패턴을 하나의 칩으로 확인할 수 있는 3K 압타머칩이 개발되었다.

압타머칩은 현재 혈액 내 단백질 양의 변화를 통해 질병 유무를 판별하고, 질병 특이적인 단백질을 탐색하는 데 활용되고 있다.

2.2 마이크로어레이/기계학습을 이용한 질병 진단

질병의 단계 또는 종류를 구분하는 데에는 기계학습(machine learning)의 한 방식인 감독학습(supervised learning), 즉, 분류(classification) 기법을 적용할 수 있다. 분류 기법은 개체가 속한 범주(category, class)가 미리 알려진 '학습' 데이터 및 '테스트' 데이터에 대해 학습 및 성능 평가를 수행한다[5]. 질병 분류 문제의 경우 환자의 의학적 표지를 "gold standard"로, 즉 범주의 정답으로 보고 범주로 활용하게 된다.

기계학습을 이용한 질병 진단은, 증거기반의료(EBM)[6], 임상의사결정지원시스템(CDSS)[7] 등의 개념이 확산됨에 따라 꾸준한 조명을 받고 있으며, 특히 마이크로어레이는 1990년대 후반 이래 지속적으로 임상분야의 응용을 위한 시도가 이어지고 있다. 암에 대해서는 분자 수준의 진단을 통해 정확하고 객관적이며 체계적인 암 분류가 가능하다는 것이 정설[8]이며, 특이적인 분자적 마커(marker)가 밝혀진 경우 확실한 분류가 가능하다. 그러나 대부분의 질병에 대해 분자적 마커를 찾는 작업은 요원한 일이다. 하지만, 명확한 마커가 없더라도 바이오칩을 통해 분자 수준의 종합적인 패턴을 얻고 다양한 기계학습 기법으로 분류하는 것이 가능하다. 순수하게 마이크로어레이의 통계적, 기계학습 기법을 통한 분자적(molecular) 분류를 통해서도 공통 성인성 악성종양 등의 진단이 가능함을 보여준 연구사례는 지속적으로 발표되고 있으며[1-3, 9], 또한 마이크

로어레이를 이용한 바이오마커 탐색도 활발한 연구가 진행되고 있다. 일반적인 마이크로어레이 데이터 분석 과정 및 개념은 [5]에 상세히 확인할 수 있다.

[9]에서는 혈구에 대한 유전자 발현 패턴 분석을 통해 헌팅턴 병(Huntington's disease)의 진단이 가능함을 보였다. 이는 압타머칩을 이용한 질병 진단의 가설인 '질병의 진행을 파악할 수 있는 정보가 혈청(serum) 중 단백질 양의 변화에 반영됨'을 뒷받침하는 한 근거사례가 된다.

3. 데이터 및 분석방법

3.1 데이터

분석에 사용한 데이터는 (주)제노프라에서 생산한 3K 압타머칩 데이터로서, 3천개의 단백질의 양을 측정할 수 있다. 전문의가 표지한 정상, 안정형 협심증(stable angina), 불안정형 협심증(unstable angina), 심근경색(myocardial infarction)의 네 단계로 분류되는 40대 남성 12명, 50대 남성 31명의 혈청에서으로부터 생성되었다. 각 단계별 샘플의 수는 [표 1]과 같다.

표 1. 분석에 사용한 심혈관 질환 관련 환자의 샘플 구성

질환단계 / 나이	정상	안정형 협심증(SA)	비안정형 협심증(UA)	심근경색 (MI)
40대	3	3	3	3
50대	9	5	7	10

GenePix Pro를 통해 생성된 데이터에 대해 'ratio of medians'값을 추출하여 샘플의 중앙값으로 imputation을 실시하고 로그변환을 실시한 후, 칩 실험시 PMT volt를 조절함에 따라 발생한 두 채널(Cy5, Cy3)의 차이효과를 제거하기 위해 각 샘플의 평균값을 0으로 평행이동한다(linear normalization).

3.2 자질 선정(feature selection)

일원배치 분산분석법(one-way ANOVA)을 이용하여 각 단백질별로 측정한 p-value가 낮은 순으로 단백질을 추출하여 추가 분석에 사용한다.

3.3 상관 분석 및 클러스터링

의학적으로 심혈관질환의 패턴은 남성의 경우 크게 나이를 기준으로 40대와 50대 이상으로 구분한다[10]. 압타머칩에서 이러한 경향이 나타나는지를 살펴보기 위하여, 40대/50대 각 샘플군에서 독립적으로 분산분석을 실시한 후 각 군에서 p-value 기준 상위 1000개의 단백질 중 중복되는 353개의 단백질을 선택하여 상관 분석 및 계층적 클러스터링을 시행한다. 상관 분석 및 계층적 클러스터링에서의 유사도 측정 기준은 모두 피어슨(Pearson) 상관계수를 적용한다. 계층적 클러스터링을 통해 샘플의 단계구분에 대한 사전 정보가 없는 상태에서 심혈관 질환 각 단계별 환자의 군집형성 여부도 확인한다.

