신경회로망을 이용한 Health-Data 최적화 처리기법에 관한 연구

연세대학교 대학원

의 과 학 과

김 도 성

신경회로망을 이용한 Health-Data 최적화 처리기법에 관한 연구

지도교수 유 선 국

이 논문을 석사 학위논문으로 제출함

2008 년 12 월

연세대학교 대학원

의 과 학 과

김 도 성

김도성의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2008 년 12 월

감사의 글

2006년 여름, 처음으로 연구실의 문을 두드릴 때의 기억이 아직 남아있는데, 두 번의 겨울과 두 번의 여름이 지나 벌써 이 글을 쓰고 있습니다. 많은 웃음과 즐거움, 그리고 고됨과 머리 쥐어뜯기가 있었던 이 시간동안 도와주신 많은 분들께 감사의 말씀을 전하고자 합니다. 제일 먼저 석사과정 중 많은 지도와 가르침을 주신 유선국 교수님께 무한한 감사의 말씀을 올립니다. 그리고 논문 심사과정의 여러 가지 일정에 바쁘신 와중에도 단 한 번도 빠지지 않고 항상 참석해주신 심사위원 지선하 교수님과 김경환 교수님, 그리고 의학공학교실에서 몸담고 있으면서 자주 뵙지는 못했지만 인사드릴 때마다 환한 웃음으로 응답해주셨던 김남현 교수님, 김덕원 교수님, 서활 교수님, 박종철 교수님께도 감사의 말씀 올립니다.

연구실 최고 맏형으로서 냉철한 지성과 따뜻한 감성을 갖고 계신 정석명 박사님. 연구실의 실무를 책임지신 김동근 박사님. course-work이 끝난 후부터는 뵙기가 정말 힘들지만, 해주시는 모든 말씀이 뼈가 되고 살이 되었던 박순만 박사님. 한 학기밖에 같이 하지 못해 아쉽지만, 기분 좋은 미소와 해박한 지식을 마음속에 남겨주신 원윤재 박사님. 내가 졸업하면서 일거리만 늘려주고 가는 것 같아 맘이 찜찜하지만, 그래도 시원시원한 웃음을 보여주는 이충기 박사. 코딩 때문에 받았던 스트레스의 감소와 댕기효과로 인해 나날이 젊어지고 있는, 하라주쿠 2회 방문의 동반자 김정채 박사. n차원의 새로운 세계로 들어서게 되지만, 꿋꿋하게 해내실 것 같은 이미희 박사님. 센터에 가면 환한 웃음과 다양한 먹거리로 무한한 기쁨을 주셨던 사무원 곽은선 님. 작은 체구에서 쏟아져 나오는 현란하고도 폭발적인 언어구사력으로 연구실 생활 중 대화의 즐거움을 새삼 일깨워 주었지만, 향후 몇 년간의 삶이 심히 걱정스러운 도윤이형. 결혼한 것이 제일 부럽고, 가격비교사이트에게 문 닫기를 권유하고 싶게 만드는 SMD 납땜의 황태자 인호형. 늦게 배운 연애질이 무섭다고, 아무도 상상도 못한 연애 이벤트를 너무나 자연스레 유감없이 보여준 상용이. 삶이 지루할 때면 화통한 웃음과 끊임없는 즐거움을 제공하고, 연구실의 몇몇 인사로 인해 탄생한 '씨・엘・유・비' 정신을 지속적으로 계승하고 있는 용귀. 말로서 사람을 재밌게 하는 법을 알아 하루에 한번 이상 큰 웃음을 주었고, 졸업 논문 쓰면서 이것저것 귀찮게 하면서 제일 많이 괴롭혀 미안했던 한규. **"어디서도 만날 수 없고, 두 번 다시 만날 수 없을 것 같은 멋진 대학원 동기들이였습니다!!"** 아무도 범접할 수 없는 그 분과 동맹 체제를 구축하였지만, 항상 불안에 떨며 각오하고 있는 동규. 1년 사이 말보다 주먹이 앞서기 시작했고, 날로 늘어가는 푸석함과 고시원라이프의 삭막함이 심히 걱정스러운 성혜. 내

