하이브리드 모형을 이용한 예측률 향상에 관한 연구

연세대학교 대학원
의학전산통계학 협동과정
의학전산통계학 전공
김 봉 섭

하이브리드 모형을 이용한 예측률 향상에 관한 연구

지도 변 혜 란 교수

이 논문을 석사 학위논문으로 제출함

2003년 12월 일

연세대학교 대학원
의학전산통계학 협동과정
의학전산통계학 전공
김 봉 섭

김봉섭의 석사학위 논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2003년 12월 일

감사의 글

입학한지 엊그제 같은데 졸업의 시간이 다가오는 저의 현 위치에 아직 실감이 나질 않습니다. 대학원에서의 새로운 환경이 처음엔 낫설었지만 지금 이 시점에서는 어느덧 그러한 환경도 돌아올수 없는 시간이기에 그리움이 남습니다.
논문의 길을 인도하고자 부족한 저에게 조언과 충고를 전해주시며 성심껏 도와주신 변혜란 교수님께 진심으로 감사를 드립니다.
조교생활을 졸업할때까지 하지 못하고 나온 점에 대해 미안한 감을 가지고 있으며 불편하신 몸에도 저의 논문에 배려와 성원을 보내주신 김동기 교수님께 진심으로 감사의 성의를 보내고자 합니다.
멀리서도 자상함과 너그러움으로 보살펴 주시며 논문에 도움을 주신 김동건 교수님께 진심으로 감사의 메시지를 전하고 싶습니다.

어렸을적부터 지금까지 저에게 모든 정성을 마다하지 않으며 보살펴주며 언제나 저를 지켜주신 부모님께 감사하며 사랑한다고 말하고 싶습니다. 바쁜 가운데도 어머니의 옆에서 도와주며 함께한 여동생, 군 입대한지 엊그제 같은데 건강하게 제대해서 예전보다 늠름해진 막내동생, 학창생활을 보내며 숙제와 고민 등을 함께하며 즐거웠던 시간을 보냈던 원열, 자기 생활에 충실하며 남에게 호의를 베풀줄 아는 장섭형, 학과의 기둥이 되며 후배들을 인도하는 기준형, 말수는 적지만 본인의 임무에 충실한 성민형, 풋풋함으로 다가서려는 무영, 이번에 겨울 신부가 된 찬미씨, 예비 박사의 길을 걸어가고 있는 미영씨, 적은 시간이었지만 같이 수업을 들었던 민지씨, 후배로서 좀더 챙겨주지 못한 정숙, 많은 애기를 하진 못했지만 학과 업무에 충실해 보였던 수옥, 은혜, 신영, 혜리, 모든 사람들에게 지금까지 있었던 일들에 아쉬움을 털고 좋은 시간만을 기억해주었음 합니다.

차 례

그림 차례

표 차례

국 문 요 약

하이브리드 모형을 이용한 예측률 향상에 관한 연구

본 논문에서는 분류 예측력을 향상시키기 위해서 기존에 제시된 단일 모형과 본 연구에서 제시된 하이브리드 Decision Tree-Neural Network 모형과 하이브리드 Genetic-Neural Network 모형을 제시함으로써 개별기법의 단점을 다른 기법으로 보완하여 분류율을 높이는데 관점이 있다.

본 연구에서 제시하는 혼합모형은 두 가지 관점에서 바라보고 있다. 첫째, 혼합모형Ⅰ은 의사결정나무의 알고리즘은 일반적으로 Small Disjunct가 아닌 Large Disjunct에 잘 적합하도록 치우치는 경향이 있다. 순수도가 낮고 노드에 속한 개체수가 적은 Small Disjunct에서 적중률이 낮아지는 경향이 있으므로 이러한 노드들에 대한 적절한 처리를 통하여 적중률을 높이고자 신경망 모형으로 보완하고자 하였다. 둘째, 혼합모형Ⅱ은 신경망 모형의 단점인 초기값이 지역적 최소값에 가까우면 지역적 최소값 에서의 모수 추정치는 실제 데이터를 정확하게 추정하지 못할 가능성이 많다는 단점을 유전자 알고리즘으로 보완하고자 하였다.

1994년~2002년까지 연세대학교 의과대학 세브란스 병원 건강증진센터에서 안과정밀 검사를 받아온 환자 실제자료를 통해 단일모형인 의사결정나무, 신경망과 하이브리드 Decision Tree-Neural Network 모형, 하이브리드 Genetic-Neural Network 모형을 시뮬레이션 한 결과 하이브리드 Decision Tree-Neural Network 모형과 하이브리드 Genetic-Neural Network 모형이 단일모형에 비해 분류 예측력을 2%정도 향상 시킬수 있었다.

핵심되는 말 : 로지스틱 회귀분석, 의사결정나무, 신경망, 유전자 알고리즘, 혼합모형, 데이터 마이닝

제 1 장 서 론

1. 1 연구 배경

최근 의료의 의미가 과거 단순한 질병의 치료에서 건강증진이나 질병의 예방으로 변화하면서 소비자의 요구에 맞는 건강검진사업은 중요한 의료서비스의 하나로 인식되고 있으며, 각 병원은 건강검진센터를 운영하면서 수동적으로 건강검진을 받으러 오는 것을 기다리기 보다 능동적으로 환자에서 만족감을 줄 수 있는 프로그램 개발이 필요하게 되었고, 건강검진 수진자의 증가가 입원, 외래 환자의 이용증가로 이어진다는 것은 건강검진센터의 운영의 필요성을 말해주고 있다.

실제로 많은 의료기관들이 종합건강검진센터를 운영 하고 있으며 개설되지 않은 의료기관들도 많은 관심을 가지고 있는 현실이다. 또한 국민소득의 증가와 이에 따른 생활환경의 변화, 인구의 증가 및 연령구조의 변화, 의료기술의 발달, 의료문화 및 의료수가의 적용범위의 변화, 국가의 보건정책 등이 맞물려 의료수요의 증가가 팽창하게 될 것이고 그에 상응하는 건강검진 및 건강증진 프로그램 등이 새로이 개발될 것으로 예상된다.

또한 WTO의 출현에 따라 막강한 자본력과 첨단 의료기술 그리고 고객중심주의적 마케팅력을 갖춘 외국의 병원들이 들어올 것으로 진단되며, 국내 굴지의 대기업들도 병원업계에 진출하여 이제는 병원도 무한경쟁시대에 접어들었다고 할 수 있다. 대부분의 병원들이 직면하고 있는 경쟁상황의 증가, 병원경영의 악화, 환자들의 보다 향상된 의료서비스의 요구 등은 비영리기관으로서의 병원경영활동에 한계를 드러내고 있는 것이라 하겠다.

이와 같은 환경의 변화로 병원에서의 마케팅활동은 생존을 위한 중요한 요소로 대두되고 있다.

병원은 정보시스템의 발전과 더불어 많은 양의 데이터가 축적되면서 필요한 정보를 찾아내어 의료정보자원의 효율적인 활용을 위한 지식경영의 비전 및 전략수립이 요구되고 있다. 과거 병원의 경영전략 수립은 정확한 정보에 근거하기 보다는 경영자의 경험에 근거하여 결정되는 경우가 많았다. 그러나 이제는 병원의 규모가 커지고 주변환경이 급변함에 따라 정확한 정보에 근거한 경영전략인 대량의 데이터를 지식으로 효과적으로 저장, 관리, 활용할 수 있는 지식 탐사 방법인 데이터 마이닝에 대한 중요도가 증대되고 있다. 많은 병원에서 전산화가 구현됨에 따라 챠트에 보관되었던 진료정보가 컴퓨터에 보관됨에 따라 데이터를 최신 정보기술인 데이터 마이닝 등의 기법을 이용하여 병원운영에 관한 정보를 얻을 수 있어서 정보기술을 활용하는 병원경영전략의 수립이 가능하게 되었다.

선진 병원에서는 진료 정보의 효율적 활용을 위한 데이터웨어하우스, 데이터마트, 데이터마이닝등의 정보기술을 기반으로 진료정보를 효율적으로 활용하고 있으나 우리나라에서는 대부분 진료정보 전산화로 구축된 데이터를 단순업무에만 활용하고 있는 실정이다.

이런 시점에서 병원의 처한 상황이나 시대의 변화에 따라 모든 원인변수를 정확히 파악하고 고객의 평생가치를 높이는데 기존의 단순한 통계적인 기법보다는 하이브리드 모형을 적용함으로 고객의 욕구변화에 적응할 수 있는 분류율 향상에 대안이 되리라 생각된다.

1. 2 연구 목적

본 연구에서는 종합건강진단센터 안과 데이터를 효과적으로 분류하기 위하여 건강 진단센터의 자료를 토대로 백내장에 걸릴 위험소지가 있는 환자들을 분류함에 있어 기존의 통계적인 모델인 의사결정나무, 신경망과 동시에 하이브리드 Decision Tree-Neural Network 모델과 하이브리드 Genetic-Neural Network 모델을 형성함으로써 보다 정확한 분류율과 예측률 향상에 도움이 될 수 있는지에 그 목적이 있다.

본 논문은 새로운 Small Disjunct 방법을 이용한 하이브리드 Decision Tree-Neural Network 모델을 사용함으로써 차별화된 접근방법 이라 언급할수 있겠다. 여기서 Small Disjunct란 의사결정나무 모형에서 나온 노드의 데이터개수가 훈련용 데이터(Training Set)의 10% 미만이면서 순수도가 낮은 모형으로 정의하였다. 혼합모형Ⅰ인 하이브리드 Decision Tree-Neural Network 모델은 의사결정나무에서 Small Disjunct로 판단되는 노드만을 신경망의 훈련용 데이터로 사용함과 동시에 의사결정나무와 신경망의 결과를 종합함으로서 두개의 모형을 결합하였다.

혼합모형Ⅱ인 하이브리드 Genetic-Neural Network 모델은 먼저 신경망 모형을 구축하고 유전자 알고리즘에서 구한 최적의 가중치를 대입함으로 두개의 모형을 결합하였다.

백내장에 걸릴 위험의 소지가 있는 환자들을 연구하기 위하여 1994년 ~ 2002년까지 연세대학교 의과대학 세브란스 병원 건강증진센터에서 검사를 받아온 환자 62,593개의 안과 자료를 토대로 분석한다. 로지스틱 회귀분석, 의사결정나무, 신경망에 관련된 이론을 정리하고 통계적인 모델과 하이브리드 모델 구축에 활용되는 기법을 SAS, E-MINER, NEUROSHELL 2.0, Evolver 4.0 등을 이용하였다.

1. 3 혼합모형에 대한 기존 연구

의사결정나무분석에서 발견된 지역적인 패턴과 로지스틱 회귀분석에 발견된 전체적인 패턴을 고려한 혼합나무-로짓모형이 각각의 단일모형에 비해 분류 예측력이 높은 것으로 알려져 있다[26].

단일의 의사결정나무 분석을 결합한 복수 의사결정나무 분석이 단일의 의사결정나무모형보다 분류 예측력이 향상되었다고 주장하였다[22].

퍼지시스템과 신경망을 결합시킨 주식 예측 시스템을 개발하였다. 입력변수를 전문가의 지식으로 변화시킨 규칙을 이용하여 퍼지시스템으로 가공한 후 인공신경망에 규칙을 입력하는 방식을 채택하여 주식수익률에 대하여 분류가 아닌 예측하는 방법을 선택하였다[29].