3.4 분류 알고리즘을 적용하여 단계 분류성능 측정

2.2절에서 설명한 바와 같이 기계학습 기법 중의 분류 알고리즘을 이용하여 심혈관 질환의 네 단계가 구분되는지를 살펴본다. 분류에 관여하는 단백질은 3.2절에서 설명한 바와 같이 일원배치 분산분석 결과 p-value가 낮은 정도를 기준으로 상위 N개를 선택한다. N의 값은 10~1,000 사이의 여러 값을 선택하여, 단백질 수에 따른 분류결과를 관찰한다. 분류 알고리즘으로는 state-of-the-art 기법인 결정트리, 인공신경망, SVM, 베이지안망을 적용하며, 각 알고리즘별 학습환경 설정은 [표 2]와 같다. 성능 평가는 정확도(accuracy)를 LOOCV (leave-one-out cross-validation) 방식으로 측정한다. 즉, 31개 샘플 중 30개를 학습에 사용하고 1개를 테스트로 사용하되, 모든 샘플이 한 번씩 테스트되도록 하여 31번 시행한 결과의 평균값을 성능의 척도로 사용한다.

표 2. 압타머칩 데이터 분류에 사용한 기계학습 알고리즘 및 환경설정내역

알고리즘	세부 알고리즘	환경변수 및 설정내역
결정트리	J4.8	pruning: ON, subtree raising: ON
인공신경망	MLP	learning rate = 0.2, momentum=0.1 hidden layer/#units = 1/3
SVM	SMO	polynomial kernel, exponent = 1.0 multi-class: pairwise
베이지안망	TAN	structure scoring: Bayes

4. 결과 및 분석

4.1 상관 분석 및 클러스터링

[그림 1]은 40대, 50대 샘플에 대한 상관분석 결과이다. 박스로 구분한 영역을 살펴보면, 샘플이 대체로 나이와 심혈관질환 단계별로 구분됨을 볼 수 있다. 즉, 40대 및 50대의 각 심혈관 질환 단계에 속하는 샘플 간에는 혈청 내 353개 단백질의 분포 패턴이 대체로 높은 상관관계를 보인다. 50대의 경우 정상/안정형(SA)의 경우와 질환이 심화된 비안정형(UA)/심근경색(MI) 간에는 음의 상관계수가 나타난다. 특기할 사항은 40대 환자의 경우 단계별 구분이 명확하지 않으며 50대 비안정형(UA)과 양의 상관관계가 비교적 크게 나타난다는 점이다.

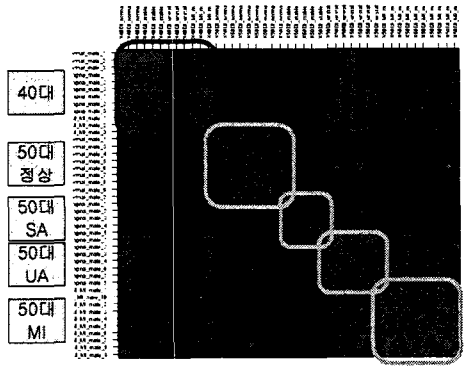

그림 1. 40대, 50대 샘플에 대한 상관분석 결과. Pearson 상관계수값을 기준으로 최소값(-1.0, 녹색)부터 최대값(1.0, 빨강색)까지 표현하였다.

[그림 2]는 40대, 50대 샘플에 대한 계층적 클러스터링 결과 덴드로그램(dendrogram)이다. 트리구조에 원형 표시를 한 바와 같이, 40대 환자는 전체적으로 하나의 클러스터를 형성하며, 50대 환자는 대체로 각 단계별 구분이 명확하다. 40대의 클러스터에 50대 비안정형 2명(화살표 표시)이 포함되어 있으며, 심근경색(MI) 환자 3명이 바로 이전 단계인 비안정형 환자와 같은 클러스터에 포함되어 있는 점이 특징적이다.

상관 분석과 클러스터링 결과를 종합하면, 압타머칩 데이터에 심혈관 질환의 40대와 50대의 차이 및 각 질환 단계별 차이가 잘 구분되어 나타난다는 점을 확인할 수 있다.

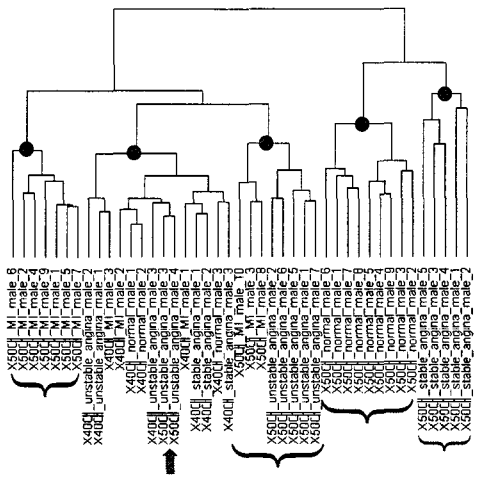

그림 2. 40대, 50대 샘플에 대한 average-linkage 계층적 클러스터링 결과. 유사도 측정 기준은 Pearson 상관계수를 적용.

