DOI: 10.3745/KIPSTB.2008.15-B.4.323

집합 결합과 신경망을 이용한 복합질환의 예측

최 현 주[†] · 김 승 현[††] · 위 규 범[†††]

요 약

복합질환은 다수의 유전자들이 상호작용하여 유발되는 질병으로서, 여러 유전자들이 관여한다는 복잡성 때문에 전통적인 분석 방법을 적용하는데 한계가 있다. 최근에는 기계학습 기법을 이용한 새로운 분석 방법들이 제안되고 있다. 신경망은 이처럼 복잡한 데이터에서 일정한 패턴을 찾아 이를 분류하는데 적합한 모델이다. 그러나 다량의 데이터가 입력으로 들어오는 경우에 학습에 오랜 시간이 걸리고 패턴을 찾기가 어려워지는 단점이 있다. 본 연구에서는 다량의 SNP 데이터로부터 질병에 연관된 소수의 중요 SNP을 찾기 위한 통계학적인 방법인 집합결합(set association)과 신경망을 결합한 모델을 제시한다. 이 모델을 천식 관련 SNP 데이터에 적용하여 천식 발병 여부를 예측한 결과, 신경망만 사용했을 때보다 실행 시간도 빠르고 예측 정확도도 높았다. 이 모델은 다른 복합질환의 예측에도 효과적으로 사용할 수 있을 것으로 기대한다.

키워드 : 신경망, 집합 결합, 단일염기다형성, 복합질환, 생물정보학

A Prediction Model for Complex Diseases using Set Association & Artificial Neural Network

Hyunjoo Choi[†] · Kim Seung-Hyun[††] · Kyubum Wee[†††]

ABSTRACT

Since complex diseases are caused by interactions of multiple genes, traditional statistical methods are limited in its power to predict the onset of a complex disease. Recently new approaches using machine learning techniques are introduced. Neural nets are a suitable model to find patterns in complex data. When large amount of data are fed into a neural net, however, it takes a long time for learning and finding patterns. In this study we suggest a new model that combines the set association, which is a statistical technique to find important SNPs associated with complex diseases, and neural network. We experiment with SNP data related to asthma to test the effectiveness of our model. Our model shows higher prediction accuracy and shorter execution time than neural net only. We expect our model can be used effectively to predict the onset of other complex diseases.

Key Words : Artificial Neural Network, Set Association, Single Nucleotide Polymorphism(SNP), Complex Disease, Bioinformatics

1. 서 론

복합질환이란 다수의 유전자들이 상호작용 하여 유발하는 질병을 말한다[1,2]. 이러한 복합 질환은 여러 유전자들과 다양한 환경과의 상호작용이라는 복잡성을 갖기 때문에 전통적인 통계적 연관 분석 방법을 사용하여 질병 유전자를 찾는데 어려움이 있다. 그러므로 현재는 단일 염기 다형성(single nucleotide polymorphism; SNP) 유전자 분석과 일배체형(haplotype) 분석으로 연구의 방향이 옮겨 가고 있다. 이러한 SNP 분석 연구는 관련된 유전자의 위치를 알려줄 뿐만 아니라 기존의 연관 분석에 비해 강하게 나타나지 않는 유전적 영향을 발견하는데 유용하다[3].

SNP은 인간의 게놈에서 약 1000개 중 하나의 빈도로 나타나는 동일하지 않은 염기를 말한다. 이는 전체 게놈의 0.1%에 해당하지만 개개인을 다르게 하는 원인이 된다. SNP이 만약 엑손(exon)에 있다면 생산해 내는 단백질이 달라질 수 있고, 변형된 단백질로 인해 질병이 유발될 수도 있다. SNP이 인트론(intron)에 존재한다면 유전자 발현 조절과정에 변화가 생겨 질병이 유도될 수도 있다. 따라서 환자와 정상인의 SNP을 비교 분석하면 질병과 관련 있는 SNP을 찾아낼 수 있을 것이다.

또한 SNP 연구는 맞춤 의학이라는 질병의 치료의 관점에서도 의의가 있다. 같은 약을 복용해도 환자마다 효과가 다르다거나, 어떤 사람에게는 부작용이 일어나는 등 약물에 대한 반응의 차이를 가져오는 것도 SNP과 관련이 있다. 따라서 SNP 정보를 이용하면 개개인에게 적합한 치료가 가능

※ 이 논문은 2006년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2006-000-10775-0)
† 정 회 원 : 코리아리서치 연구원
†† 정 회 원 : 아주대학교 의과대학 연구 조교수
††† 종신회원 : 아주대학교 정보및컴퓨터공학부 교수
논문접수 : 2007년 11월 13일
수 정 일 : 1차 2008년 2월 25일, 2차 2008년 4월 25일
심사완료 : 2008년 4월 29일

해 진다[3].

이처럼 질병과 관련된 SNP을 찾는 연구가 큰 의의를 지니는 시급한 과제임에도 불구하고 데이터의 방대함과 복잡도로 인해 연구에 많은 어려움이 있다. 현재 알려진 SNP 연구 방법에는 회귀분석(logistic regression)[4], 신경망을 이용한 방법[5,6,7], multifactor dimensionality reduction (MDR)[8,9], set association[10,11], random forests[12,13] 등을 이용한 방법들이 있다[14].