변수선정을 위해 유전자알고리즘을 이용하여 인공신경망과 결합한 모형이 통계적 기법이나 CART에 의한 변수선택을 이용하여 인공신경망과 결합한 모형보다 분류 예측력이 향상되었다고 주장하였다[2].

로지스틱 회귀분석이 선형구조를 쉽게 파악할 수 있고, 각 관찰치에 대해서 유일하게 예측확률을 갖을 수 있을 뿐만 아니라 기존의 분류분석 방법에 비해 매우 뛰어난 성능을 갖는 것으로 알려져 있다[5].

유전자 알고리즘은 다른 분류기법과 결합함으로 단일 모형보다 향상된 결과를 보였다고 주장하였다[7].

제 2 장 이론적 배경

2. 1 Logistic Regression

로지스틱 회귀분석은 종속변수가 범주형 자료인 이항변수로 구성된 일반화 선형모형의 특수한 경우로 여러 분야에서 사용되고 있다.

여러 설명변수로부터 두 범주만을 가지는 종속변수를 예측하는데 사용되는 로지스틱 회귀분석은 모형구조에 의해 연관성 및 교호 작용 유형을 설명할 수 있으며 모수의 추론을 통해서 반응 값에 대한 독립변수의 영향력을 평가할 수 있다. 또한 사후확률을 바탕으로 판별분석과 같은 판별 및 분류 분석의 기법으로도 사용할수 있는데 독립변수들의 동일한 공분산 행렬과 다변량 정규분포를 가정하는 판별분석에 비해 로지스틱 회귀분석은 독립변수에 대한 제약 조건이 적기 때문에 공분산 행렬과 다변량 정규분포의 가정들이 만족되지 못한 경우에 로지스틱 회귀분석을 사용하는 것이 더 좋은 결과를 가져다 준다고 한다.

그러나 로지스틱 회귀분석은 일반적인 회귀분석 기법이 가지고 있는 단점인 독립변수간의 교호효과와 독립변수의 수에 대한 한계를 극복하지 못하고 있다는 한계점을 지니고 잇다. 즉, 변수들간의 상관관계가 높은 경우에는 그 효과를 반영하지 못한다는 점과 독립변수들의 수가 증가함에 따라 설명력이 계속 증가한다는 점이다.

2. 2 Decision Tree

의사 결정 나무는 나무 구조(Tree Structure)를 이용하여 분류하므로 다른 기법들 보다 결과를 쉽게 이해할 수 있다는 장점을 가지고 있다.
Morgan 과 Songuist가 1960년대 초 AID(Automatic Interaction Detection)프로그램으로 회귀분석에 나무를 사용하였다. 그 후세 격인 분류 프로그램은 1970년대 초 Morgan과 Messenger에 의해 발전된 THAID(Theta Automatic Interaction Detection)이다. 그 후 CHAID, CART과 C4.5가 이 초기 방법들을 확장시키고 강하게 발전시켜왔다.

2. 2. 1 분류나무와 회귀나무

의사결정나무는 종속변수의 성격에 따라 분류나무(Classification Tree)와 회귀나무(Regression Tree)로 나눌 수 있다.

첫째, 분류나무는 종속변수가 이산형인 경우, 종속변수의 각 범주에 속하는 빈도에 기초하여 분리가 일어난다. 이산형 종속변수가 잘 구별되는 정도는 카이제곱 통계량 또는 지니계수(Gini Index), 엔트로피(Entropy)등의 분리기준에 의해 측정된다.

둘째, 회귀나무는 종속변수가 연속형인 경우, 종속변수의 평균에 기초해서 분리가 일어난다. 연속형 종속변수가 잘 구별되는 정도는 분산분석에서의 F-검정값 또는 분산의 감소(Variance Reduction)등의 분리기준에 의해 측정된다.

2. 2. 2 의사결정나무의 구성요소

의사결정나무의 일반적인 구성요소는 뿌리마디, 자식마디, 부모마디, 끝마디, 중간마디, 가지등으로 구성되어 있으며 구체적인 설명은 다음과 같다.

첫째, 뿌리마디(Root Node)는 나무구조가 시작되는 마디로써 전체로 이루어져 있다. 둘째, 자식마디(Child Node)는 하나의 마디로부터 분리되어 나간 2개 이상의 마디들을 말한다. 셋째, 부모마디(Parent Node)는 자식마디의 상위마디를 의미한다. 넷째, 끝마디(Terminal Node)는 각 나무줄기의 끝에 위치하고 있는 마디로써 잎(Leaf)이라고도 하며 결국 끝마디의 개수만큼 분류규칙이 생성되는 것이다. 다섯째, 중간마디(Internal Node)는 나무구조의 중간에 있는 끝마디가 아닌 마디들을 의미한다. 여섯째, 가지(Branch)는 하나의 마디로부터 끝마디까지 연결된 일련의 마디들을 의미하며, 이때 가지를 이루고 있는 마디의 개수를 깊이라고 한다.

2. 2. 3 의사결정나무의 분석과정

의사결정나무는 일반적으로 의사결정나무의 형성, 가지치기, 타당성 평가, 해석 및 예측의 단계로 나누어진다.

첫째, 의사결정나무의 형성단계는 분석의 목적과 자료구조에 따라 적절한 분리기준과 정지규칙을 지정하여 의사결정나무를 얻는다. 둘째, 가지치기단계는 오분류율을 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지를 제거한다. 셋째, 타당성 평가단계는 이익도표(Gains Chart)나 위험도표(Risk Chart) 또는 테스트용 자료에 의한 교차타당성 평가 등을 이용하여 의사결정나무를 평가한다. 넷째, 해석 및 예측단계는 의사결정나무를 해석하고 예측 모형을 설정한다.

2. 2. 4 분리기준

분리기준(Splitting Criterion)은 하나의 부모마디로부터 자식마디들이 형성될 때 예측변수의 선택과 범주의 병합이 이루어지는 기준을 의미한다. 즉 어떤 예측변수를 이용하여 어떻게 분리하는 것이 종속변수의 분포를 가장 잘 구별해 주는지를 파악하여 자식마디가 형성되고 분리기준으로 순수도나 불순도가 사용되는데 순수도는 각 노드에서의 종속변수의 분포가 얼마나 동질적인지를 측정하는 함수를 말한다.

2. 2. 5 가지치기

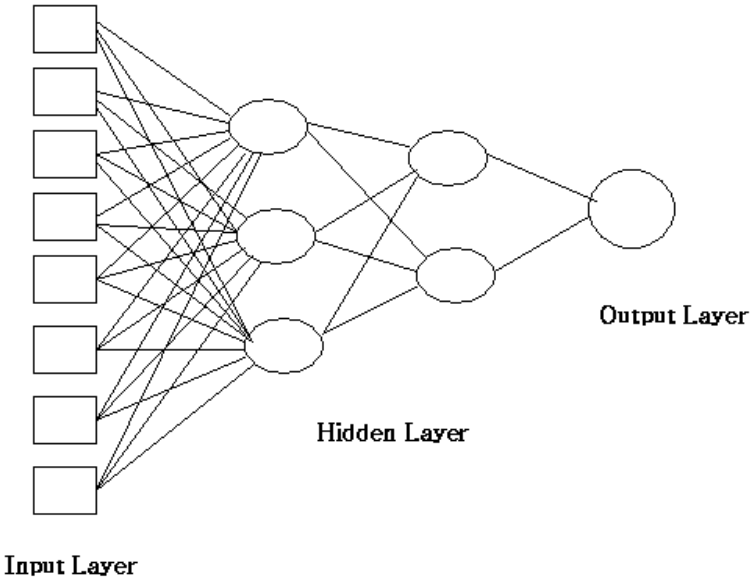
지나치게 많은 마디를 가지는 의사결정나무는 새로운 자료에 적용할 때 예측오차가 매우 클 가능성이 있다. 따라서 형성된 의사결정나무에서 적절하지 않는 마디를 제거하여 적당한 크기의 부나무(Subtree)구조를 가지는 의사결정나무를 최종적인 예측모형으로 선택하는 것이 바람직하다. CART, C4.5와 같은 알고리즘에서는 가지치기를 의사결정나무의 형성과정에 포함시키기도 하지만 실제로는 연구자가 적절한 가지치기를 수행해주는 것도 필요하다.

2. 3 Neural Network

두뇌 활동의 메커니즘을 수학적으로 재현한 인공지능의 한 분야이다. 여기서 인공지능이란 인간의 지각이나 경험에 의하여 가지게 되는 선형적 체험과 인공신경망(Artificial Neural Network)은 인간의 두뇌를 모방하여 같은 지적능력을 학습을 통하여 컴퓨터에 지식베이스를 구축하고, 구축된 지식베이스를 이용하여 주어진 자료를 추론하고 그 결과를 예측하고 설명하는 기능을 말한다.

신경망 모형을 구성하는 가장 기본적인 단위는 뉴런(Neuron)이며 기본적인 정보처리의 단위이다. 이는 입력값들을 가중 합산하여 그 결과를 전이함수(Transfer Function)로 전환하여 결과를 전달하는 기능을 수행한다. 신경망에서 일반적으로 사용되는 전이함수는 가중합산된 값을 그대로 사용하는 선형함수, 특정한 임계값을 기준으로 -1 이나 +1, 또는 0 이나 1의 값을 취하는 임계함수(Threshold Function), 0이하에서는 0의 값을 갖고 1 이상에는 1의 값을 가지며 그 사이의 값에서는 선형인 함수, 그리고 S자 형태의 변환을 가하는 함수(Sigmoid 또는 Hyper-Tangent Function) 등이 있다.

신경망은 뉴런의 연결방식과 학습방법에 따라 여러 종류로 구분된다. 그 중에서 가장 많이 사용되는 신경망 모형은 다층 퍼셉트론(Multilayer Perceptron)으로서, [그림 2-1]은 3층 구조를 가진 다층 퍼셉트론을 도식화하고 있다. 각 층은 입력값을 갖는 입력층, 정보처리 과정이 일어나는 은닉층, 출력값을 나타내는 출력층으로 구분된다. 신경망은 충분한 수의 은닉층이 있으면 어떠한 함수라도 표현할 수 있는 보편적인 함수식이라고 할 수 있다.



[그림 2-1] 3층 구조를 가진 다층 퍼셉트론

신경망이 주어진 자료의 특성을 학습하는데 사용되는 학습 알고리즘(Learning Algorithms)에는 여러 가지가 있으나 그 중에서 오차를 최소화시켜 나가는 BPN 방법이 흔히 사용된다. 이 알고리즘은 최소자승 알고리즘의 비선형적 확장으로 볼 수 있다. 즉, 입력층의 각 노드에 입력패턴을 주면 이 신호는 각 노드에서 변환되어 은닉층에 전달되고 계산과정을 거쳐 출력층에서 신호를 출력하게 된다. 이때 출력값과 목표값을 비교하여 둘 사이의 차이, 즉 오차를 줄여나가는 방향으로 가중치를 반복적으로 조정해 나가는 방법이다.