가 자리를 옮기고 나서부터 괴롭히는 횟수와 가중치가 exponential로 증가하였지만 능글능글하게 모두 받아낸 광수. 캐스퍼처럼 언제 왔다가 언제 가는지 모를 정도로 소리 없이 돌아다녀서 사람 기겁시키는데 재주 많은 수호. 수업도 다르고 일 하느라 한 학기도 넘게 얼굴도 못 보다가 엉뚱한 곳에서 오랜만에 보게 된 은정이. 생각할수록 맘이 짠해지지만, 분명 더 좋은 길이 기다리고 있을 것이라 확신하는 현택이. 본의 아니게 사지로 몰아 넣어버린 것 씁쓸한 마음이 남고, 항상 밝고 강해보이지만 아직은 더 많은 노력을 주문하고 싶은 주현이. 나에게 어마어마한 폭탄을 넘기고 간 윤정이. RFID만 남기고 떠나가 버린 정진이. 처음 연구실 알아볼 때 너무나 친절(?)하게 연구실 추천해 주었던 국진이형. 졸업한 후 더 친해진 것 같은 시사프로 전문 출연 연예인 영재. 모니터 앞에서의 올바른 연구생의 자세에 대해 알려준 민규형. 나이를 초월한 외모와 스타일을 자랑하는 봉문이형 등 모든 연구실 사람들에게 감사를 전합니다. 또한 오랜만에 만날 때면 항상 신촌에서 만나서 절대 빠질 수 없도록 만들어준, 고마운(?) 연세대학교 의공학과 16대 학생회 멤버들과 00학번 동기들, 그리고 내 오랜 친구들인 '골프GTI오너이자 마이크로소프트 대리 초특급 진급자인 훈남 친구'라고 해달라고 한 희수, 영경양과 새로운 인생을 시작한 the First 유부남 경종이, 잠시나마 아메리칸 워터에 흠뻑 젖어 온 뉴요커 윤수, 올림픽 개최년도에만 얼굴 보는 것 같은 진우 등 모든 친구들에게도 감사의 마음 전합니다.

그리고 아들에게 매일 따뜻한 아침 먹여 보내시려고 2년 동안 수고스러운 아침 시간을 마다하지 않으신 사랑하는 어머니. 항상 믿어주시는 아버지. 바쁘다고 애들이랑 잘 놀아주지 못해 미안했던 큰누나와 술 한 잔 같이 해 본 적이 언제인지 기억도 잘 안나 죄송스러운 매형, 가난한 동생에게 단비와 같은 물질적 혜택을 게을리 하지 않아서 항상 고마운 작은누나. 그리고 늦은 밤에 지친 몸으로 집에 들어가면 엄청난 속도로 마중 나와, 강렬한 애교와 재롱으로 피로 대신 웃음과 활력을 채워주었던 우리 이쁜 조카들, 윤자매 지유・지효에게도 감사의 마음을 전합니다.

마지막으로 신경도 많이 써주지 못 했고, 많은 시간 함께하지 못해 미안하지만, 3년 넘게 변함없이 항상 따뜻한 웃음과 편안한 휴식, 그리고 진심어린 위로와 격려가 되어준 여자친구 진미에게도 감사의 마음을 전합니다.

2008년 12월

김 도 성 드림.

차 례

그림 차례

표 차례

국문요약

신경회로망을 이용한 Health-Data 최적화 처리기법에 관한 연구

당뇨병은 인슐린 작용의 부족에 의한 만성 고혈당 증세와 그 외에 여러 가지 대사이상 합병증을 수반하는 질환이다. 본 연구에서는 당뇨병의 발병가능성을 예측함에 있어서 유전 알고리즘을 이용하여 발병진단과 밀접한 관계가 있는 데이터를 선별한 후, 신경회로망 기법을 이용하여 가장 최적화된 모델을 찾아 그 정확도를 높일 수 있는 방법에 대하여 연구하고자 한다.

230명의 환자데이터와 3945명의 정상데이터, 총 4175명의 데이터를 각각 3:1의 비율로 나누어 Training Data와 Test Data로 구성하였다. 실험에 사용한 데이터 종류는 Adiponectin, Triglyceride, HDL-Cholesterol, Sex, Age, Waist, BMI, HTN, Smoke, Exercise, Alcol이다. 이 데이터 중에서 질병 예측에 최적화된 데이터만을 유전알고리즘을 통해 추출한 후, 신경회로망 모델을 이용하여 발병가능성을 예측해본 결과, 위의 11가지 데이터 전체를 Initial-Feature로 선정한 경우에는 67.8%의 정확도를 보였고, 혈액검사를 통해 구할 수 있는 데이터를 제외한 나머지 8개의 데이터를 Initial-Feature로 선정한 경우에는 63.9%의 정확도를 나타내었다.