신경망은 복잡한 데이터에서 일정한 패턴을 찾아 이를 분류하는 능력이 뛰어난 모델로 특별히 비선형의 복잡한 데이터에 좋은 성능을 보인다. 그러나 입력 데이터의 종류가 너무 많으면 패턴을 잘 찾지 못할 뿐만 아니라 작업에 오랜 시간이 걸리는 단점이 있다.

본 연구에서는 이를 극복하기 위해 set association과 신경망을 결합한 새로운 모델을 제시하고 이의 효율성을 입증한다.

본 논문의 구성은 다음과 같다. 제2장에서는 신경망, set association, 그리고 본 논문에서 제안하는 모델에 대하여 설명한다. 제3장에서는 제안한 모델을 사용하여 대표적인 복합질환인 천식을 유발하는데 관련된 SNP 데이터를 분석한다. 제4장에서는 분석 결과를 설명하며, 제 5 장에서는 결론을 서술한다.

2. 집합결합 전처리 신경망 모델

2.1 신경망

인공 신경망은 뇌의 구조나 동작방식을 단순화시켜 수학적으로 모델링한 계산 모형으로서, 뇌의 신경세포에 해당하는 노드(node)와 노드 간을 연결하는 에지(edge) 그리고 각 연결에 부여되는 연결강도(weight)로 구성된다. 노드는 실제적인 처리 단위로서 여러 개의 층을 형성할 수 있으며 외부로부터 입력을 받아들이는 층을 입력층, 외부로 출력을 내는 층을 출력층, 입력층과 출력층 사이의 중간층을 은닉층(hidden layer)이라고 한다.

본 실험에서는 입력층, 은닉층, 출력층으로 구성되는 전방향(feedforward) 네트워크를 사용하였다. 전방향 네트워크란 입력층, 은닉층, 출력층의 방향으로 연결되어 있고, 출력층에서 입력층으로의 직접적인 연결은 존재하지 않으며, 각 층의 노드들은 같은 층 내의 노드들과는 연결되지 않고 오직 다음 층의 모든 노드들과 연결되어 있는 네트워크를 말한다.

같은 층 내의 노드들은 기능적으로 같은 작업을 수행하며 신경망의 동작의 기본 단위가 된다. 은닉층은 신경망에 비선형의 특성을 부여하여 네트워크의 능력을 향상시키는 층으로서, 적당한 은닉 노드들만 있다면 기존의 선형 통계모형들로 풀 수 없었던 복잡한 비선형 문제들을 푸는 것이 가능하다. 그러나 적당한 은닉 노드의 수를 결정하는 것은 쉬운 문제가 아니다. 만약 너무 적은 수의 히든 노드를 사용하면 학습이 충분히 되지 않아 언더피팅(underfitting)과 높은 바이어스가 생길 수 있고, 반대로 너무 많은 히든 노드를 사용하면 트레이닝 에러(training error)는 작아질 수 있으나 과학습으로 인한 오버피팅(overfitting)과 높은 편차 때문에 일반화에 있어서 여전히 높은 에러율을 갖게 되기 때문이다[15]. 최적의 은닉노드 수를 결정하는 일반적인 방법은 알려져 있지 않다[16].

본 연구에서는 천식 발병 예측 모델에서 은닉 노드의 수를 변화시켜 가며 최적의 정확도를 찾아보았다.

2.2 학습

신경망이 입력으로 들어오는 데이터에 대해 목적에 맞는 출력을 내도록 연결 가중치를 변화시키는 과정을 학습이라고 한다. 본 실험에서는 감독 학습(supervised learning) 방법 중 하나로 가장 많이 이용되는 오류 역전파 알고리즘을 이용하였다. 오류 역전파(back propagation) 알고리즘은 경사 하강법(gradient descent)과 미분의 연쇄규칙(chain rule)을 적용하여 오류를 역으로 전파하는 방법으로서, 전방향으로 출력값을 구하는 과정과 출력값과 교사값(teacher value)을 비교하여 오차를 구하고 이를 역으로 전파하여 연결강도를 변화시켜 그 차이를 감소시키는 과정으로 나뉜다.

오류 역전파 알고리즘은 신경망에서 가장 많이 이용되는 학습방법이지만 몇 가지 문제점을 가지고 있다. 먼저 이 방법이 기울기가 큰 경사면을 따라가는 경사 하강법을 사용하기 때문에 오차가 0이 아닌 지역 최소점에 머무를 수 있다는 점이다. 곧 전역적 최소점이 아니라 미분값이 0인 지역 최소점에 머무를 위험이 있다[15, 16]. 이를 막기 위한 방법 중 하나가 은닉 노드의 수를 증가시키는 것인데 본 연구에서는 은닉 노드의 수를 변화시키면서 실험을 행하므로 이를 피할 수 있을 것이다. 또 다른 문제는 학습이 완료되기까지 많은 횟수의 반복 학습이 필요하며 학습의 완료 시점을 예측할 수 없다는 것이다. 이를 해결하기 위해 반복 횟수 또한 변화시켜 가면서 여러 번 실험을 행하였다.