한편, 신경망 모형과 같은 비선형 모형은 학습패턴 이외에 대해서는 좋은 예측성과를 내지 못하는 과적합 문제가 제기된다. 이러한 문제는 연구결과의 일반화 가능성을 약화시키는 요인이 되는데 이러한 한계를 극복하고 신경망 모형의 일반화 성능을 향상시키기 위한 방법으로 흔히 사용되는 것이 결합 신경망(Combined

Neural Network)이다. 신경망의 오류는 편차와 분산으로 나누어 진다고 볼 때, 결합 신경망은 오류의 편차를 줄여줌으로써 일반화 성능을 향상시키는 방법이다.

인공 신경망은 공식적이고 정형화된 입력 자료를 요구하지 않기 때문에 여러 분야에 이상적으로 적용될 수 있다. 즉 적합한 자료뿐만 아니라 부적합한 자료를 식별하여 걸러내고, 적합한 정보만을 이용할 수 있다는 장점이 있다. 또한 잘 학습된 인공 신경망은 입력 자료에 이러한 부분이 포함되어 있다고 하더라도 정확한 결과를 산출할 수 있다는 장점이 있다. 반면에 인공신경망은 하나의 블랙박스처럼 기능 한다는 한계점을 가지고 있다. 즉 은닉층에서 자료처리가 이루어 지는데, 이용자의 입장에서는 입력자료를 어떻게 처리해서 어떻게 결론에 이르는지 추적할 수 있는 방법이 없다는 한계가 있다. 또한 학습과정에서 실질적으로 많은 시간과 노력이 요구된다는 문제점이 있다.

신경망의 추정에 있어서 비수렴성 문제는 비선형함수를 최적화할 때 매우 보편적으로 발생하는 문제이다. 만약 초기값이 국부최소값(Local Minimum)에 가깝다면 추정치는 전역 최소값(Global Minimum)이 아닌 국부최소값에서 수렴할 것이며, 국부최소값에서의 모수 추정치는 실제 데이터를 정확하게 추정하지 못할 가능성이 크다.

신경망의 적용분야는 데이터베이스 마케팅, 생화학분석, 광학문자인식, 의약품 개발, 주식평가, 의료진단, 음성인식 등에서 활발히 사용되고 있다.

2. 4 Genetic Algorithm

유전자 알고리즘(Genetic Algorithm)은 자연의 법칙인 '적자생존의 원리'에 근거를 두고 있다[7]. 즉 환경에 맞추어 생물들이 진화해 가는 과정을 알고리즘화한 것으로서 생물학에서 사용하는 용어를 그대로 사용하고 있다.

최근에 발전된 기법으로 자연선택과 유전인자를 기초로 하여 탐색과 최적해를 구하는 데 사용되는 진화연산에 의한 알고리즘이다. 본 기법은 비선형, 불연속 매개변수로부터 지속적으로 해를 향상시킴으로써 전체 최적해를 얻을수 있는 장점을 가진다.

유전자 알고리즘이 다른 탐색이나 최적화 방법과 달리 다양한 분야에 적용되는 이유로는 첫째, 알고리즘의 단순성과 일반성이다. 아무리 복잡한 문제라도 일단 염색체 형태로 표현이 되면 염색체의 복제와 재결합 및 적합도의 평가 등과 같은 비교적 단순한 연산과정의 반복을 통해 계산이 수행되어 빠르고 신빙성있는 해답을 준다.

둘째, 기존의 문제 해결방법과 결합하여 사용하기가 쉽다는 것이다. 이것은 유전자 알고리즘이 문제해결에 특수한 정보를 많이 사용하지 않고 또한 문제에 대한 배경지식이 있으면 이를 쉽게 수용할 수 있기 때문이다.

셋째, 설계변수의 형태에 관계없이 다양한 종류의 문제들에 효과적으로 적용가능하며 전체 최적점에 가까운 해를 구할 수 있다.

유전자 알고리즘은 진화론의 적자생존과 자연도태의 유전학에 근거한 적응탐색기법으로, 여러 가지 어려운 조합문제에 대한 효율적인 탐색을 수행하고 최적해에 근사한 해를 구할 수 있는 방법이다. 세대를 거듭함에 따라 최적의 해에 수렴하고 전세대의 생존자(우수개체)로부터 새로운 개체들의 집합이 형성된다.

후보해(Candidate Solution)들의 집합을 개체군(Population)이라고 하며, 이 후보해들은 염색체(Chromosome)라고 불린다. 염색체는 유전인자(Gene)로 구성되고, 각 유전인자들은 값을 갖는데, 이 값을 대립형질(Allele)이라고 한다. 실제 문제영역에

서 하나의 후보해가 X_1, X_2, X_3로 구성되어 있다고 하면, 유전적 알고리즘에서는 이것들의 값을 하나의 염색체로 표현한다. 그러므로 하나의 염색체는 X_1, X_2 및 X_3의 값을 표현하는 부분으로 구성되어 있는데, 각 X_i 의 값에 해당하는 염색체의 부분을 유전자형(Genotype)이라고 하고, 그 유전자형에 대응되는 실제 문제영역에서의 값, 즉 X_i의 값을 표현형(Phenotype) 이라고 한다. 하나의 염색체는 그에 따른 X_i 의 값에 의하여 실제문제에서의 적응함수값(Fitness value)을 갖는다.

2. 4. 1 유전 연산자

유전연산자(Genetic Operator)란 다음 세 가지 기능에 의해 진화와 도태의 과정을 반복하는데, 첫째는 재생(Reproduction)으로 잘 적응한 해들은 살아남고 잘 적응하지 못한 해들은 도태되도록 유도하는 조작법이다. 둘째는 교배(Crossover)로 두 부모해의 유전 정보를 임의의 위치에서 부분적으로 교환함으로써 새로운 자손해를 생성한다. 셋째는 돌연변이(Mutation)로 부모해로부터 자손해로 전달되는 특정한 유전정보에 대하여 무작위적인 변형을 시도함으로써 전체 해 집단에서 배제된 새로운 개체를 발생시키거나 진화과정에서 상실한 특정 유전정보의 재현을 시도하는 조작방법이다.

2. 4. 1. 1 Selection

선택 연산자(Selection Operator)는 적자생존의 개념을 구현하기 위해 의도된 것이다. 기본적으로, 선택 연산자는 한 시점의 모집단에 있어 이진열들 중 어떤 것들이 그들의 유전 형질(Genetic Material)을 다음 세대에 전할 것인지를 결정한다. 비율 선택(Proportional Selection)이라고 불리는 표준 선택 연산자는 모집단 p 중에서 복원으로 n을 임의로 선택하는 것으로 이루어진다. 현 모집단의 주어진 원소 j가 한 특별한 선택에서 뽑힐 확률은 그 이진열의 적합값(이진열이 적합함수에 의

해 계산된 값) 에 비례한다. 이러한 종류의 선택은 매우 자연적인 접근을 취하고 있는데, 왜냐하면 어떤 다른 이진열보다 적합값이 2배 높은 이진열은 다음 세대에서 다른 이진열보다 2배정도 많이 복사될 것이라고 예상된다.

2. 4. 1. 2 Crossover

교차 연산자는 모집단에서 새로운 개체들을 생성하는 핵심적인 연산자이다.

교차는 더욱 좋은 개체를 생성하기 위해서, 높은 적합값을 가진 이진열의 유전 형질을 합치도록 의도되었다. 교배집단은 나누어지고, 다음의 연산자가 교차확률 x로 각 쌍에 적용된다. 일반적으로 x의 값은 0.6보다 크고, 때로 x=1도 사용된다. 두 이진열에 확률 1-x로 아무런 변화도 주어지지 않지만, 확률 x로 두 부모간에 유전 형질이 교환된다.

가장 단순한 단위점 교차는 하나의 교차점이 임의로 선택된다. 그 후에 교차점의 오른편 이진값들은 두 부모간에 교환된다.

유전자 알고리즘의 여러 문헌에는 많은 다른 교차기법들이 소개되어 있다. 어떤 연구자들은 두점이나 다점교차도 사용한다. 이러한 기법들은 이진열의 길이가 상대적으로 길 때 특별히 유용하다.

2. 4. 1. 3 Mutation

돌연변이 연산자는 유전자 알고리즘에서 초기 모집단에 존재하지 않는 이진값을 포함하는 해를 발견하도록 한다. 이 연산자를 조절하는 모수는 돌연변이 확률이라고 하고 μ로 표기한다. 교차를 적용한 후, 어떤 이진열의 각 이진값은 확률μ로 반전된다(즉 원래 0 이었으면 1로 바뀌고 1 이었으면 0으로 바뀐다). 선택 연산자가 모집단의 다양성을 감소시키는 반면에 돌연변이 연산자는 다시 증가시킨다.

돌연변이 확률이 높을수록 조기 수렴의 위험이 작아지지만, 높은 돌연변이 확률은 유전자 알고리즘을 순수한 확률 탐색(Stochastic Search) 알고리즘으로 바꾼다. 그래서, 일반적으로 돌연변이 확률은 작은 값으로 설정된다.

2. 4. 1. 4 응용분야

유전자 알고리즘은 전통적인 통계적 방법론이 부적당한 큰 해공간(Solution Spaces)을 효율적으로 탐색할 수 있는 알고리즘이다. 생물학, 특히 집단유전학에서의 개념을 이용하여 작동하여, 잡음이 있을 때의 로버스트성과 높은 비선형성, 다변수 문제에 대한 적응력 때문에 관심을 끌고 있다.

모형화(Modelling), 추론, 분류 등 통계학에서 공통적으로 마주치는 문제는 상황이 복잡하고, 잠재적인 해가 많을 때 최선의 해를 선택하는 것이다. 다른 분야와 마찬가지로, 통계학에서도 기본적으로 이용되는 문제 해결 절차가 사용될 수 없는 특별한 형태의 문제들이 있다. 이러한 문제들은 일반적으로 다음과 같은 특성들을 가지고 있다.

- ○ 잠재적인 해공간이 너무 커서, 전체 탐색 절차를 이용할 수 없다.
- ○ 독립변수와 종속변수의 관계가 높은 비선형성을 나타내어, 국소최적값들이 많아 실제 최적값을 결정하기 힘들다.
- ○ 결정적으로 해를 찾는 절차가 잡음이 있으면 사용되기 힘들다.
- ○ 높은 차원, 또는 보조 정보가 부족

이러한 맥락에서 유전자 알고리즘을 이용한 접근은 매우 유용할 수 있다. 물론 이러한 종류의 문제들이 새로운 문제들은 아니며, 기존의 다른 방법들에 비해, 유전자 알고리즘이 특별히 가진 세 가지 이점은 다음과 같다.

○ 유전자 알고리즘은 하나의 해가 아니라, 여러 해들의 집단을 탐색한다. 그래서 해공간의 더욱 많은 부분을 탐색할 수 있으므로 국소 최적값에 빠질 가능성이 적다.

○ 유전자 알고리즘은 미분같은 보조 정보보다는 잠재적인 해의 가치를 평가하기 위해 유용한 점수화 정보를 이용한다. 최대하강법(Steepest Descent Method)과 같은 기울기를 이용하는 전통적인 최적화 기법들은

그러한 보조 정보에 의존하므로, 예를 들어 미분 불가능 함수를 최소화하려는 경우등에는 적용될 수 없다. 따라서 유전자 알고리즘이 더 넓은 응용성을 가지고 있다.