이는 대조군으로 설정한 Logistic Regression Model의 결과보다는 약간 낮은 정확도를 보였는데, 이는 환자데이터와 정상데이터간의 수치적 편중에 의한 것으로 판단된다. 두 데이터의 양을 동일하게 구성한 실험에서는 각각 약 80% 정도의 월등히 높은 정확도를 보였기 때문이다.

추후 추가적인 환자데이터가 확보되고, 다양한 신경회로망 모델을 개발되어 매우 우수한 정확도를 보인다면, 의료 현장에서도 적용 가능한 예측모델로서 정립될 수 있을 것이라 판단된다.

핵심되는 말 : 당뇨병, 신경회로망, 유전 알고리즘

신경회로망을 이용한 Health-Data 최적화 처리기법에 관한 연구

<지도교수 유 선 국>

연세대학교 대학원 의과학과

김 도 성

Ⅰ. 서 론

당뇨병(糖尿病, Diabetes Mellitus)은 인슐린 작용의 부족에 의한 만성 고혈당증을 특징으로 하면서 여러 특징적인 대사 이상을 수반하는 질환이다. 인슐린은 주로 탄수화물 대사에 관여하므로, 당뇨병은 탄수화물 대사의 이상이 기본적인 문제이나, 이로 인해 체내의 모든 영양소 대사가 영향을 받게 되므로, 또한 총체적인 대사상의 질병이라고 할 수 있다.

당뇨병은 인슐린 부족이나 인슐린에 대한 세포 저항으로 인한 고혈당이 근본적인 원인으로 고혈당이 지속됨에 따라 대사상의 변화가 초래된다. 인슐린 작용이 저하되면 과다한 당을 섭취하였을 때 일정한 혈당 수준을 유지하는 내당능력이 감소하므로 혈당이 높아지고, 따라서 당을 소변으로 배설하는 포도당 낭비 현상을 보인다. 당뇨병 초기의 특징적인 증상으로는 다뇨(polyuria), 갈증(polydipsia), 식욕항진(polyphagia), 체중감소를 들 수 있다. 임상적인 증상으로는 요를 통한 당의 배설(glucosuria), 고혈당(hyperglycemia), 결구 내당능 검사의 이상(abnormal glucose tolerance test), 무력증(asthenia) 등이 있다.

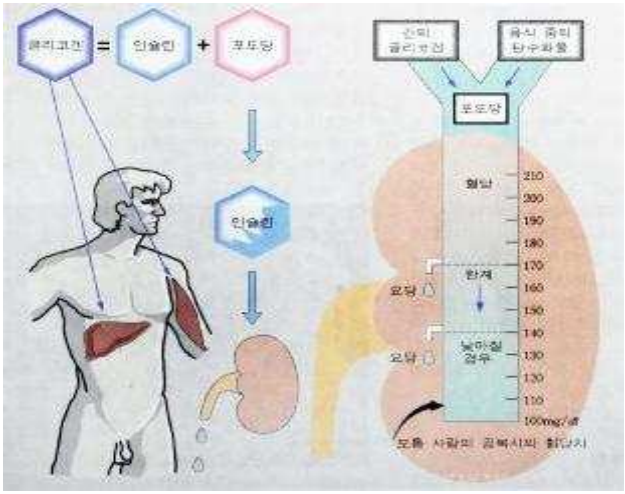


Figure 1. 당뇨병 발생기전.

당뇨병은 크게 제1형 당뇨병과 제2형 당뇨병으로 분류한다. 제1형 당뇨병은 췌장 β 세포의 파괴성 병변에 의해 인슐린이 결핍되어 생기는 당뇨병으로, 다시 면역 매개형, 원인 불명형으로 나뉜다. 제2형 당뇨병은 인슐린 분비 저하와 인슐린 저항성으로 인해 생기며, 이 두 가지 인자의 관여 정도에 따라 인슐린 분비부족 우위 당뇨병과 인슐린 저항성 우위 당뇨병으로 나뉜다. 한국인에게 많은 제2형 당뇨병의 경우 인슐린저항성으로 생기며, 이 때문에 초기에는 췌장에서 인슐린 분비를 늘림으로써 별다른 증상이 없지만, 췌장의 능력에 한계가 오게 되면 증상이 나타나기 때문에 진단이 늦어진다. 보통 이 기간이 5년 이상으로 알려져 있으며 따라서 당뇨진단 당시에 합병증이 이미 와 있을 수 있기 때문에 합병증에 대한 검사가 필요하다.