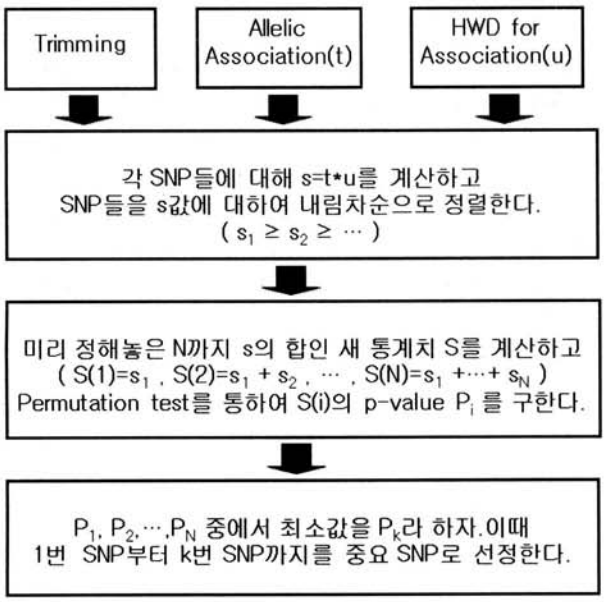
2.3 Set association

Set association은 다량의 SNP 데이터로부터 질병에 연관된 소수의 중요 SNP을 찾기 위한 통계학적인 방법으로서 2002년 Ott와 Hoh에 의하여 제안되었다[10,11]. 이를 위해서는 각 SNP에 대한 allelic association (AA) 통계치와 Hardy-Weinberg Disequilibrium(HWD) 통계치가 필요하다. AA 통계치란 각 SNP에 대하여 SNP와 질병유무(disease outcome)과의 관계 정도를 측정한 χ^2-통계치이고, HWD 통계치란 Hardy-Weinberg equilibrium을 귀무가설로 하여 각 SNP의 분산 정도를 측정한 χ^2-통계치이다.

환자 그룹에서 HWD 편차가 높은 SNP들은 질병과 관련이 있다는 의미이다. 정상인 그룹에서 HWD 편차가 높은 SNP들은 genotyping error를 의미하므로 해당 SNP을 제거하거나 HWD값을 0으로 맞춰 주는 트리밍(trimming) 작업이 필요하다.

AA 통계치와 HWD 통계치를 곱하여 새로운 통계치를 만들고 이 값에 따라 내림차순 정렬을 한 후, 정렬되어 있

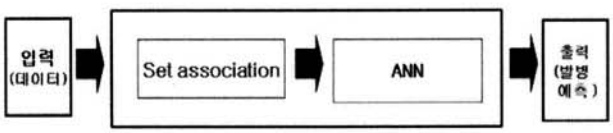
는 통계치들을 갖는 SNP들을 순서대로 하나씩 더해가며 미리 정해진 개수 N만큼의 SNP들의 집합들을 만들어 간다. Permutation test로 각 집합의 p-value를 구하고, p-value가 minimum이 될 때의 SNP 집합을 질병과 관련있는 SNP 집합으로 선택한다. 자세한 과정은 (그림 1)와 같다.



(그림 1) Set association 방법의 요약

2.4 Set association과 신경망을 결합한 모델(SA+NN)

복합질환을 유발하는 데 연관된 것으로 의심되는 다수의 SNP 중에서 일차적으로 set association을 이용하여 소수의 중요 유전자를 선별한 후, 선별된 SNP들만을 가지고 신경망을 학습시켜 이들 SNP들의 유전형(genotype)에 의해서 복합질환 발병을 예측한다 (그림 2 참조).



(그림 2) set association과 신경망을 결합한 모델

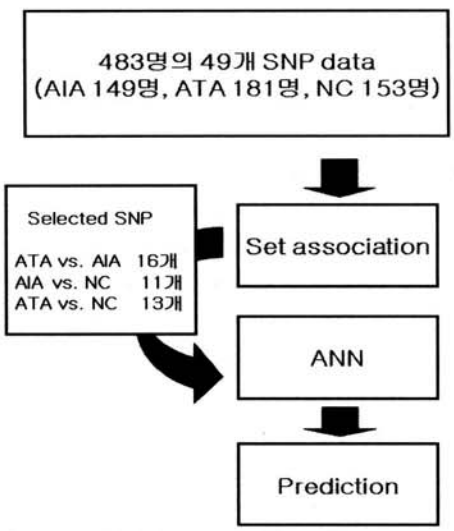
본 연구에서 제시하는 모델의 유효성을 입증하기 위하여 대표적인 복합질환인 천식을 유발하는데 연관된 SNP 데이터를 분석하였다. 실험은 아스피린 복용으로 유도되는 아스피린 과민성 천식(AIA), 아스피린 내인성 천식 (ATA), 정상인(NC)의 세 그룹으로 나누어 수행하였다. 천식 환자군을 두 그룹으로 나눈 이유는 AIA와 ATA 사이의 특징적인 SNP을 찾으려 함이다. 실제 임상에서는 AIA인지 ATA인지에 따라 각기 다른 치료가 행해지는데, 병력만으로 이러한 진단을 내리기 힘든 경우가 많다. 또한 천식 발작을 전후하여 진통 소염제를 많이 복용해 아스피린과의 직접적인 인과관계를 알아내기 힘들기 때문에 약물 유전체학(pharmacogenetics) 관점에서 빠르고 정확하게 이 둘을 구분할 수 있는 SNP을 찾으려는 것이 본 연구의 또 다른 목적이다[3].