○ 유전자 알고리즘은 결정적인 전이 규칙보다는 확률적인 규칙을 이용한다. 즉, 유전자 알고리즘은 잡음이 있을 때라도, 국소 최적에 빠질 가능성이 적다. 결정적 기법들은 좋은 초기값의 선택에 큰 영향을 받지만, 유전자 알고리즘은 연속적으로 전이를 하는데 있어서 확률적 규칙을 이용함으로 이러한 초기값에 대한 종속을 줄이고 로버스트성을 증가시킨다.

제 3 장 모형에 관한 연구

전통적인 분류분석으로 모수적인 방법인 판별분석이나 로지스틱 회귀분석이 사용되었는데, 정규성 가정 위반시에는 로지스틱 회귀분석이 판별분석에 비해 더 좋은 예측력을 갖는 것으로 알려져 있다[20]. 비모수적인 방법으로는 특별한 가정을 요하지 않는 의사결정나무와 데이터의 형태에 영향을 적게 받는 신경망이 사용되고 있다. 최근에는 해석상의 편의는 다소 저하되더라도 정확도를 향상시키기 위하여 혼합모형에 관한 연구가 이루어지고 있다.

본 연구에서 제시하는 혼합모형 이외에도 나무-유전자 혼합 모형, 혼합 나무-최인접 모형, 혼합 나무-로짓 모형등이 있는데 혼합 나무-유전자 모형[8]이 분류 예측력 향상의 결과를 가져옴을 보였고 혼합 나무-로짓 모형의 결과가 유의적으로 좋다고 주장한 바 있다[26].

본 연구에서 제시하는 혼합 모형은 두 가지 관점에서 접근하고 있다. 첫 번째 제시하는 혼합모형 Ⅰ은 의사결정나무의 단점을 신경망 모형에 의해서 해결하고자 하였고, 두 번째 제시하는 혼합모형 Ⅱ는 신경망 모형의 단점을 유전자 알고리즘으로 보완하였다.

3. 1 혼합모형 Ⅰ: Hybrid Decision Tree-Neural Network

본 연구에서 제시한 첫번째 혼합모형은 의사결정나무의 단점을 신경망 모형에 의해 해결하고자 하였다. 단일모형으로는 의사결정나무와 신경망 모형을 사용하였고 의사결정나무와 신경망 모형을 결합한 혼합모형 Ⅰ의 모형을 구축하여 비교를 실시하였다.

의사결정나무는 순수도를 이용하여 의사결정규칙을 관심대상이 되는 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 또한 분석의 결과를 비교적 쉽게 해석하고 이로부터 유용한 정보를 얻을 수 있다. 순수도는 목표변수의 특정 범주에 개체들이 포함되어 있는 정도를 의미한다. 의사결정나무 분류 시 노드에 속한 개체수가 많고 순수도가 높은 경우에 잘 분류해 내지만 노드에 속한 개체수가 적고 순수도가 낮을 때는 잘 분류해 내지 못하는 점과 분리기준에 의해 분류의 결과가 확연히 달라질 수 있기 때문에 실제 자료의 예측 시에 불안정한 모형을 만들 수 있다는 단점을 가지고 있다.

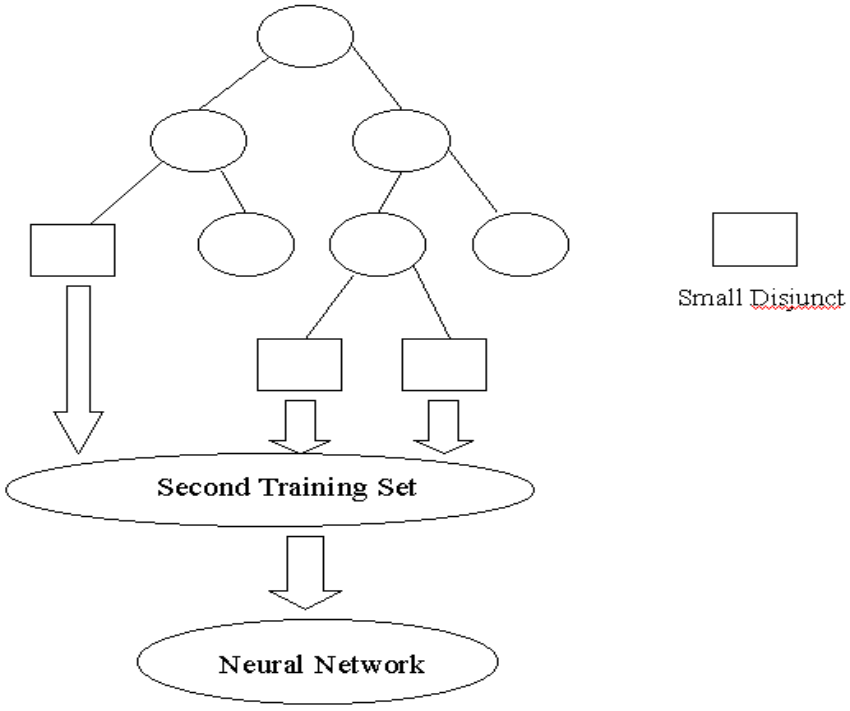
의사결정나무 결과 각 노드에 속한 개체수에 따라 Large Disjunct와 Small Disjunct로 구분한다. Small Disjunct란 훈련용 데이터 분석결과 노드에 속한 개체수가 작고 순수도가 낮은 노드를 말한다[16]. 본 연구에서는 훈련용 데이터의 분석결과 노드에 속한 개체수가 훈련용 전체 데이터 개수의 10% 미만의 데이터를 포함하면서도 종속변수의 사후확률 추정치가 0.4 ~0.6 사이인 경우인 노드로 정의한다.

의사결정나무 알고리즘은 일반적으로 Small Disjunct가 아닌 Large Disjunct에 잘 적합하는 경향이 있다. 각각의 Small Disjunct 들은 분류 정확도에 작은 영향을 미치는 것처럼 보이지만 각각의 Small Disjunct에 속한 개체들이 모이면 훈련용 데이터 에서 큰 수를 가짐으로 보다 나은 분류율을 가져 올수 있다[27]. 노드에 속한 개체수가 적은 Small Disjunct에서는 적중률이 낮아지는 경향이 있으므로 이

러한 노드들에 대한 적절한 처리를 통하여 적중률을 높일 수 있다.

본 연구에서 Small Disjunct에 속하는 개체들로 이루어진 2차 훈련용 데이터의 처리를 위해 신경망 모형을 이용하여 적중률을 높이고자 하였다.

따라서 의사결정나무 모형은 개체수가 적은 노드인 경우 잘 분류해 내지 못하는 단점을 가지는데 신경망 모형의 장점을 활용하여 새로운 접근법인 하이브리드 Decision Tree-Neural Network 모형을 구축하여 보완하려 하였다. 하이브리드 Decision Tree-Neural Network 모형의 전반적인 흐름의 형태는 [그림 3-1]과 같다.



[그림 3-1] 하이브리드 Decision Tree-Neural Network Training 구조 [8]

[그림3-1]에서 의사결정나무 분석에 사용된 훈련용 데이터를 원래 훈련용 데이터라 하였고 의사결정나무 분석 결과 Small Disjunct에 속하는 개체들을 모아 신경망의 훈련용 데이터로 사용된 것을 두 번째 훈련용 데이터로 표시하였으며 정사각형으로 나타내었다.

하이브리드 Decision Tree-Neural Network 구축 방법은 첫 번째가 통계적 기법에 의한 변수 선택을 하고 두 번째는 잘 알려진 의사결정나무 알고리즘에서 CART로 운영하였으며 세 번째는 의사결정나무 분석 결과 Small Disjunct로 판단된 노드의 데이터를 다루기 위해 신경망을 사용하였다.

3. 1. 1. 변수 선택 방법

변수선택 방법으로는 로지스틱 회귀분석의 선택적 방법에 의한 통계적 기법을 사용하였다. 사용된 소프트웨어는 SAS, E-miner를 이용하였다.

입력 변수 선택 방법을 정리하면 [표 3-1]와 같다.

[표 3-1] 입력 변수 선택 방법

선택 방법	선택 기준
통계적 기법(Stepwise)	로지스틱 회귀분석에 의한 선택

3. 1. 2. 트리 모형 구축

본 연구에서는 다양한 트리 모형중 이진분리를 사용하는 CART 알고리즘을 적용하였다. 분리기준으로는 불순도함수(Impurity Function)의 일종인 지니지수를 선택하였고 Leaf 노드 최소값으로는 25로 하였다.

3. 1. 3. 신경망 모형 구축

본 연구에서 신경망 모형은 MLP(다층 퍼셉트론)를 사용하여 은닉층의 노드수는 시행착오를 통해 최적의 값을 찾는 것이 원칙이지만 은닉층의 노드수가 결과에 미치는 영향을 배제시키기 위해서 은닉층의 노드수를 3으로 정하였다. MLP 모형의 출력은 1개이며 출력값은 [0, 1]의 범위에 존재하게된다. 신경망 실험시 Combination Function으로는 Linear, Activation Function으로는 Logistic을 사용하였다. 학습 반복횟수는 10,000번으로 하였다.

3. 1. 4. 하이브리드 Decision Tree-Neural Network 모형 구축

의사결정나무 모형에서 Small Disjunct로 판단된 노드만을 선택하여 신경망의 입력으로 사용하였고 의사결정나무에서 나온 결과와 2차 훈련용 데이터셋을 이용한 신경망의 결과를 종합하여 하이브리드 Decision Tree-Neural Network 모형의 오분류율을 계산하였다.

본 연구에서 하이브리드 Decision Tree-Neural Network 결합 방법은 [표 3-2]와 같다.

[표 3-2] 하이브리드 Decision Tree-Neural Network 결합 방법

개체가 속한 노드	하이브리드 Decision Tree-Neural Network
Large Disjunct 인 경우	의사결정나무의 예측에 따름
Small Disjunct 인 경우	신경망의 예측에 따름

의사결정나무 분석 결과 개체가

1) Large Disjunct 에 속하는 경우

의사결정나무를 이용하여 예측한다.

2) Small Disjunct 에 속하는 경우

신경망을 이용하여 예측한다.

결과적으로 하이브리드 Decision Tree-Neural Network 모형은 의사결정나무 분석결과에서 나온 Disjunct 형태에 따라 의사결정나무의 결과를 따를 것인지 신경망의 결과를 따를 것인가 결정된다. 모형을 구축하기 위하여 E-miner, SAS를 사용하였다.