당뇨병의 진단은 정맥혈의 혈장 포도당 농도를 기준으로 이루어진다. 서로 다른 날 2번 검사하여 공복시(보통 검사 전 8~12시간 금식 상태) 혈장 포도당 농도가 둘 다 126 mg/dL 이상이면 당뇨로 진단한다. 일반적으로 정상 혈장 포도당 농도는 보통 100 mg/dL 미만이며, 100~125 mg/dL 사이를 경계형 당뇨병라고 한다. 경계형 당뇨병의 경우 당뇨로 진행하는 경우가 25~30%로 알려져 있으므로, 자주 혈당

을 측정하여 당뇨병으로의 진행여부를 조기에 파악해야 한다.

당뇨병의 치료에서 가장 중요하고, 급성 또는 만성 합병증의 예방법은 혈당을 철저하게 조절하는 것이다. 처음에는 운동요법, 식이 요법으로 조절하고, 심화되면 먹는 약이나 인슐린 주사, 혹은 인공 췌장이나 췌장 이식술 등을 시도할 수 있다.

그러나 모든 질병이 그러하듯이, 당뇨병도 치료보다는 예방하는 것이 제일 중요하다. 당뇨병을 예방하기 위해서는 과식과 약물남용을 삼가하고, 담도나 담낭, 췌장에 생긴 감염증을 예방하고 조기치료를 해야 한다. 또한 정기적으로 혈당 및 요당 검사를 통해 조기 발견하는 것이 건강생활을 유지하는 가장 지름길이라 할 수 있다.

또한 당뇨병을 조기발견하기 위하여 통계학적·역학적 측면에서의 연구도 시도되고 있다. 그 중 가장 보편적인 연구방법은 당뇨병뿐만 아니라 뇌졸중과 같은 다양한 질병에 대하여 적용되고 있는 방법인데, 정상인과 환자들의 성별, 연령, BMI 지수, 허리둘레, 고혈압, 흡연여부, HDL 콜레스테롤 등의 각종 데이터를 바탕으로 Multi-variable Logistic Regression Model과 같은 통계·역학적인 방법을 적용하여 그 질병의 발병가능성을 예측하는 연구가 지속적으로 시도되고 있다.[1,2] 이와 더불어 최근 시도되고 있는 방법은 인공지능 시스템 이론 중 하나인, Neural Network(신경회로망)를 이용한 방법이다. Neural Network은 인체 사고의 주체인 뇌의 가장 기본단위인 뉴런의 동작 및 기능을 공학적으로 모델화한 인공지능 이론으로서, 패턴 인식 분야나 분류, 제어 등과 같은 다양한 응용분야에서 적용되고 있다.

Neural Network를 의학 분야, 특히 당뇨병과 같은 질병의 발병가능성을 판단하고자 하는 연구는 이미 존재하고 있다.[3,12,13,14,15] 그러나 선행 연구에서 시도된 방법은 성, 연령, 혈압, 키, 몸무게 등과 같은 기본 신체 데이터뿐만 아니라 Triglyceride, HDL 콜레스테롤, Adiponectin, BMI 지수, 흡연, 운동 등과 같은 방대한 종류의 데이터를 모두 Neural Network 모델에 입력시키는 방법을 사용하여 데이터

간 중복되는 데이터 종류도 있고, 불필요한 데이터가 입력되는 단점이 존재하였다.

따라서 본 연구에서는 당뇨병의 발병가능성을 예측함에 있어서 생물의 유전법칙을 기반으로 발전한 인공지능 시스템 이론 중 하나인 Genetic Algorithm(유전 알고리즘), 또한 신경계의 기본단위인 뉴런의 동작을 모델화하여 발전한 Neural Network(신경회로망) 기법을 이용함에 있어서, 가장 최적화된 모델을 찾아 그 정확도를 높일 수 있는 방법에 대하여 연구하고자 한다.