3. 실 험

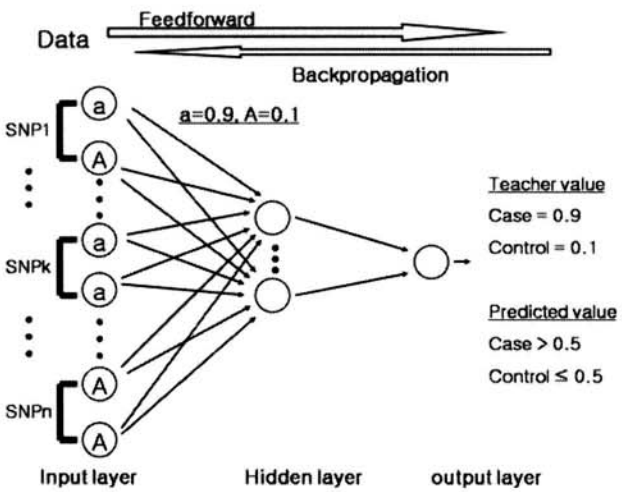
실험은 49개 SNP에 대한 483명(AIA 149명, ATA 181명, NC 153명)의 데이터를 AIA vs. ATA, AIA vs. NC, ATA vs. NC의 세 그룹으로 나누어 진행하였다. 실험에 사용된 SNP 데이터는 본 연구팀의 소속 대학 병원으로부터 제공받은 실제 데이터이다.

(그림 3)에서와 같이 각 그룹에서 일차적으로 set association을 이용하여 중요한 SNP들을 선별하였다. 일차적으로 set association을 이용한 경우에는 AIA vs. ATA 그룹의 경우 16개, AIA vs. NC 그룹의 경우 11개, ATA vs. NC의 경우 13개의 SNP이 중요 SNP으로 선별되었다. 선별된 SNP들을 비교해 본 결과 그룹간 중요 SNP들이 다름을 알 수 있었다. 선별된 중요 SNP들을 신경망의 입력으로 설정하였다.

신경망은 (그림 4)와 같이 전방향(feedforward) 네트워크를 사용하였고 학습에는 오류 역전파 알고리즘을 사용하였다.



(그림 3) 신경망 기반 복합질환 발병 예측 실험



(그림 4) 실험에 사용된 신경망 모델

신경망에 입력 값을 줄 때 각 SNP의 유전형(genotype)에 따라 다수의 대립 유전자(major allele)에는 0.1, 소수의 대립 유전자 (minor allele)에는 0.9의 입력 값을 준다.

예를 들어서, (그림 4)에서 SNP_1의 major allele이 C이고 minor allele이 A 이라면, SNP1의 유전형이 CC인 경우에는 SNP_1을 인코딩하는 두 개의 입력노드에 각각 0.1이 입력된다. 유전형이 CA이면 0.1, 0.9가 입력되며, 유전형이 AA이면 0.9, 0.9가 입력된다.

출력값은 control(정상)은 0.1, case(환자)는 0.9의 값을 주어서 학습하며, 예측 시에는 출력값이 0.5 이하이면 정상, 0.5보다 크면 환자로 분류한다 (그림 4 참조).

은닉노드는 1개에서 16개까지, 반복횟수(epoch)는 100번에서 2000번까지 변화시켜가며 신경망을 구성하였다. 은닉노드 수와 반복 횟수를 고정한 각 경우에 10-fold 교차확인(cross validation)을 통해 민감도(sensitivity), 특이도(specificity), 정확도(accuracy)를 측정하였다.

실험에 사용한 신경망은 C언어로, set association은 java로 작성하였고 linux기반의 intel xeon 2.8GHz CPU와 2G 메모리의 환경에서 실험하였다.

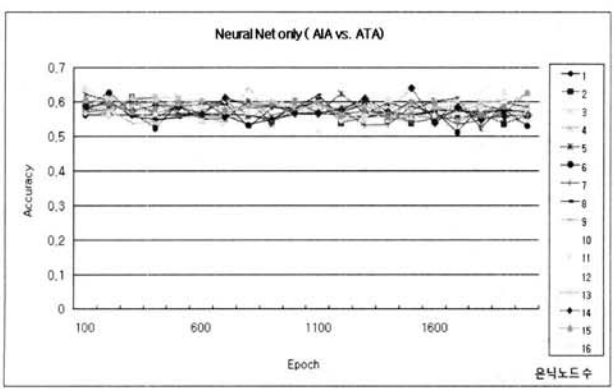
4. 결 과

각 시행 방법은 정확도(accuracy)와 실행 시간을 기준으로 평가된다. 정확도는 10-fold 교차확인을 통해 얻은 결과들을 평균 내어 얻은 평균정확도이다. 실행 시간은 은닉노드를 1개에서 16개까지, 반복횟수(epoch)를 100번에서 2000번까지 변화시켜가며 신경망을 구성하면서 발병 예측 정확도를 측정하는데 걸린 초 단위 시간이다.

4.1 신경망만을 사용한 경우(NN-only)

49개의 전체 SNP을 신경망에 입력 값으로 넣고 천식 발병 여부를 예측해보았다. AIA vs. ATA, AIA vs. NC, ATA vs. AIA의 모든 경우, 예측 정확도의 그래프 패턴은 일정한 값에 수렴하지 않고 진동하는 모습을 보였다.

