신경망을 이용한 천식 발병 예측 모델

최현주*, 김승현**, 위규범*
*아주대학교 정보통신전문대학원
**아주대학교 의과대학
e-mail:{yoyo194, kimsh, kbwee}@ajou.ac.kr

A Prediction Model for Asthma using ANN

Hyunju Choi*, Seung-hyun Kim**, Kyubum Wee*
*Graduate School of Information & Communication, Ajou University
**School of Medicine, Ajou University

요 약

신경망은 복잡한 데이터에서 일정한 패턴을 찾아 이를 분류하는 능력이 뛰어난 모델이다. 그러나 다량의 데이터가 입력으로 들어오면 연산에 오랜 시간이 걸리고 패턴을 찾기가 어려워진다는 한계가 있다. 본 연구에서는 set association과 의사결정나무를 이용하여 신경망에 입력되는 데이터의 수를 줄여서 다량의 데이터에도 적용 가능하며 예측의 정확도를 높인 신경망 모델을 구성하였다. 이 모델을 천식 관련 SNP 데이터에 적용하여 천식 발병 여부를 예측한 결과, 각각의 방법을 독립적으로 사용했을 때보다 높은 예측 정확도를 얻었다.

1. 서론

SNP은 인간의 게놈에서 약 1000개 중 하나의 빈도로 나타나는 동일하지 않은 염기를 말한다. SNP이 단백질을 코딩하고 있는 부분에 있는 경우에는 그 유전자가 생산해 내는 단백질이 달라질 수 있고, 변형된 단백질로 인해 질병이 유발될 수 있다. 또한 SNP이 단백질을 코딩하고 있지 않은 조절 부위에 존재하는 경우에는 유전자 발현 조절과정에 변화가 생겨 질병이 유도 될 수도 있다. 따라서 환자와 정상인의 SNP을 비교 분석하여 질병과 관련 있는 SNP을 찾아낼 수 있을 것이다. SNP 연구는 맞춤 의학이라는 질병의 치료 관점에서도 의의가 있는데 개개인의 약물에 대한 반응의 차이도 SNP과 관련이 있기 때문이다. 그러므로 SNP 정보를 이용하면 개개인에게 적합한 치료 또한 가능해 질 것이다.

그러나 이처럼 질병과 관련된 SNP을 찾는 연구가 큰 의의를 지니는 시급한 과제임에도 불구하고 SNP 데이터의 방대함과 복잡도로 인해 연구에 많은 어려움이 있는 실정이다.

2. 신경망을 이용한 SNP 연구 목적

신경망은 복잡한 데이터에서 일정한 패턴을 찾아 이를 분류하는 능력이 뛰어난 모델로 특별히 비선형의 복잡한 데이터에 좋은 성능을 보인다. 이러한 이유로 본 실험에서는 신경망을 통해 천식 관련 유전자들을 찾고 나아가 천식 발병 여부를 예측해 보았다.

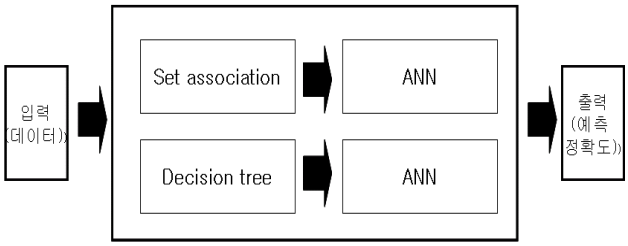
이러한 접근 방법의 이전연구는 신경망에 PDM (parameter decreasing method) 기법을 적용하여 CAA(childhood allergic asthma)와 관련된 SNP을 제시한 것이 있다 [1][2].

그러나 데이터의 수가 많아지면 전체 SNP에서 모든 SNP를 돌아가면 하나씩 제거하여 모든 경우를 관찰하는 PDM 방법을 적용하기가 힘들어진다. 신경망 자체도 입력 데이터의 종류가 너무 많으면 연산에 오랜 시간이 걸릴 뿐만 아니라 패턴을 찾기가 어렵기 때

이 논문은 2006년도 정부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2006-000-10775-0)

문에 다량의 데이터에 PDM을 적용하는 데는 더더욱 한계가 있다.