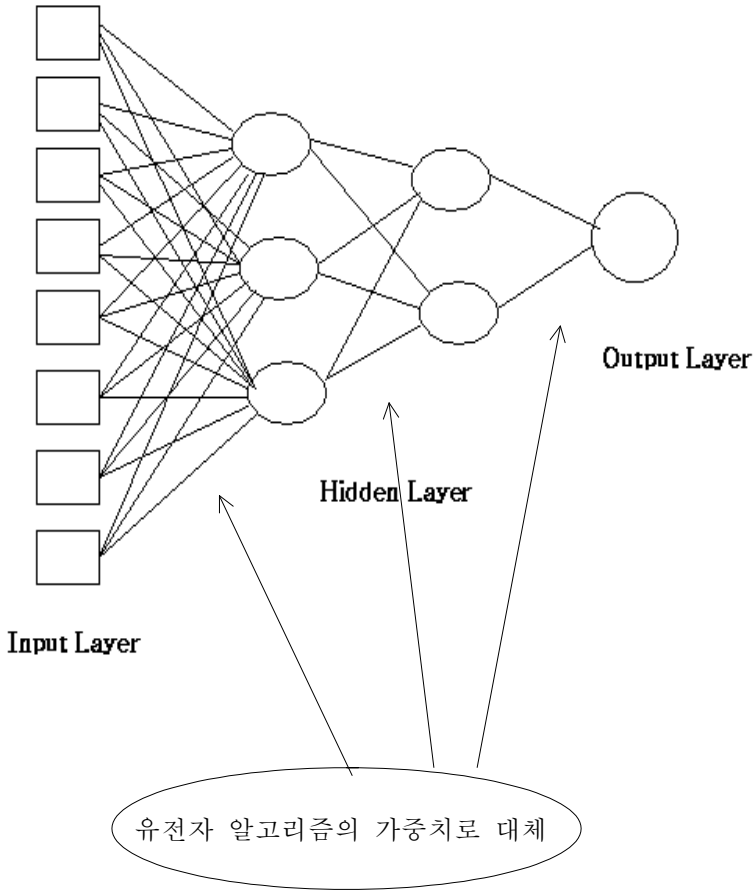
3. 2 혼합모형 Ⅱ : Hybrid Genetic-Neural Network

혼합모형 Ⅱ는 신경망 모형의 단점을 유전자 알고리즘으로 보완하여 최적의 연결 가중치를 사용하여 결합하고자 하는 것이다.

신경망에서는 일반적으로 사용되는 복잡한 비선형 최적화 문제에 적용되는 역전파 알고리즘의 단점인 기울기 검색기법의 한계로 자주 일관성이 없고 예측 불가능한 성능을 보인다. 초기값이 지역 최소값에 가까우면 추정치는 전역 최소값이 아닌 지역 최소값으로 수렴할 것이며, 지역 최소값에서의 모수 추정치는 실제 데이터를 정확하게 추정하지 못할 가능성이 많다는 단점을 가지고 있다.

반면 유전자 알고리즘은 전역 최소값을 찾고 쉽게 지역 최소값에 빠지지 않으며, 광범위하고 넓은 검색공간에서 효율적이라는 장점을 가진다. 유전자 알고리즘은 에러를 최소화 하는데 있어 전역 최소값으로 접근하며 최적화문제의 해결에 많이 사용되고 있다. 유전자 알고리즘은 일반적으로 임의의 선택점을 시작으로 하여 에러를 급격하게 줄일수 방향으로 가중치가 조정된다.

신경망 모형은 지역 최소값에서의 모수추정치는 실제 데이터를 정확하게 추정하지 못하므로 유전자 알고리즘으로 보완하여 새로운 하이브리드 Genetic-Neural Network모형을 구축하였다. 하이브리드 Genetic-Neural Network 모형의 흐름의 형태는 [그림 3-2]와 같다.



[그림 3-2] 하이브리드 Genetic-Neural Network 구조 [29]

하이브리드 Genetic-Neural Network 구축 방법은 첫 번째는 로지스틱 회귀분석의 선택적 방법에 의한 통계적 기법을 사용하여 변수를 선택하였고 두 번째는 신경망 모형에 의해 오분류율을 최소화하는 모형을 구축하고 세 번째는 신경망의 가중치를 유전자 알고리즘에 구한 최적의 가중치로 대체하는 것이다.

3. 2. 1. 변수 선택 방법

변수선택 방법으로는 로지스틱 회귀분석의 선택적 방법에 의한 통계적 기법을 사용하였다.

3. 2. 2. 유전자 모형 구축

본 연구에서 사용한 유전자 모형은 오분류율을 최소화하기 위해 집단의 크기 1000, 교배율 0.5, 돌연변이율 0.06로 고정시켜 사용하였으며 유전자 알고리즘을 이용하기 위해 Evolver 4.0을 사용하였다.

3. 2. 3. 신경망 모형 구축

혼합모형Ⅱ의 분류 정확도를 혼합모형Ⅰ과 비교하기 위해 신경망 모형에서의 옵션 설정은 혼합모형Ⅰ과 동일하게 설정하여 사용하였다.

3. 2. 4. 하이브리드 Genetic-Neural Network 모형 구축

본 연구에서는 이미 모형화된 신경망 모델에 대해서 최적의 가중치를 찾는 것이다. 신경망 모형에 연결할 수 있는 최적의 가중치를 유전자 알고리즘에 의해 훈련용 자료에서 구한 뒤, 이를 검증용 자료에 적용하여 분류력을 비교해 보는 것이다. 하이브리드 Genetic-Neural Network 가중치 조절 규칙은 관측치와 예측치 사이의 오차제곱합(SSE)를 최소화 하는 것인데, 다시 말해서 오분류율을 최소화 하고자 하는 것이다. 신경망에서 선택된 히든 노드(Hidden Node)수를 그대로 사용하며 단지 중간 단계에 들어가는 가중치만을 유전자 알고리즘을 통해 조절하게끔 하였다.

모델 구축에 앞서 자료의 전처리 작업을 시행하였으며 변수들의 범위는 -50 에서 50사이로 설정 집단의 크기 1000, 교배율 0.5, 돌연변이율 0.06로 고정시켜 사용하였다.

신경망에서 나온 결과와 유전자 알고리즘에 가중치를 수정하여 나온 결과를 결합함으로 하이브리드 Genetic-Neural Network 모형의 오분류율을 계산하였다. 모형을 구축하기 위하여 Neuroshell 2.0, Evolver 4.0 을 사용하였다.

제 4 장 실험 및 결과

4. 1 실험 계획

기존의 분류목적에 사용된 의사결정나무, 신경망을 실시한 결과와 본 연구에서 제시한 혼합모형의 결과와 비교하고자 한다. 실험 절차에 대해 설명하면 다음과 같다. 단일 모형으로는 의사결정나무, 신경망을 실시한다.

단일 모형으로 분류 예측한 결과와 혼합모형 I 인 하이브리드 Decision Tree-Neural Network 모형과 혼합모형 II인 하이브리드 Genetic-Neural Network의 예측 결과를 비교하여 어떠한 성능을 가져다주는지 보고자 한다.

1994년 ~ 2002년까지 연세대학교 의과대학 세브란스 병원 건강증진센터에서 검사를 받아온 환자 62,593개의 안과 데이터를 이용하여 두가지의 경우로 나누어 실험해 보았다. 첫째, 결측치를 제외한 완전한 자료를 토대로 분석하는 경우와 둘째, 결측치를 가지고 있는 데이터를 그대로 사용하여 분석하는 경우이다. 결측치를 제외한 완전한 자료는 Sampling을 100%로 시도하였고 결측치를 포함한 자료는 Sampling을 50%로 하여 데이터의 개수를 비슷하게 하였다. 변수선택 방법으로 통계적 기법인 로지스틱 회귀분석 방법을 이용하였다.

결측치는 연속형 변수에서는 평균, 범주형 변수에서는 카이제곱에 의한 방법으로 처리하였으며 43개의 변수 중에서 1개의 종속변수와 42개의 독립변수를 사용하여 분석을 시도하였다.

모형이 과적합 되는 것을 막기 위해 데이터 분할을 훈련용과 테스트 자료로 7:3 비율로 나누어 사용하였고 안정된 결과를 얻기 위해 표본을 랜덤하게 추출하여 시뮬레이션을 10번 하였다.

최종적으로 시뮬레이션 실험 결과를 가지고 T-검정을 실시하여 유의적인 차이를 보이는지 확인해 보았다.

4. 2 실험 자료

본 연구에서는 두가지의 경우로 나누어 실험해 보았다.

첫째, 62,593개의 안과데이터에서 하나 이상의 결측치를 포함한 데이터를 제외시켜 분석하는 경우와 둘째, 결측치를 가지고 있는 데이터를 그대로 사용하여 분석하는 경우이다.

4. 2. 1 완전한 자료(결측치를 제외한 자료)

[표 4-1] 자료 설명(완전한 자료)

※ 안과 정밀검사 자료

데이터 출처	연세대학교 의학통계학과
데이터 내용	1994년 ~ 2002년까지 연세대학교 의과대학 세브란스 병원 건강증진센터에서 안과정밀 검사를 받아온 환자
데이터 개수	26,083개
데이터 형태	연속형(33개), 범주형(10개)
종속 변수	Diagt(0:정상 1:정밀검사요함) - 범주형(Missing 0%)

4. 2. 2 불완전한 자료(결측치를 포함한 자료)

[표 4-2] 자료 설명(불완전한 자료)

※ 안과 정밀검사 자료

데이터 개수	62,593개
데이터 형태	연속형(33개), 범주형(10개)
종속 변수	Diagt(0:정상 1:정밀검사요함) - 범주형(Missing 0%)

4. 2. 3 독립변수 설명

[표 4-3] 독립변수 설명

	변 수 명	변수형태	설 명	Missing%
독립 변수	A01_Height	연속형	신장	0
	A01_Weight	연속형	체중	0
	A01_Stdweight	연속형	표준체중	0
	A01_Bmi	연속형	비만도	0
	A11_Rbc	연속형	Rbc	15
	A11_Hb	연속형	Hb	15
	A11_Hct	연속형	Hct	15
	A11_Wbc	연속형	Wbc	15
	A11_Plt	연속형	혈소판	15
	A11_bioNa	연속형	나트륨	15
	A11_bioK	연속형	칼륨	15
	A11_bioCL	연속형	염소	15
	A11_bioCo2	연속형	이산화탄소	15
	A11_bioCa	연속형	칼슘	15
	A11_bioP	연속형	인	15
	A11_bioGlucose	연속형	혈당	15
	A12_Eat_vitamin_a	연속형	비타민 A	47
	A12_Eat_vitamin_c	연속형	비타민 C	47
	A11_bioBun	연속형	혈중요소질소	15

	변 수 명	변수형태	설 명	Missing%
독립 변수	A11_bioCreat	연속형	크레아티닌	15
	A11_bioUric	연속형	요산	15
	A11_bioProtein	연속형	총단백	15
	A11_bioAlb	연속형	알부민	15
	A11_bioBil	연속형	총빌리루빈	15
	A11_bioAst	연속형	Ast	15
	A11_bioAlt	연속형	Alt	15
	A11_bioRgt	연속형	R-Gt	15
	A11_bioChol	연속형	총콜레스테롤	15
	A11_bioTrig	연속형	중성지방	15
	A11_bioHDL	연속형	고밀도 콜레스테롤	15
	A11_bioLDH	연속형	LDH	15
	A11_Ph	연속형	신도	15
	Sage10	연속형	나이	0
	A11_Bilirubin	범주형	빌리루빈	16
	A11_Protein	범주형	단백	16
	A11_Glucose	범주형	요당	16
	A11_Ketone	범주형	케론체	16
	A11_Blood	범주형	잠혈	16
	A11_Urobil	범주형	요빌리노겐	15
	A11_Nitrite	범주형	아질산염	16
	A11_Uwbc	범주형	백혈구	16
	Sex	범주형	성별	0

4. 3 연구 모형 구축

4. 3. 1 완전한 자료(결측치를 제외한 자료)

4. 3. 1. 1 데이터 분할

데이터 분할 시 7:3의 비율로 훈련용, 테스트용 자료로 사용하였다.
데이터 분할에 사용된 자료의 개수는 [표 4-4]에 제시된 바와 같다.