(그림 5)는 AIA vs.ATA 그룹의 예측 정확도 그래프이다. 반복횟수에 따라 전체적으로 상승하거나 하강하는 모습은 보이지 않으며 진동하는 모습만이 확인된다. 은닉노드가 14개이고 반복횟수가 1500 epoch일 때 0.6406의 가장 높은 예측 정확도를 보였으며 다른 두 그룹에 비해 전체적으로 높은 예측 정확도를 보였다.

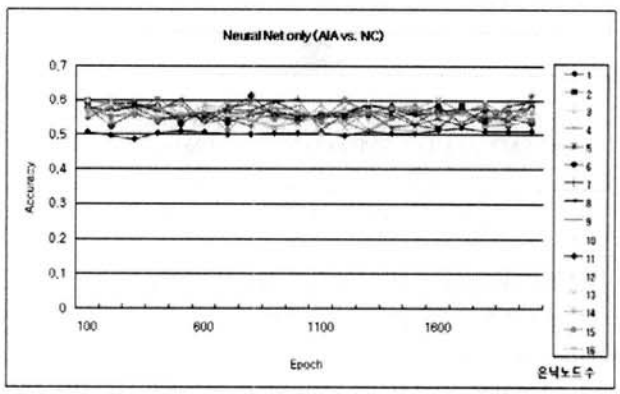


(그림 5) AIA vs. ATA 그룹에 신경망을 적용한 경우의 예측 정확도

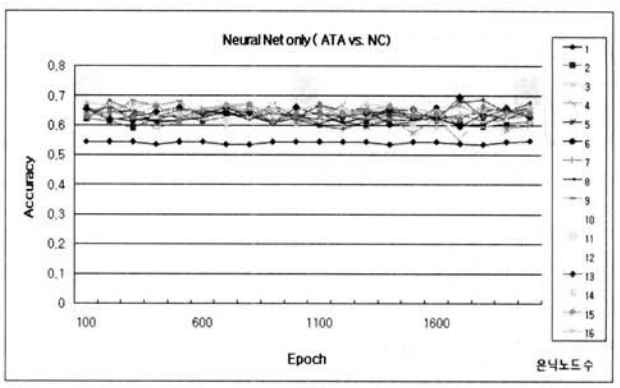
AIA vs. NC 그룹의 경우 (그림 6)과 같이 은닉노드가 1개일 때는 0.5 정도의 정확도를 보이다가 은닉노드가 2개일 때부터 진동하는 모습을 보였다. 최고 정확도는 은닉노드가 11개 반복횟수가 800 epoch일 때 0.6138로 나타났다.

ATA vs. NC 그룹의 예측 정확도는 (그림 7)과 같다. 은닉노드가 1개일 때 AIA vs. NC의 경우보다 조금 높은 0.55 정도의 정확도를 보였다. 은닉노드가 2개 이상일 때부터 진동하는 패턴이 보이는데 은닉노드 수에 따른 예측 정확도 증가 모습은 확연히 구분하기 힘들다. 가장 높은 정확도는 0.6970으로 은닉노드가 13개 반복횟수가 1700 epoch일 때 나타났다.

각 그룹에서 최고 예측 정확도는 <표 1>과 같다. AIA



(그림 6) AIA vs. NC 그룹에 신경망을 적용한 경우의 예측 정확도



(그림 7) ATA vs. NC 그룹에 신경망을 적용한 경우의 예측 정확도

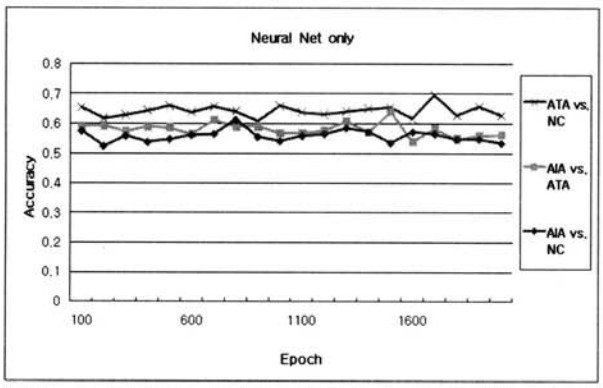
〈표 1〉 전체 SNP을 신경망에 입력한 경우의 예측 정확도

	은닉 노드 수	반복횟수 (epoch)	정확도 (%)
AIA vs. ATA	14	1500	64.06
AIA vs. NC	11	800	61.38
ATA vs. NC	13	1700	69.70

vs. ATA의 경우 약 64.06%의 정확도를 보였고 AIA vs. NC의 경우 약 61.38의 정확도를 보였으며 ATA vs. NC의 경우 69.70%의 예측 정확도를 보였다.

전체 SNP을 신경망의 입력으로 하고 천식 발병 여부를 예측한 경우에 각 그룹 간 정확도를 비교해 보았다.(그림 8 참조)

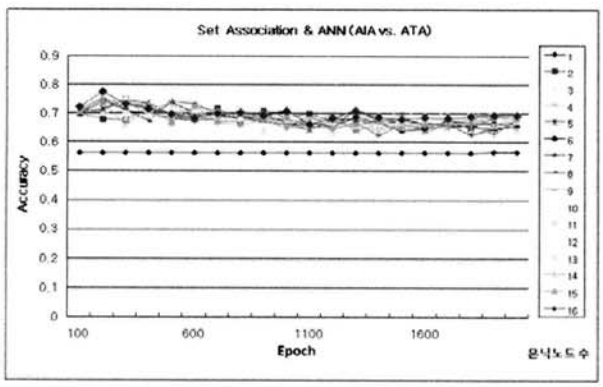
은닉노드 수를 최고 정확도를 나타낼 때와 같은 수로 고정하고 반복횟수를 변화시켜가며 각 그룹 간 정확도를 비교를 해보았다. (그림 8)에서 보면 반복횟수에 상관없이 최고 정확도가 높은 그룹이 평균적으로 높은 정확도를 나타내고 있음을 알 수있다. 즉, ATA vs. NC, AIA vs. ATA, AIA vs. NC 순으로 높은 예측 정확도를 보였다.