본 연구에서는 이를 극복하기 위해 set association과 의사결정나무를 이용하여 신경망에 입력되는 SNP의 수를 줄였다. 즉, (그림 1)에서와 같이 일차적으로 set association과 decision tree를 이용하여 중요 SNP을 선별한 후 이를 신경망의 입력 값으로 주어, 다량의 데이터를 이용한 천식의 발병 여부 예측을 가능하게 하고 예측에 소요되는 시간을 줄이며 궁극적으로 천식 발병 예상의 효율성을 높이고자 하였다.



(그림 1) 신경망 기반 천식 발병 예측 모델 과정

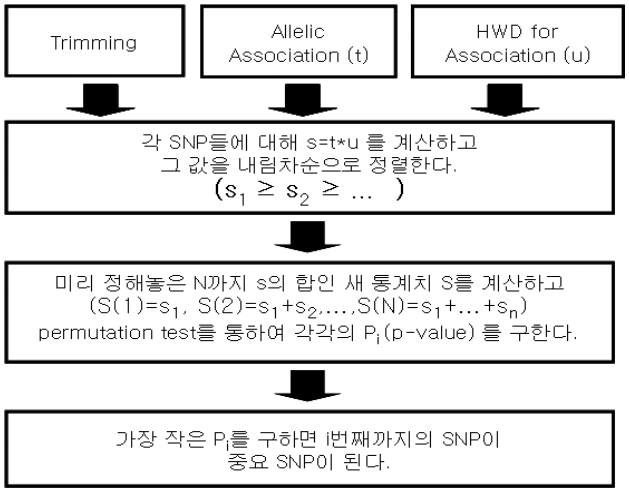
실험은 아스피린 복용으로 유도되는 아스피린 과민성 천식(AIA)과 아스피린 내인성 천식(ATA), 정상인(NC)의 세 그룹으로 나누어 행해진다. 특별히 천식 환자군을 두 그룹으로 나눈 이유는 AIA와 ATA 사이의 특징적인 SNP을 찾으려 함이다. 실제 임상에서는 AIA인지 ATA인지에 따라 각기 다른 치료가 행해지는데 병력만으로 이러한 진단을 내리기 힘든 경우가 많기 때문에 약물 유전체학 관점에서 빠르고 정확하게 이 둘을 구분할 수 있는 SNP를 찾으려는 것이 본 실험의 또 다른 목적이다[3].

2. 방법

2.1. Set association

Set association은 allelic association과 Hardy-weinberg equilibrium 정보를 통해 데이터에서 중요한 SNP의 집합을 찾는 방법이다. 이를 위해 각 SNP에 대하여 SNP과 질병유무와의 관계 정도를 측정한 χ^2-통계치인 AA 통계치와 Hardy-Weinberg equilibrium을 귀무가설로 하여 각 SNP의 분산 정도를 측정한 χ^2- 통계치인 HWD 통계치가 필요하다.

특별히 환자 그룹에서 HWD 편차가 높은 SNP들은 그 SNP이 질병과 관련이 있다는 의미이다. 반면에 정상인 그룹에서 HWD 편차가 높은 SNP들은 genotyping error를 의미하므로 해당 SNP를 제거하거나 '0'으로 맞춰주는 트리밍(trimming) 작업이 필요하다. 본 연구에서는 49개 SNP 중 X 염색체 위에만 존재하여 이미 평형이 깨어진 SNP가 있어서 이를 제거하고 실험을 행하였다. set association 방법은 (그림 2)와 같다[4].