[표 4-4] 데이터 분할

안과진단자료	데이터 분할
훈련용 자료	18,258
테스트용 자료	7,825
총 자료	26,083

4. 3. 1. 2 이상치 제거

범주형 변수는 각 변수의 범주중 1이하의 빈도를 갖는 범주를 분석에서 제외시켰으며 연속형 변수는 중위수를 중심으로 ±9밖의 값을 이상치로 간주하여 분석에서 제외시켰다.

[표 4-5] 이상치를 제거한 자료

안과진단자료	이상치	이상치를 제거한 최종자료
훈련용 자료	4,911	13,347

4. 3. 2 불완전한 자료(결측치를 포함한 자료)

4. 3. 2. 1 데이터 분할

데이터 분할 시 7:3의 비율로 훈련용, 테스트용 자료로 사용하였다.
데이터 분할에 사용된 자료의 개수는 [표 4-6]에 제시된 바와 같다.

[표 4-6] 데이터 분할

안과진단자료	데이터 분할
훈련용 자료	21,847
테스트용 자료	9,363
총 자료	31,210

4. 3. 2. 2 이상치 제거

범주형 변수는 각 변수의 범주중 1이하의 빈도를 갖는 범주를 분석에서 제외시켰으며 연속형 변수는 중위수를 중심으로 ±9밖의 값을 이상치로 간주하여 분석에서 제외시켰다.

[표 4-7] 이상치를 제거한 자료

안과진단자료	이상치	이상치를 제거한 최종자료
훈련용 자료	5,766	16,081

4. 4 분석 결과

4. 4. 1 혼합모형 I : Hybrid Decision Tree-Neural Network

4. 4. 1. 1 변수 선택 결과

[표 4-8] 입력변수 선택방법 결과

※ 안과 정밀진단 자료

선택 방법		결 과
통계적 기법	Stepwise방법	Sage, A11_bioglucose, Sex1, A11_biop, A11_plt, A11_bioco2, A12_vitamin_a, A11_ph, A11_bioalb, A11_biorgt, A11_biochol, a11_bioldh, A11_bioca, A11_biobun

나이(Sage), 혈당(A11_bioglucose), 성별(Sex1)이 종속변수에 가장 많은 영향을 미치고 다음으로 인(A11_biop), 혈소판(A11_plt), 이산화탄소(A11_bioco2), 비타민A(A12_vitamin_a), 신도(A11_ph), 알부민(A11_bioalb), R-gt(A11_biorgt), 총콜레스테롤(A11_biochol), Ldh(A11_bioldh), 칼슘(A11_bioca), 혈중요소질소(A11_biobun)이 독립변수로 선택이 되어 단일모형인 트리와 신경망 모형의 입력

변수로 들어가는 것이다.

4. 4. 1. 2 분석 결과

4. 4. 1. 2. 1 완전한 자료(결측치를 제외한 자료)

본 연구에서 결측치를 제외한 완전한 자료에서의 형태는 [그림 4-1]과 같다.

STAT	DIAGT	==> 1	==> 0	TOTAL
N	1	7	1701	1708
N	0	2	11637	11639
N	+	9	13338	13347
Row%	1	0	100	100
Row%	0	0	100	100
Row%	+	0	100	100
Col%	1	78	13	13
Col%	0	22	87	87
Col%	+	100	100	100
%	1	0	13	13
%	0	0	87	87
%	+	0	100	100

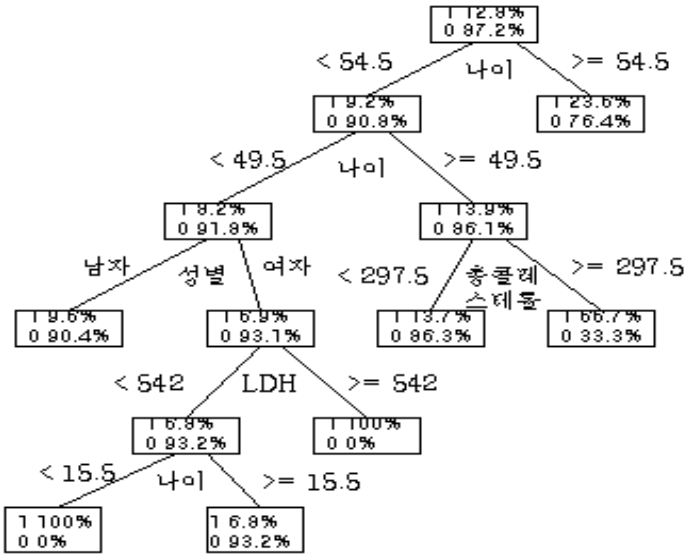
[그림 4-1] 결측치를 제외한 자료

Name	Importance
SAGE	1,0000
SEX1	0,5639
A11_BIOCHOL	0,0612
A11_BIOLDH	0,0433

[그림 4-2] 변수의 중요도(의사결정나무 결과)

[그림 4-2] 결과에 의하면 의사결정나무 분석결과, 나이, 성별, 총콜레스테롤, LDH가 중요한 변수로 선택되어 분류가 이루어지고 있다. 시뮬레이션 10번중 오분

류율이 가장 낮은 최적의 의사결정나무 분류 결과를 살펴보면 [그림 4-3]과 같다.



[그림 4-3] 의사결정나무 분류 결과(네번째 시뮬레이션)

[그림 4-3]을 살펴보면 백내장에 걸릴 위험의 소지가 높은 경우는 첫째, 나이가 54.5세 이상인 사람, 둘째, 나이가 49.5 이상이고 총콜레스테롤이 297.5 이상인 사람, 셋째, 나이가 49.5세 미만이고 여자이면서 LDH가 542이상인 사람이 백내장에 걸릴 위험의 소지가 높았다.

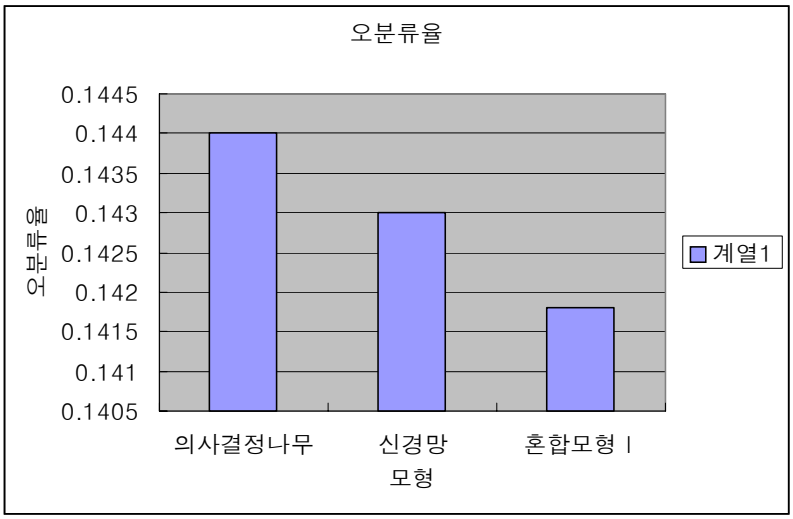
본 연구에서 의사결정나무의 분석결과 Small Disjunct로 판단된 개체에 대해 분류의 결과를 향상시키고자 하이브리드 Decision Tree-Neural Network 모델을 구축하여 의사결정나무에서 제대로 분류하지 못한 부분을 신경망의 2차 훈련용 데이터로 받아들여 다시 분류하였다. 하이브리드 Decision Tree-Neural Network 모델의 성능 결과를 단일 모형과 비교해 보고자 하였다.

SAS에서 Small Disjunct에 해당되는 노드를 추출하기 위해 노드 번호(Node

Identification Number)를 참조하여 Small Disjunct로 판단된 데이터를 뽑아서 신경망의 입력으로 사용하였다. 표본을 랜덤하게 추출한 결과 Small Disjunct에 해당되는 데이터의 개수가 250~500개 사이의 범위를 가졌다. 10번의 시뮬레이션 결과 최적의 의사결정모형으로 판단된 Small Disjunct의 데이터의 개수는 290개였고 이것이 신경망의 2차 훈련용 데이터로 사용되었다.

[표 4-9] 안과 자료에 대한 오분류율 - 테스트 자료 사용

시뮬레이션	의사결정나무	신경망	Tree-Neural Network
1	0.132	0.139	0.133
2	0.154	0.145	0.142
3	0.151	0.149	0.156
4	0.147	0.155	0.147
5	0.138	0.146	0.149
6	0.156	0.137	0.138
7	0.145	0.135	0.142
8	0.135	0.141	0.133
9	0.139	0.147	0.137
10	0.143	0.136	0.141
평균	0.144	0.143	0.1418



[그림 4-4] 오분류율 평균 - 테스트 자료 사용(혼합모형 I)

[표 4-9]은 안과 자료에 대한 결과인데, 10번의 시뮬레이션 평균 결과에서 단일 모형으로는 신경망이 의사결정나무보다 약간 더 낮은 오분류율을 갖는 것으로 나타났다. 단일모형과 혼합모형을 비교해볼 때 하이브리드 Decision Tree-Neural Network 모형이 단일모형보다는 더 낮은 오분류율을 가짐으로써 보다 정확한 분류를 할수 있다는 것을 알수 있었다.

4. 4. 1. 2. 2 불완전한 자료(결측치를 포함한 자료)

본 연구에서 결측치를 포함한 불완전한 자료에서의 형태는 [그림 4-5]과 같다.

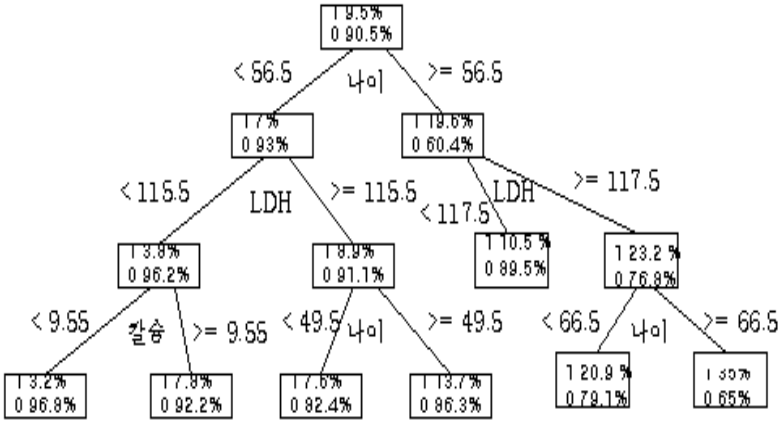
STAT	DIAGT	==> 1	==> 0	TOTAL
N	1	12	1509	1521
N	0	0	14560	14560
N	+	12	16069	16081
Row%	1	1	99	100
Row%	0	0	100	100
Row%	+	0	100	100
Col%	1	100	9	9
Col%	0	0	91	91
Col%	+	100	100	100
%	1	0	9	9
%	0	0	91	91
%	+	0	100	100

[그림 4-5] 결측치를 포함한 자료

Name	Importance	Role
SAGE	1,0000	input
A11_BIOLDH	0,7402	input
SEX1	0,4746	input
A11_BIOCA	0,2356	input
A11_PLT	0,0601	input
A11_BIOBUN	0,0564	input

[그림 4-6] 변수의 중요도(의사결정나무 결과)

[그림 4-6] 결과에 의하면 의사결정나무 분석결과, 나이, LDH, 성별, 칼슘, 혈소판, 혈중요소질소가 중요한 변수로 선택되어 분류가 이루어지고 있다. 시뮬레이션 10번중 오분류율이 가장 낮은 최적의 의사결정나무 분류 결과를 살펴보면 [그림 4-7]과 같다.