(그림 8) 전체 SNP을 신경망에 입력된 경우 예측 정확도의 그룹 간 비교

4.2 Set association과 신경망을 결합한 모델을 사용한 경우 (SA+NN)

(그림 9)는 AIA vs. ATA 그룹에서 예측 정확도 그래프이다. 예측 정확도는 은닉노드가 1개 일 때 0.5625로 반복횟수에 관계없이 일정하였다. 은닉노드가 2개 이상일 때부터 그래프가 진동하는 모습이 나타났는데 은닉노드 수가 증가할수록, 반복횟수가 증가할수록 진폭이 작아지는 경향을 보였다. 반복횟수가 증가하면서 오히려 전체적으로 정확도가 약간 감소하는 경향을 보이는데 이는 신경망이 오버피팅(overfitting)된 것으로 볼 수 있다. 오버피팅이란 주어진 학습 데이터에 대해서는 에러를 최소화 시키는 해를 찾았지만 테스트 데이터에 대해서는 추론 기능이 별로 우수하지 못하게 되는 현상이다. 은닉노드가 16개이고 반복회수가 200 epoch일 때 0.775의 가장 높은 정확도를 보였다.

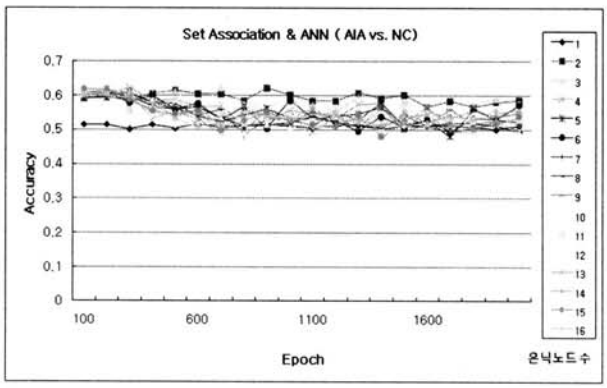


(그림 9) AIA vs. ATA 그룹에 set association과 신경망을 이용한 경우의 예측 정확도

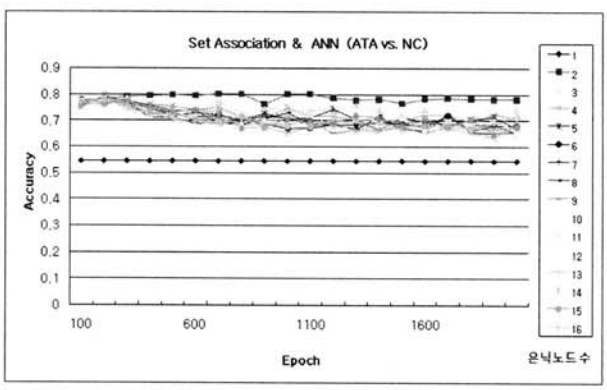
AIA vs. NC 그룹의 경우는 은닉노드가 1개일 때부터 진동하는 패턴을 보였는데 여기서도 은닉노드 수가 많아질수록 오버피팅 되는 경향을 확인할 수 있었다. 은닉노드 수가 2개일 때 평균적으로 가장 높은 정확도를 보였으나 최고 정확도는 은닉노드가 4개이고 반복횟수가 300 epoch일 때 0.6207였다. 따라서 AIA와 NC를 구분하는 데에는 상대적으로 적은 수의 은닉노드가 필요하다고 할 수 있겠다. 예측 정확도 그래프는 (그림 10)과 같다.

ATA vs. NC 그룹에서는 오버피팅 모습을 확실히 관찰할 수 있다. (그림 11)을 보면 은닉노드가 2개일 때 거의 모든 경우 최고의 정확도가 나타나고 있다. 최고 정확도도 은닉노드가 2개이고 반복횟수가 700 epoch일 때 0.8030이었다. 또한 반복횟수가 많아질수록 정확도가 낮아지는 경향을 보이므로, 오버피팅을 막고 높은 정확도를 얻기 위해서는 적은 수의 은닉노드와 반복횟수가 필요할 것으로 보인다.

Set association을 이용하여 중요 SNP를 선별하고 이것을 신경망의 입력으로 하는 방법에서는 AIA vs. ATA 경우 약 77.5%, AIA vs. NC의 경우 약 62.07%, ATA vs. NC의 경



(그림 10) AIA vs. NC 그룹에 set association과 신경망을 이용한 경우의 예측 정확도



(그림 11) ATA vs. NC 그룹에 set association과 신경망을 이용한 경우의 예측 정확도

우 약 80.30%의 높은 예측 정확도를 보였다. <표 2>는 정확도가 가장 높은 경우를 보여준다.