(그림 2) Set association 방법의 요약

2.2. 의사결정나무(decision trees)

의사결정나무는 관심대상을 적절한 기준에 의해 몇 개의 소집단으로 분류하는 방법이다. 관심대상 전체를 시작 노드에 넣고 이를 적절한 기준으로 이용하여, 비슷한 특성을 공유하는 소집단으로 반복적으로 나누다가 더 이상 나누어지지 않는 노드에 이르면 수행을 종결하는데, 나무의 맨 끝에 있는 마디가 의사결정나무 수행 결과 생성되는 클래스이다. 이러한 의사결정나무는 분석 과정을 쉽게 이해할 수 있고 분류 성능 또한 좋기 때문에 다양한 분야에서 이용되고 있다[5].

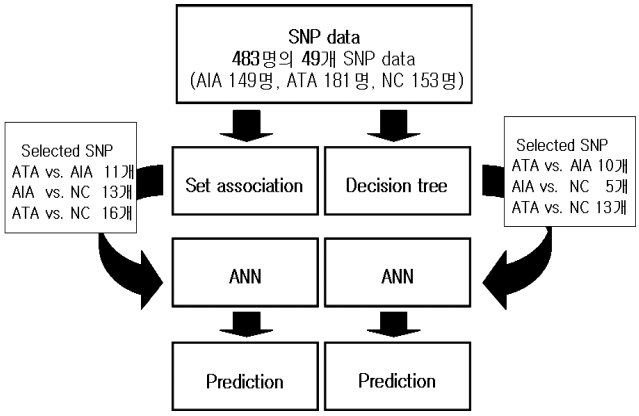
2.3. 신경망

신경망은 뇌의 구조와 동작방식을 단순화 시켜 수학적으로 모델링한 잘 알려진 계산모형이다. 특별히 은닉층은 신경망에 비선형의 특성을 부여하여 적당한 은닉 노드들만 있다면 복잡한 비선형 문제들을 푸는 것을 가능하게 한다고 알려져 있다[5].

그러나 적당한 은닉 노드의 수를 결정하는 것이 쉬운 문제가 아니다. 만약 너무 적은 수의 히든 노드를 사용하면 학습이 충분히 되지 않아 언더피팅(underfitting)과 높은 바이어스(bias)가 생길 수 있고, 반대로 너무 많은 히든 노드를 사용하면 과학습으로 인한 오버피팅(overfitting)과 일반화에 있어서 여전히 높은 에러를 갖게 되기 때문이다. 현재까지 몇 개의 네트워크를 훈련시키고 각각의 에러를 일반화시켜 추

정하는 방법 외에는 대부분의 상황에서 최적의 히든 노드수를 결정하는 방법은 없다고 알려져 있다.

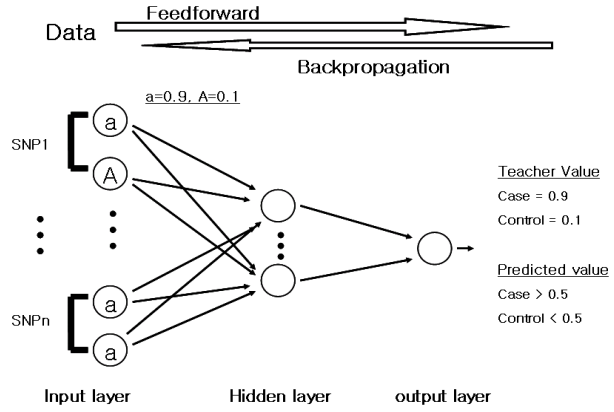
3. 실 험



(그림 3) 신경망 기반 천식 발병 예측 실험의 요약

실험은 49개 SNP에 대한 483명(AIA 149명, ATA 181명, NC 153명)의 데이터을 AIA vs. ATA, AIA vs. NC, ATA vs. NC의 세 그룹으로 나누어 진행되었다.

(그림 3)에서와 같이 각각의 그룹에서 일차적으로 set association과 의사결정나무를 이용하여 중요한 SNP들을 선별하였고 이를 신경망의 입력 값으로 입력하였다.