[그림 4-7] 의사결정나무 분류 결과(첫번째 시뮬레이션)

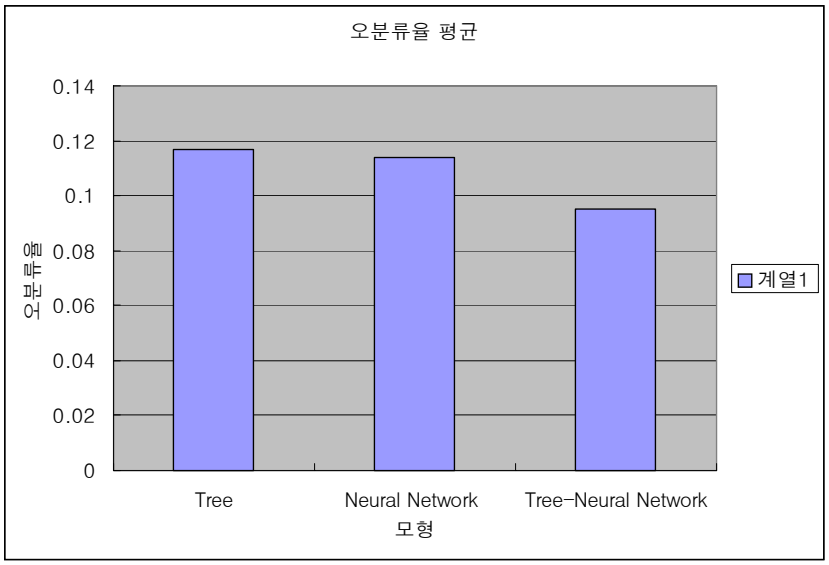
[그림 4-7]을 살펴보면 백내장에 걸릴 위험의 소지가 높은 경우는 첫째, 나이가 56.5세 이상인 사람, 둘째, 나이가 56.5 이상이고 LDH가 117.5 이상인 사람, 셋째, 나이가 56.5세 미만이고 LDH가 115.5 미만이면서 칼슘이 9.55 이상인 사람이 백내장에 걸릴 위험의 소지가 높았다.

본 연구에서 의사결정나무의 분석결과 Small Disjunct로 판단된 개체에 대해 분류의 결과를 향상시키고자 하이브리드 Decision Tree-Neural Network 모델을 구축하여 의사결정나무에서 제대로 분류하지 못한 부분을 신경망의 2차 훈련용 데이터로 받아들여 다시 분류하였다. 하이브리드 Decision Tree-Neural Network 모델의 성능 결과를 단일 모형과 비교해 보고자 하였다.

표본을 랜덤하게 추출한 결과 Small Disjunct에 해당되는 데이터의 개수가 300~700개 사이의 범위를 가졌다. 10번의 시뮬레이션 결과 최적의 의사결정모형으로 판단된 Small Disjunct의 데이터의 개수는 552개였고 이것이 신경망의 2차 훈련용 데이터로 사용되었다.

[표 4-10] 안과 자료에 대한 오분류율 - 테스트 자료 사용

시뮬레이션	의사결정나무	신경망	Tree-Neural Network
1	0.122	0.116	0.092
2	0.115	0.120	0.085
3	0.131	0.121	0.106
4	0.102	0.105	0.083
5	0.126	0.132	0.107
6	0.114	0.112	0.092
7	0.12	0.105	0.102
8	0.115	0.114	0.100
9	0.114	0.111	0.087
10	0.116	0.112	0.103
평균	0.117	0.114	0.095



[그림 4-8] 오분류율 평균 - 테스트 자료 사용(혼합모형 I)

[표 4-10]은 안과 자료에 대한 결과인데, 10번의 시뮬레이션 평균 결과에서 단일 모형으로는 신경망이 의사결정나무보다 약간 더 낮은 오분류율을 갖는 것으로 나타났다. 단일모형과 혼합모형을 비교해볼 때 하이브리드 Decision Tree-Neural Network 모형이 단일모형보다는 더 낮은 오분류율을 가짐으로써 보다 정확한 분류를 할수 있다는 것을 알수 있었다.

4. 4. 1. 3 유의성 검정

위의 결과를 토대로 단일모형인 의사결정나무, 신경망 모형과 혼합모형인 하이브리드 Decision Tree-Neural Nework모형이 통계적으로 차이가 있는지에 대해서 알아보기 위해 각 모형간의 비교를 실시해 보았다.

10번의 시뮬레이션 결과에 대해서 각 모형끼리 T-test를 통해서 모형간의 차이가 통계적으로 의미가 있는지에 대해서 알아보고자 한다.

[표 4-11] 오분류율에 대한 기술통계량(결측치를 제외한 자료)
안과 자료에 대한 10번 시뮬레이션 결과

모형구분	N	오분류율 평 균	오분류율 표준편차
의사결정나무	10	0.144	0.026
신경망	10	0.143	0.025
Tree-Neural Network	10	0.1418	0.023

[표 4-12] 오분류율에 대한 기술통계량(결측치를 포함한 자료)

안과 자료에 대한 10번 시뮬레이션 결과

모형구분	N	오분류율 평 균	오분류율 표준편차
의사결정나무	10	0.117	0.019
신경망	10	0.114	0.017
Tree-Neural Network	10	0.095	0.013

[표 4-11]의 결과는 안과 자료의 오분류율에 대한 요약자료로서 하이브리드 Decision Tree-Neural Network 모형이 의사결정나무나 신경망에 비해서 오분류율이 0.002정도 낮음을 알 수 있었다.

[표 4-12]의 결과는 하이브리드 Decision Tree-Neural Network 모형이 의사결정나무나 신경망에 비해서 오분류율이 0.02정도 낮음을 알 수 있었다. 따라서 하이브리드 Decision Tree-Neural Network모형이 단일 모형에 비해 더 나은 분류 예측력을 갖는다고 말할 수 있다.

다음으로는 이러한 차이의 정도가 통계학적으로 의미가 있는 차이인지 알아보기 위해 각 단일모형인 의사결정나무, 신경망과 하이브리드 Decision Tree-Neural Network모형간의 T-test를 실시해 보았다.

[표 4-13] 오분류율 결과를 기초로 한 모형간의 T-test 결과

(결측치를 제외한 자료)

모형간 비교	오분류율간 평균 차이	차이의 표준오차	T값	유의확률
의사결정나무 과 Tree-Neural Network	0.0022	0.00042	5.42	0.000
신경망 과 Tree-Neural Network	0.0012	0.00036	3.33	0.031

[표 4-14] 오분류율 결과를 기초로 한 모형간의 T-test 결과

(결측치를 포함한 자료)

모형간 비교	오분류율간 평균 차이	차이의 표준오차	T값	유의확률
의사결정나무 과 Tree-Neural Network	0.022	0.0058	3.79	0.000
신경망 과 Tree-Neural Network	0.019	0.0055	3.46	0.000

우선 T-test를 하기에 앞서 두 개의 독립집단이 분산이 동일한지에 대해서 살펴보았다. 그 결과 분산이 동일하지 않는 이분산으로 나타났다. 따라서 두 개의 집단은 분산이 동일하지 않다는 가정하에 T-test를 실시하였다.

[표 4-13], [표 4-14]의 결과를 통해 유의수준 0.05하에 단일모형과 하이브리드 Decision Tree-Neural Network 모형간의 차이는 통계학적으로 의미가 있는 차이임을 알 수 있었다.

위의 내용을 종합해 보면 안과 자료에 대해서 단일모형에 비해 트리의 단점을 신경망으로 보완한 하이브리드 Decision Tree-Neural Network 모형이 더 나은

분류력을, 이러한 분류력의 차이가 통계학적으로 의미가 있는 차이라고 말할 수 있겠다.

4. 4. 2 혼합모형Ⅱ : Hybrid Genetic-Neural Network

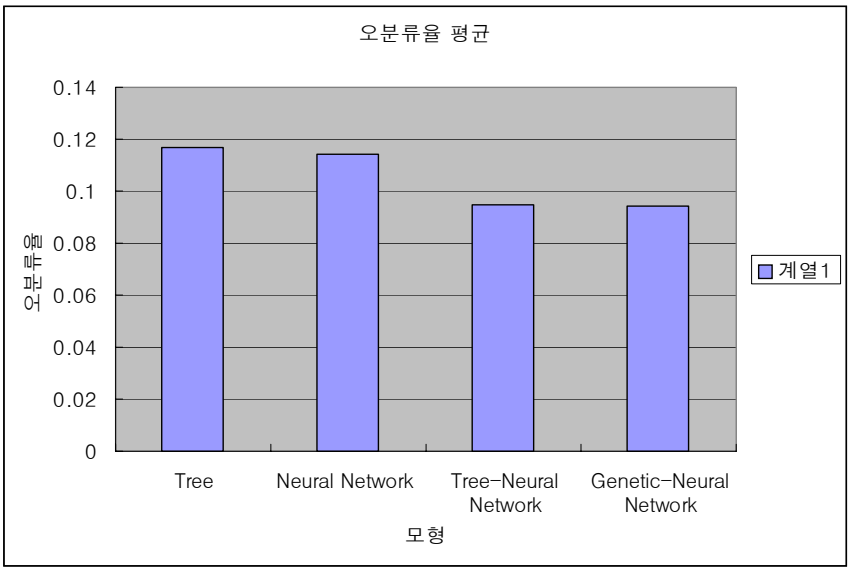
4. 4. 2. 1 분석 결과

본 연구에서 신경망의 분류의 결과를 향상시키고자 하이브리드 Genetic-Neural Network 모델을 구축하였다. 신경망 모형에서 지역적 최소값 에서의 모수 추정치는 실제 데이터를 정확하게 추정하지 못할 가능성이 많다는 단점을 지역적 최소값에 쉽게 빠지지 않는다는 유전자 알고리즘으로 보완하여 신경망의 가중치를 수정하였다. 하이브리드 Genetic-Neural Network모델 성능의 결과를 단일 모형과 비교해 보고자 하였다.

[표 4-15] 안과 자료에 대한 오분류율 - 테스트 자료 사용

(결측치를 포함한 자료)

시뮬레이션	의사결정나무	신경망	Tree-Neural Network	Genetic-Neural Network
1	0.122	0.116	0.092	0.091
2	0.115	0.120	0.085	0.087
3	0.131	0.121	0.106	0.110
4	0.102	0.105	0.103	0.096
5	0.126	0.132	0.107	0.092
6	0.114	0.112	0.092	0,091
7	0.12	0.105	0.102	0.105
8	0.115	0.114	0.100	0.102
9	0.114	0.111	0.087	0.082
10	0.116	0.112	0.083	0.090
평균	0.117	0.114	0.095	0.094



[그림 4-9] 오분류율 평균 - 테스트 자료 사용(혼합모형Ⅱ)

[표 4-15]은 안과 자료에 대한 결과인데, 10번의 시뮬레이션 평균 결과에서 혼합모형Ⅰ,Ⅱ가 단일모형 보다 더 낮은 오분류율을 갖는 것으로 나타났다. 혼합모형Ⅰ과 혼합모형Ⅱ를 비교해볼 때 혼합모형 Ⅱ가 혼합모형Ⅰ 보다 약간 더 낮은 오분류율을 가짐을 알수 있었다.