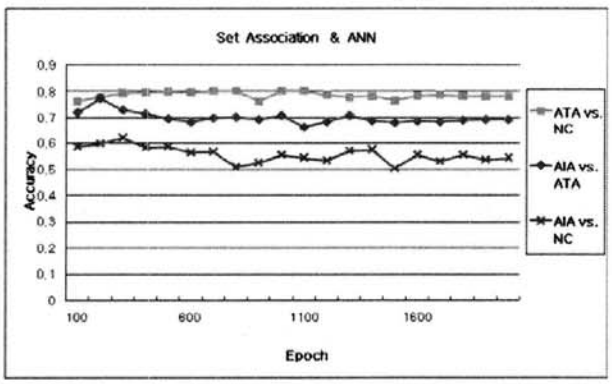
은닉노드 수를 최고 정확도를 나타낼 때와 같은 수로 고정하고 반복횟수를 변화시켜가며 각 그룹 간 정확도를 비교해 보았다. 일차적으로 set association을 사용한 경우에도 반복횟수에 상관없이 최고 정확도를 갖는 그룹 순으로 높은 정확도를 나타내었다. ATA vs. NC, AIA vs. ATA, AIA vs. NC 순으로 높은 예측 정확도를 보임을 (그림 12)에서 확인할 수 있다.

정확도와 실행 시간을 기준으로 각각의 방법의 성능을 비교해 보았다(<표 3> 참조). 전체 SNP을 신경망의 입력으로 한 경우에는 세 그룹에서 모두 60%대의 정확도를 보였으나, set association을 이용하여 중요 SNP을 선별한 후에 신경망에 입력한 경우에는 이보다 높은 60~80%의 정확도를 보였다. 일차적으로 중요한 SNP들을 선별해 이들을 신경망의 입력으로 하여 예측하는 방법이 전체 SNP들을 이용하여 예측하는 방법보다 모든 그룹에서 높은 예측도를 보였다.

실행시간에서도 set association을 이용하는 방법이 효율적임을 확인할 수 있었다. <표 3>을 보면 set association 방법을 이용할 경우 최대 약 26000초의 실행시간이 걸리지만 신경망만을 사용할 경우 최대 76000초 이상의 시간이 걸림을 알 수 있다. Set association을 통해 중요 SNP을 선별하는 데에 약 250초의 시간이 걸렸음을 감안하더라도, 일차적으로 주요 SNP들을 통해 발병을 예측하는 것이 3배 이상 효율적임을 확인할 수 있었다. 이러한 실행시간 단축 효과는 신경망에 입력되는 SNP의 수가 많을수록 증대될 것이다.

〈표 2〉 Set association과 신경망을 이용한 예측 정확도

	은닉 노드 수	반복횟수 (epoch)	정확도 (%)
AIA vs. ATA	16	200	77.50
AIA vs. NC	4	300	62.07
ATA vs. NC	2	700	80.30



(그림 12) Set association과 신경망을 이용한 예측 정확도의 그룹 간 비교

〈표 3〉 두 가지 방법의 성능비교

	ANN only		Set association & ANN	
	정확도 (%)	실행시간 (단위: 초)	정확도 (%)	실행시간 (단위: 초)
AIA vs. ATA	64.06	76609	77.50	26043
AIA vs. NC	61.38	64177	62.07	18289
ATA vs. NC	69.70	74256	80.30	22559

4.3 NN-only 와 SA+NN 의 성능 비교

제안한 모델의 효율성을 제1장에서 소개한 multifactor dimensionality reduction(MDR)과 비교하였다[7, 8]. MDR은 SNP과 질병 연관성 연구에 널리 사용되고 있으며 공개 소프트웨어로 구현되어 있다[17].

MDR은 SNP 데이터 중에서 missing value들 처리하는 방법이 유연하지 못하므로, 데이터를 이루는 49 개의 SNP 중에서 특히 missing value가 많은 3 개의 SNP을 제외하고 46개의 SNP을 사용하였다. 그 이외에 간헐적으로 발생하는 missing value들은 해당 SNP이 가장 흔하게 가지는 genotype으로 값을 채워 넣었다. ANN에서는 missing genotype을 (0.5, 0.5)로 입력하였다.

Set association & ANN 과 MDR의 성능 비교 결과는 <표 4>와 같다.

MDR은 주어진 n개의 SNP 중에서 모든 1-combination, 2-combination(pair), 3-combination(triple), …, n-combination을 다 검사한다. 그러므로 SNP이 개수가 많으면 n-combination까지 다 계산하는 것은 실제로 컴퓨터 실행시간과 메모리의 제약 때문에 불가능하다. <표 4>의 결과는 실험 환경이 허용하는 최대 범위인 4-combination(quadruple) 까지 계산한 결과를 보인 것이다. 더 좋은 환경에서 실험한다면 5-comb. 혹은 6-comb. 까지 계산 가능하겠지만 정확도가 대폭 증가하지는 않을 것으로 생각된다.