(그림4) 실험에 사용된 신경망 모델

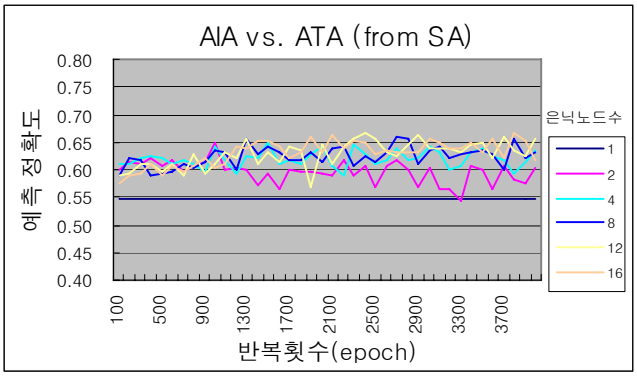
신경망은 (그림 4)와 같이 전방향 (feedforward) 네트워크를 사용하였고 학습에는 오류 역전파 알고리즘을 사용하였다. 신경망에 입력 값을 줄때 각 SNP의 지노타입에 따라 다수의 대립 유전자(major allele)에는 0.1, 소수의 대립유전자 (minor allele)에는 0.9의 입력 값을 주게 된다. 각 조건에서 10-fold 크로스확인을 통해 10번의 학습이 이루어진다. 각 경우마다 최고의 예측을 위해 은닉노드는 1개에서 14개까지, 반복횟수(epoch)는 100번에서 4000번까지 변화시켜가며 신경망을 구성하여 민감도(sensitivity),특이도(specificity), 정확도(accuracy)를 측정하였다.

4. 결과

각 시행방법의 비교는 정확도(accuracy)에 의한다. 여기서의 정확도는 10-fold 크로스확인의 결과를 평균내어 얻은 평균정확도이다.

4.1. set association과 신경망을 이용한 결과

AIA vs. ATA 그룹의 경우, 은닉노드가 2개 이상일 때부터 비슷한 패턴을 보이며 은닉노드가 증가 할수록 진폭이 작아지는 경향을 보인다. 반복 횟수는 1000번(1000epoch)까지는 정확도가 조금씩 증가하는 모습을 보이나 그 이후에는 별다른 증가를 보이지 않는다.



(그림 5) set association과 신경망을 이용한 천식 발병 예측 정확도(AIA vs. ATA)

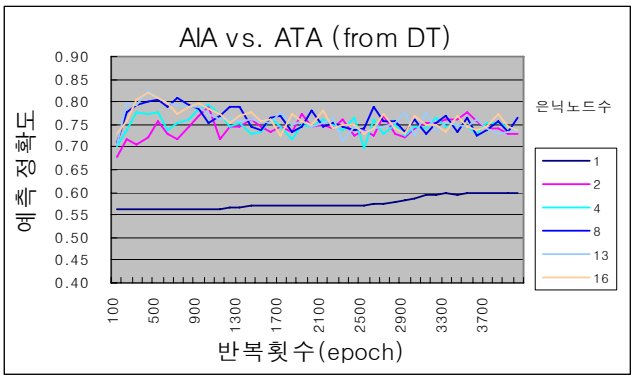
AIA vs. NC 그룹의 경우도 역시 은닉노드가 2개 이상일때부터 비슷한 정확도 패턴을 보이고 은닉노드가 증가 할수록 진폭이 작아지는 경향을 보인다. 반복 횟수가 증가할수록 오히려 정확도가 감소되는 경향을 보이는데 이는 신경망이 과학습으로 오버피팅 된 것으로 볼 수 있다. ATA vs. NC의 경우 은닉노드의 수에 따른 정확도 차이는 확연히 구분되지 않았으며 오버피팅이 반복 초기에 나타났다.