4.2 분류 알고리즘을 이용한 심혈관 질환 단계 예측

[그림 3]은 3.2절에서 설명한 자질 선정 방법으로 선택한 단백질 수에 따른 각 분류알고리즘의 심혈관 질환 단계별 예측 정확도 측정 결과이다. 임의(random) 분류시의 정확도(25%)를 크게 웃도는 정확도를 보인다. 다만, 샘플의 수 31은 각 단계별 패턴을 종합적으로 반영하기에는 상당히 부족한 수라는 한계가 있으며, 이 분류 결과에 대해서는 각 단계별 구분이 어느 정도 명확히 된다는 점에만 의미를 두는 것이 옳다. 분류에 사용하는 단백질의 수는 100~200 정도가 적절한 것으로 판단된다. 이보다 적은 경우 분류 결과를 살펴보면 판별의 안정성이 부족해 보이며, 많은 경우 샘플 수에 비해 자질의 양이 커지므로 오버피팅(overfitting)의 우려가 있다. [표 3]에는 각 알고리즘별로 최적의 성능을 보이는 경우의 단백질 수를 정리하였다.

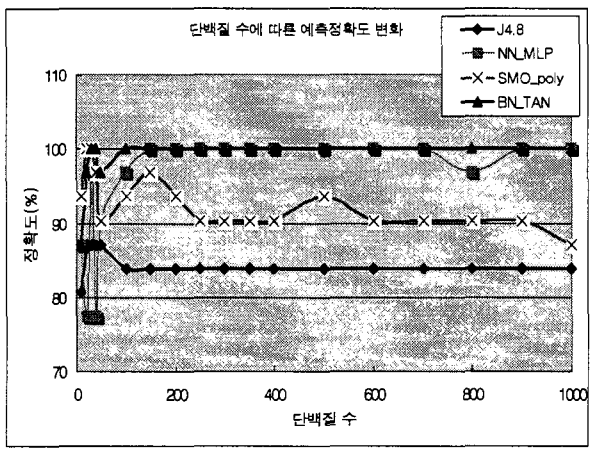

그림 3. 네 가지 분류 알고리즘을 이용한, 단백질 수에 따른 심혈관질환 예측성능 예측결과.

표 3. 각 알고리즘별 최고 예측정확도 및 해당 단백질 수. 동일한 예측정확도이면 단백질 수가 적은 경우를 선택

알고리즘	세부 알고리즘	예측정확도	단백질 수
결정트리	J4.8	87.1%	20
인공신경망	MLP	100%	150
SVM	SMO	100%	20
베이지안망	TAN	100%	30

5. 논의 및 결론

본 논문에서는 심혈관 질환 환자의 압타머칩 데이터에 대해, 40대 환자와 50대 환자의 차이가 데이터에 나타나며, 50대 환자 샘플이 심혈관 질환 각 단계별로 구분됨을 확인하였다. 이러한 확인을 바탕으로 50대 환자 31명에 대한 여러 분류(classification) 알고리즘의 예측 정확도(accuracy)를 LOOCV 방식으로 평가한 결과 77%~100%의 정확도를 얻었다.

요컨대, 분류(classification) 알고리즘을 이용한 심혈관 질환 각 단계별 예측 가능성을 확인하였다

샘플의 수가 보충되어 보다 정밀하고 객관적인 성능 평가를 내리는 것은 차후의 과제이다. 나아가 다양한 질병 샘플이 확보되면 각 질병의 예측 가능성을 본 논문의 절차를 이용하여 확인할 수 있을 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실사업(NRL)에 의하여 지원되었음. 김병희는 서울과학장학생 사업에서 지원받았음.

참고문헌

[1] Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**:531-537, 1999.
[2] U. Sherf, *et al.*, A gene expression database for the molecular pharmacology of cancer, *Nature Genetics*, **24**:236-244, 2000.
[3] S. Ramaswamy, *et al.*, Multiclass cancer diagnosis using tumor gene expression siagnatures, *PNAS*, **98**(26): 15149-15154, 2001.
[4] M. T. Barrett, Stacking the chips for biological discovery, *Nature Genetics Supplement*, **37**:S1, 2005.
[5] D. B. Allison, *et al.*, Microarray data analysis: from disarray to consolidation and consensus, *Nature Reviews Genetics*, **7**:55-65, 2006.
[6] D. L. Sackett, *et al.*, Evidence based medicine: what it is and what it isn't: It's about integrating individual clinical expertise and the best external evidence, *British Medical Journal*, **312**:71-72, 1996.
[7] D. Hunt, *et al.*, Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review, *JAMA*, **280**:1339-46, 1998.
[8] J. I. Connoly, *et al.*, in Cancer Medicine, eds. J. F. Holland *et al.*, Williams & Wilkins, 1997, pp.533-555.
[9] F. Borovecki, *et al.*, Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease, *PNAS*, **102**(31):11023-11028, 2005.
[10] R. M. Conroy, *et al.*, Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project, *European Heart Journal*, **24**(11):987-1003, 2003.