〈표 4〉 MDR vs. Set Association & ANN

	MDR		Set association & ANN	
	정확도 (%)	실행시간 (단위: 초)	정확도 (%)	실행시간 (단위: 초)
AIA vs. ATA	61.66	675	77.50	26043
AIA vs. NC	52.98	672	62.07	18289
ATA vs. NC	54.79	665	80.30	22559

5. 결 론

복합질환의 예측을 위하여 set association과 신경망을 결합한 새로운 모델을 제시하였다. 신경망은 패턴분류 능력이 뛰어난 모델이지만 입력 노드 수가 너무 많으면 패턴을 잘 찾지 못할 뿐만 아니라 오랜 실행시간이 걸린다는 단점이 있다. 본 연구에서는 set association을 이용하여 일차적으로 의미 있는 SNP들만 신경망에 입력하였다.

이러한 결과 다량의 데이터에서 보다 빠른 실행시간과 높은 예측 정확도를 갖는 신경망을 구성할 수 있었다. 실제 천식 SNP 데이터에 이 모델을 적용해 보았을 때 AIA vs.

ATA, AIA vs. NC, ATA vs. NC 세 경우에 모두 60%이상, 최고 약 83%에 이르는 예측 정확도를 보였으며 실행시간도 3배 이상 단축되었다.

제안하는 모델의 또 다른 장점은, 기존의 신경망을 단독으로 사용하는 방법은 어떤 입력이 중요한 입력인지 알 수 없는 블랙박스 형태인 반면에, 이 모델에서는 set association을 통해 선택되는 입력들을 보면 어떤 입력이 보다 중요한지 혹은 그렇지 않은지 알 수 있다는 것이다. 실험에서 set association에 의해서 선별된 주요 SNP들은 실제로 AIA, ATA, NC의 세 그룹을 구분 가능하게 해주는 특징적인 SNP들이라 생각된다.

본 연구에서 제시하는 set association과 신경망을 결합한 모델은 천식 이외의 다른 복합질환의 예측에도 좋은 성능을 보일 것으로 기대한다.

참 고 문 헌

[1] J. Y. Dai, I. Ruczinski, M. LeBlanc and C. Kooperberg, "Imputation methods to improve inference in SNP association studies," Genetic Epidemiology, Vol.30, pp.690-702, 2006.

[2] D. Bostein, N. Risch, "Discovering genotypes underlying human phenotypes: past success for Mendelian disease, future approaches for complex disease," Nat. Genet. Suppl. Vol.33, pp.228-237, 2003.

[3] 임성빈, "약물유전체학(Phamacogenomics)", 월드사이언스, 2004.

[4] N. Nagelkerke, J. Smits, S. Le Cessie, H. Van Houwelingen, "Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting," Stat. Med. Vol.24, pp.121-130, 2005.

[5] Y. Tomita, S. Tomida,Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi and H. Homita, "Artificial neural network approach for selection of susceptible single nucleotide polymorphism and construction of prediction model on childhood allergic asthma," Bioinformatics, Vol.5, pp.120-132, 2004.

[6] M. D. Ritchie, B. C. White, J. S. Parker, L. W. Hahn and J. H. Moore, "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human disease," BMC Bioinformatics, Vol.4, pp.28-42, 2003.

[7] A. A. Motsinger, S. L. Lee, G. Mellick and M. D. Ritchie, "GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease," BMC Bioinformatics, Vol.7, pp.39-49, 2006.

[8] L. W. Hahn, M.D. Ritchie and J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions," Bioninformatics, Vol.19, No.3, pp.376-382, 2003.

[9] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney and B. C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of numan disease susceptibility," J. Theor. Biol. Vol.241, pp.252-261, 2006.

[10] J. Hoh, A. Wille and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," Genome Res. Vol.11, pp.2115-2119, 2001.

[11] J. Ott and J. Hoh, "Set association analysis of SNP case-control and microarray data", J. Comput. Biol. Vol.10, pp.569-574, 2003.

[12] K. L. Lunetta, L. B. Hayward, J. Segal and P. Van Eerdewgh, "Screening large-scale association study data: exploiting interactions using random forests," BMC Genetics, Vol.5, pp. 32-45, 2004.

[13] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith and P. Van Eerdewegh, "Identifying SNPs predictive of phenotype using random forests," Genet. Epidemiol. Vol.28, pp.171-182, 2005.

[14] A. G. Heidema, J. M. Boer, N. Nagelkerke, E.C. Mariman and 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, Vol.7, pp.23-38, 2006.

[15] S. Kumar, "Neural Networks: A Classroom Approach," McGraw Hill, 2004.

[16] R. Rojas, "Neural Networks: A Systematic Introduction," Springer, 1991.

[17] http://sourceforge.net/projects/mdr/

최 현 주

e-mail : yoyo194@kric.com
2005년 아주대학교 생명과학과 (학사)
2008년 아주대학교 컴퓨터공학과 (석사)
2008년~현 재 코리아리서치 연구원
관심분야 : 생물정보학

김 승 현

e-mail : kimsh@ajou.ac.kr
1990년 숙명여자대학교 화학과 (학사)
1993년 서울대학교 화학과 (이학석사)
1996년 서울대학교 화학과 (이학박사)
2003년~2006년 아주대의대 연구 전임강사
2007년~현 재 아주대의대 연구 조교수
관심분야 : 정보처리, 데이터베이스 등

위 규 범

e-mail : kbwee@ajou.kr

1978년 서울대학교 수학과 (학사)

1985년 University of Wisconsin 전산학과 (석사)

1992년 Indiana University 전산학과 (박사)

1993년~현 재 아주대학교 정보및컴퓨터 공학부 교수

관심분야 : 컴퓨팅이론, 생물정보학