Ⅱ. 재 료 및 방 법

1. Genetic Algorithm

Genetic Algorithm(유전 알고리즘, 이하 GA)은 적자생존과 유전의 메커니즘을 바탕으로 하는 탐색 알고리즘이다. 다시 말해 주어진 환경에 잘 적응하는 유전자만을 선택하고, 교배하고, 경우에 따라 돌연변이도 하며 다음 세대에 우수한 유전 형질이 전달되도록 하는 것이다. 따라서 진화가 거듭될수록 주어진 환경에 더 적합한 유전자들만이 남아있게 된다.

가. Genetic Algorithm 이란?

Genetic Algorithm은 생물의 진화과정, 즉 자연선별과 유전법칙을 모방한 확률적 탐색기법이다. 이 알고리즘은 1975년 Holland의 연구 “Adaptation in Natural and Artificial System”에서 처음으로 소개되었다. 그는 자연시스템의 한 메커니즘으로 생물의 진화과정을 추상화하여 인공시스템을 설계하고자 했다. 그 후 20여 년 동안 GA의 이론과 응용에 관하여 활발한 연구가 이루어져 왔다. GA의 가장 큰 특징은 복수개의 잠재해들로 이루어진 해의 집단(Population)을 운용한다는 것이다. 이러한 해집단에 자연선별과 유전법칙의 메커니즘을 적용하여 세대(Generation)를 진행시키면서 해공간을 탐색한다.

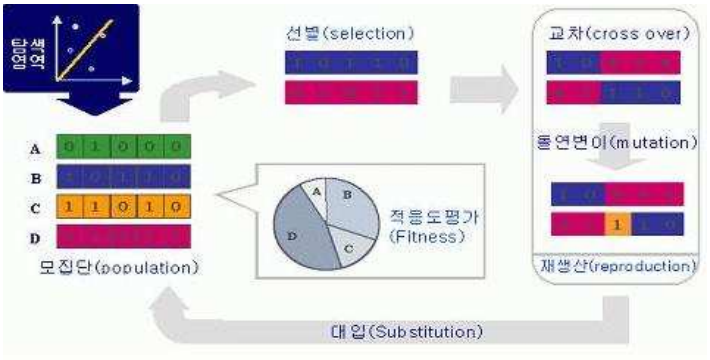


Figure 2. Genetic Algorithm

그리고 GA은 그 개념과 이론이 단순하고, 해의 탐색능력이 우수하며, 특히변수와 제약이 많은 대형 수리문제를 해결하는데 적합한 기법이고, 또한 모형에 대한 유연성이 높아 제약 첨가나 목적함수의 변경이 용이하다는 장점을 가지고 있다.

GA은 문제의 잠재해를 표현한 개체들로 이루어진 모집단으로 시작한다. 모집단은 각 세대에서 일정수의 개체를 유지하고 개체의 적응도를 평가하여 다음 세대에 생존할 개체들을 확률적으로 선별한다. 선별된 개체들 중 일부의 개체들이 임의로 짝을 지어 교차하여 자손을 생성한다. 이때 교차에 의해 부모의 유전자가 자손에게 상속되고 돌연변이가 발생할 수 있다. 자손은 부모로부터 좋은 유전형질을 상속받는다고 가정함으로써 다음 세대의 잠재해들은 평균적으로 전 세대보다 더 좋아진다는 원리이다. 이러한 진화과정은 종료조건을 만족할 때까지 반복된다. GA의 요소로는 유전자 표현, 초기 모집단, 적응도 평가, 선별, 교차, 돌연변이, 유전 파라미터 등이 있다

나. Genetic Algorithm Procedure

① 유전자 표현 (Gene Representation)

GA을 표현하는 첫 단계로 잠재해를 개체(유전자 표현)로 표현한다. 이 유전자 표현(Gene Representation)은 GA이 수행되는 동안 적응도(Fitness Function)와 유전연산과정에 영향을 미치므로 문제의 특성을 잘 반영할 수 있어야 한다.

② 초기모집단 (Initial Population)

GA은 개체들로 구성된 모집단을 운영하므로 초기에 모집단이 생성되어야 한다. 초기의 모집단을 만드는 방법으로는 문제의 특성을 이용한 기존의 발견적 방법이나, 임의생성방법이 있다. 일반적으로 GA에서는 발견적 방법에 의한 초기해보다는 임의생성방법을 사용한다.