4. 4. 2. 2 유의성 검정

위의 결과를 토대로 단일모형인 신경망모형과 혼합모형Ⅰ인 Decision Tree-Neural Network, 혼합모형Ⅱ인 Genetic-Neural Nework모형이 통계적으로 차이가 있는지에 대해서 알아보기 위해 각 모형간의 비교를 실시해 보았다.

10번의 시뮬레이션 결과에 대해서 각 모형끼리 T-test를 통해서 모형간의 차이가 통계적으로 의미가 있는지에 대해서 알아보고자 한다.

[표 4-16] 오분류율에 대한 기술통계량(결측치를 포함한 자료)
안과 자료에 대한 10번 시뮬레이션 결과

모형구분	N	오분류율 평 균	오분류율 표준편차
신경망	10	0.114	0.017
Tree-Neural Network	10	0.095	0.013
Genetic-Neural Network	10	0.094	0.014

[표 4-16]의 결과는 안과 자료의 오분류율에 대한 요약자료로서 혼합모형Ⅱ 모형인 Genetic-Neural Network 모형은 신경망에 비해 오분류율이 0.02정도 낮음을 알 수 있었다.

다음으로는 이러한 차이의 정도가 통계학적으로 의미가 있는 차이인지 알아보

기 위해 단일모형인 신경망과 혼합모형Ⅰ인 Decision Tree-Neural Network와 혼합모형Ⅱ인 Genetic-Neural Network모형간의 T-test를 실시해 보았다.

[표 4-17] 오분류율 결과를 기초로 한 모형간의 T-test 결과
(결측치를 포함한 자료)

모형간 비교	오분류율간 평균 차이	차이의 표준오차	T값	유의확률
신경망 과 Genetic-Neural Network	0.02	0.0061	3.28	0.03
Tree-Neural Network 와 Genetic-Neural Network	0.001	0.00026	3.85	0.000

우선 T-test를 하기에 앞서 두 개의 독립집단이 분산이 동일한지에 대해서 살펴보았다. 그 결과 분산이 동일하지 않는 이분산으로 나타났다. 따라서 두 개의 집단은 분산이 동일하지 않다는 가정하에 T-test를 실시하였다.

[표 4-17]의 결과를 통해 유의수준 0.05하에 신경망모형과 혼합모형Ⅰ인 Decision Tree-Neural Network와 혼합모형Ⅱ인 Genetic-Neural Network 모형간의 차이는 유의확률 0.05하에 통계학적으로 의미가 있는 차이임을 알 수 있었다.

위의 내용을 종합해 보면 안과 자료에 대해서 단일모형과 혼합모형Ⅰ모형에 비해 신경망모형의 단점을 유전자 알고리즘으로 보완한 하이브리드 Genetic-Neural Network 모형이 더 나은 분류력을 갖았으며, 이러한 분류력의 차이가 통계학적으로 의미가 있는 차이라고 말할 수 있겠다.

5. 결론 및 고찰

본 연구는 병원의 고객이라 얘기할 수 있는 건강증진센터의 자료를 토대로 하여 개별기법을 통합한 하이브리드 모형을 제시하여 예측력의 향상을 꾀하였으며 접근 시도는 아래와 같다.

첫째, 의사결정나무의 알고리즘은 일반적으로 Small Disjunct가 아닌 Large Disjunct에 잘 적합하도록 치우치는 경향이 있으므로 Small Disjunct인 단점을 신경망모형으로 보완한 하이브리드 Decision Tree-Neural Network 모형이 단일 모형보다 더 나은 분류율을 보여주었으며 이들 간의 분류력의 차이가 통계적으로 의미가 있음을 확인하였다.

둘째, 신경망 모형시 초기값이 지역적 최소값에 가까우면 모수 추정치는 실제 데이터를 정확하게 추정하지 못할 가능성이 많다는 단점을 유전자 알고리즘으로 보완한 하이브리드 Genetic-Neural Nework 모형이 단일모형과 하이브리드 Decision Tree-Neural Network 모형보다 더 나은 분류율을 보여주었으며 이들 간의 분류력의 차이가 통계적으로 의미가 있는 차이임을 알 수 있었다.

본 연구에서의 한계점은 다음과 같다. 첫째, 하이브리드 모형은 개별기법의 모형을 결합하여 사용함으로 사용하기에 불편한점이 있다는 것이다. 둘째, 신경망 모형과 결합되는 하이브리드 모형은 정확성면에서 좋지만 해석상의 어려움을 가지고 있다는 것이다. 셋째, 본 연구에서 시도하지 못했던 의사결정나무의 Small Disjunct 단점을 추가적으로 유전자 알고리즘, SVM, 퍼지 등을 결합한 하이브리드 모델로 최적화하여 비교해 보는 것이다. 넷째 두 개의 혼합모형 중 어느 모형이 더 우수하다고 얘기하는 것은 어렵지만 향후 연구 분야에서 좀더 구체적으로 접근한다면 두개 모형의 비교가 가능할 것으로 보인다.

향후의 연구 방향은 첫째, 신경망 모형 분석시 보다 다양한 신경망 기법을 사용해 보는 것이다. 둘째, 분류 정확도와 해석을 동시에 고려 할수 있는 의사결정나무와 다른 모형과의 결합을 시도해 보는 것이 앞으로의 좋은 연구 과제가 될 것으로 생각된다.

참고 문헌

[1] 박찬욱, 데이터베이스 마케팅, 연암사, 18-24, 1996

[2] 배장섭, 혼합모형을 이용한 예측방법에 관한 연구, 연세대 대학원, 2003

[3] 장남식, 홍성완, 장재호, 데이터 마이닝, 대청미디어, 156-157, 2000

[4] Berry, Michael J. A., and Gordon Linoff, *Data Mining Techniques for Marketing Sales and Customer Support*, Wiley&sons, Inc. 1997

[5] Breiman, L., Friedman J., Olshen, R. A. and Stone, C. J., "Classification and Regression Trees", Chapman&Hall, 1011-1033, 1984

[6] Contana, D. J., "A Weighted Probabilistic Neural Network", Advances in Neural Information Processing Systems 4, 1110-1117, 1992

[7] David E. and Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning,* ADDISON_WESLEY PUBLISHING COMPANY, INC, 1989.

[8] Deborah R. and Carvalho "New results for a hybrid Decision Tree/Genetic Algorithm for Data Mining", Comendador Franco, 223-240, 2000

[9] Fletcher, D. and Goss, E., "Forecasting with neural networks : An application using bankruptcy data", Information and Management, 24, 159-167, 1993

[10] Freeman, J., "Simulating a basic genetic algorithm", The Mathematica Journal, 52-56, 1993

[11] Friedman, J., *Data Mining and Statistics; What's The Connection?*, Standford University, 1992

[12] Gately, E., *Neural Networks for Financial Forecasting,* John Wiley & Sons, Inc., 1996

[13] Greg Rogers and Ellen Joyner, *Mining data for Health care quality improvement*. SAS Institute, Inc., Cary, NC., 1998

[14] Han, Ingoo , Park, Cheolsoo and Kim, Chulhong, "Bankruptcy Predictions for Korea Medium-sized Firms using Neural Networks and Case Based Reasoning", Proceedings of Korea Management Science Institute Conference, 87-92, 1995

[15] Heistermann. J., "Learning in Neural Nets by Genetic Algorithms", Proceeding of Parallel Processing in Neural Systems and Computers(ICNC), Eckmiller, R. et al.(eds.), Elsevier. 165-168, 1990

[16] Holte, R.C., Acker, L.E. and Porter, B.W. "Concept Learning and the Problem of Small Disjuncts", Proceedings of the Eleventh International Joint Conference, 813-818, 1989

[17] Huang, J., Bala, J. and Vafaie, H., "Hybrid using Genetic Algorithms and Decision Tree for Pattern Classification" Proceedings of the Eleventh International Joint Conference, 250-278, 1995

[18] Lawrence, O. H., "Decision Tree Learning on Very Large Data Sets", Proceeding of IEEE International Conference,76-92, 1998

[19] Lee Kun chang , Han Ingoo and Kwon Youngsig, "Hybrid Neural Network models for bankruptcy predictions", Decision Support Systems, 18 , 63-72, 1996

[20] McFadden, D., "A Comment on Discriminate Analysis versus Logit Analysis", Annuals of Economics and Social Measurement 5, 511-523, 1976

[21] Press, S. J. and Wilson, S., "Choosing Between Logistic Regression and Discriminant Regression", Journal of the American Statistical Association, 73, 335-352, 1978.

[22] Rajeev, R. and Sim, K., "A Decision Tree Classifier that Integrates Building and Pruning", Data Mining and Knowledge Discovery, 4, 315-344, 2000

[23] Samuel E. B. and Ciril, K., "Using k-nearest-neighbor classification in the leaves of a tree", Computational Statistics & Data Anaylsis 40, 27-37, 2002

[24] Schwarrz, G., "Estimating the dimension of a model", Annals of Statistics, 6, 461-464, 1978

[25] Shin, K. S. and Han, I., "Bankruptcy Prediction Modeling Using Multiple Neural Networks Models", Proceedings of Korea Management Science Institute Conference, 240-265, 1998

[26] Steinberg, D., "Hybrid Cart-Logit and Cart-Neural Nets for Classification

and Regression", American Statistical Association, 562-574, 1999

[27] Weiss, G.M. and Hirsh, H. "A Quantitative Study of Small Disjuncts", Proc. of seventeenth National Conference on Artificial Intelligence. Austin, Texas, 665-670, 2000

[28] Weiss, S. M. and Kulikowski, C. A., *Computer Systems That Learn*, Morgan Kaufmann, 1991

[29] Wong, F. and Tan C., "Hybrid neural, genetic and fuzzy systems," In Deboeck, G. J. (ed.), Trading on the edge, New York:John Wiley, 245-247, 1994

ABSTRACT

A Study on Improving Classification Predictive Power using the Hybrid Model

Kim, Bong-Sob
Department of Biostatistics and Computing
The Graduate School
Yonsei University

This study proposes a hybrid decision tree-neural network model and a hybrid genetic-neural network model to improve classification predictive power. The purpose of this dissertation is to supplement defect and expand advantage about simple models.

In this research paper, two perspective of hybrid models are introduced. First perspective suggests hybrid model Ⅰ, decision tree algorithms have a bias towards generality that is well suited for large disjuncts, but not for small disjuncts. This makes them a promising solution for the problem of small disjuncts, which tries to supplement lacking parts of decision tree with neural network. Second perspective suggests hybrid model Ⅱ, artificial intelligence applied to complex nonlinear optimization problem have often resulted in inconsistent and unpredictable performance. This makes them a promising solution for the problem of local minimum, which tries to supplement lacking parts of neural network with genetic algorithms.

This study was researched based on screening test data accumulated from

1994 to 2002. Hybrid decision tree-neural network and hybrid genetic-neural network were applied to real data in this research. Result of hybrid decision tree-neural network and hybrid genetic-neural network was improved in classification predictive power with single model.

Key words : Logistic Regression, Decision Tree, Neural Network, Genetic Algorithm, Hybrid Model, Data Mining