<표 1> set association과 신경망을 이용한 천식 발병 예측 최고 정확도

	은닉노드 수	반복횟수 (epoch)	최고 정확도
AIA vs. ATA	12	2400	0.6667
AIA vs. NC	2	2500	0.6082
ATA vs. NC	8	3100	0.5798

4.2. 의사결정나무와 신경망을 이용한 결과

AIA vs. ATA 그룹의 경우 반복횟수 400이나 500에서 최고의 정확도를 보인다. 따라서 학습은 적은 수의 반복 횟수와 은닉노드로 충분해 보인다.



(그림6) 의사결정나무와 신경망을 이용한 천식 발병 예측 정확도(ATA vs. ATA)

ATA vs. NC 그룹의 경우 은닉노드가 2개일 때 거의 모든 반복 횟수에서 최고의 정확도를 보였고, 반복횟수가 증가함에 따라 약간의 과학습 경향을 보였다. 그러므로 2개의 은닉노드와 적은 반복횟수 만으로도 충분한 학습이 가능할 것으로 보인다. AIA vs. NC 그룹의 경우 전반적으로 은닉노드가 많을수록 약간 더 높은 정확성을 보이고 반복횟수도 1000이상 일 때 조금 더 향상된 정확도를 보였다.

<표 2> 의사결정나무와 신경망을 이용한 천식 발병 예측 최고 정확도

	은닉노드 수	반복횟수 (epoch)	최고 정확도
AIA vs. ATA	13	400	0.825
AIA vs. NC	6	3400	0.6005
ATA vs. NC	2	3100	0.775

특별히 의사결정나무를 이용하여 중요 유전자를 택했을 때가 더 좋은 결과가 나왔는데 이는 set association에서 X 염색체 위에 있기 때문에 트리밍에서 제거했던 SNP에 의한 것으로 생각된다. Set association에서 제거된 SNP은 의사 결정 나무에서 모든 경우에 가장 중요한 SNP으로 선택되었다.

5.결론

본 연구는 set association과 의사결정나무를 이용하여 일차적으로 의미 있는 SNP들만 신경망에 입력하여 다량의 데이터에서 보다 빠르고 높은 예측 정확도를 갖는 신경망을 만드는 것이 목적이다. 실제로 천식 SNP 데이터에 이 모델을 적용해 보았을 때 모든 경우 60%이상, 최고 82%에 이르는 예측 정확도를 보였으며 이는 각각의 방법들을 독립적으로 수행하였을 때보다 좋은 결과이다. 이 모델의 또 다른 장점은 기존의 신경망을 단독으로 사용하는 방법이 어떤 입력이 중요한 입력인지 알 수 없는 블랙박스 형태인 반면, 이 모델에서는 set association이나 decision tree를 통해 선택되는 입력들을 보면서 어떤 입력이 보다 중요한지 혹은 그렇지 않은지 알 수 있다는 것이다. 또한 서로 다른 그룹 간 입력들을 비교해 보아 각 그룹에서 중요한 입력들을 찾아낼 수도 있다.

그러나 이 모델에서 다룰 수 있는 다량의 데이터란 set association이나 의사결정나무에서 처리 가능한 정도를 말하기 때문에 이용범위가 제한적일 수 있다. 그러므로 방대한 양의 데이터를 다룰 수 있기 위한 연구가 계속되어야 할 것이다.

참고문헌

[1] Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, H. Honda, “Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma”, BMC Bioinformatics 5:120, 2004.

[2] A. G. Heidema, J. M. Boer, N. Nagelkerke, E.C. Mariman, D. L. van der A, E. J. Feskens, "The Challenge for Genetic Epidemiologists : How to Analyze Large Numbers of SNPs in Relation to Complex disease", BMC Genetics 7:23, 2006.

[3] 최원일, "아스피린과 천식", http://www.tgma.org/sub/info_guide.htm?no=17847&page=1&action=read&category=&keyword=&code=.

[4]J. Hoh, A. Wile, J. Ott, "Trimming, Weighting, and Grouping SNPs in Human Case-Control Association Studies", Genome Res. 11: 2115-2119, 2001.

[5]이종태, 왕지남, "신경망 개론", 아주대학교 출판부, 2004.