③ 적합도 평가 (Evaluation Fitness)

적합도(Fitness)는 자연개체의 생존능력을 나타내는 것으로 최적화

문제에서 적합도는 목적함수(Objective Function)에 의해 평가될 수 있다. 적합도를 평가하는 함수로는 최적화문제의 목적함수 자체를 흔히 사용한다.

④ 선택 (Selection)

선택은 적자생존의 자연법칙에 기초한다. 즉 환경에 대한 적합도에 의해 현 세대의 모집단으로부터 다음 세대에 생존할 개체를 선택하는 과정이다. GA에서 선택은 모집단의 다양성과 선별압력이 조화를 이룰 수 있어야 한다. 즉, 적합도가 높은 우수개체가 열성개체보다 생존확률이 아주 높은 강한 선택압력은 모집단의 개체들을 조기 수렴시키는 경향을 갖는데, 이는 모집단의 다양성을 약화시켜, 해공간의 다양한 탐색을 방해하는 결과를 가져온다. 선택방법에는 Roulette Wheel Selection, Ranking Selection, Tournament Selection, Elite Strategy 등이 있다.

⑤ 교차 (Crossover)

교차는 유전자 알고리즘의 대표 연산자로 두 부모가 갖는 유전자를 조합하여 자손을 생성하는 과정으로 좋은 해를 이용하는 역할을 한다. 이를 위해서 교차는 부모의 좋은 형질이 가능한 파괴되지 않고 자손에게 상속될 수 있어야 한다. 교차방법에는 일점교차, 이점교차, 순서교차, 균등교차 등 다양한 방법이 있다.

⑥ 돌연변이 (Mutation)

돌연변이는 개체에 새로운 유전자가 생성되는 것으로, 한 개체에서 아주 작은 수의 유전자를 임의로 변화시키는 과정이다. 유전자 알고리즘에서 돌연변이는 해 공간을 다양하게 탐색하는 역할을 한다. 돌연변이 방법에는 점 돌연변이, 삽입, 교환, 역치 등 다양한 방법 등의 일반적인 돌연변이 방법과 세대수가 증가되면서 해의 개선이 일어나지 않으면 돌연변이율을 증가시켜주는 변동 돌연변이가 있다.

⑦ 유전 파라미터

GA에 사용되는 파라미터로는 모집단의 크기(Population Size), 교차율(Crossover Probability), 돌연변이율(Mutation Probability), 종료조건 등이 있다. 모집단의 크기는 모집단을 이루는 개체의 수를 의미하며, 교차율과 돌연변이율은 각 개체가 교차 및 돌연변이 될 확률을 나타낸다. 알고리즘의 종료조건으로는 진행된 세대수 또는 생성된 개체 수, 해의 개선이 이루어지지 않고 진행된 세대수, 생성된 개체 수, 계산 소요시간 또는 목표로 하는 적합도 및 결과 값 등이 사용된다. 이러한 유전 파라미터는 문제의 특성과 알고리즘의 구조에 따라 해결하고자 하는 문제의 특성을 잘 반영하면서, 가장 빠르고 효율적으로 해공간의 탐색이 가능한 값들을 실험을 통하여 결정할 수 있어야 한다.

Table 1. Genetic Algorithm Procedure

단계	내 용
1	종료조건 설정
2	최초 유전자개체 모집단 구성
3	최초 모집단의 유전자개체에 대한 적합도 평가
4	교차율에 의해 모집단에서 두 유전자개체 선택
5	선택된 두 개체에 대한 교차 실시
6	돌연변이율에 의해 임의개체 변형
7	교차 및 돌연변이 된 유전자개체의 적합도 평가
8	종료조건 만족시 계산 종료. 불만족시 단계 4로 회귀.

다. Genetic Algorithm의 응용분야

① 설계문제 : VLSI 레이아웃 설계, 컴퓨터의 캐시 시스템 설계, 통신 네트워크, 엔진 설계, IC 설계, 디지털 필터 설계, 형상 설계, 퍼지 제어기 설계 등

② 스케줄링 문제 : 태스크 할당(멀티프로세서 컴퓨터), 작업

스케줄링 등

③ 조합에 의한 최적화 문제 : 자동차의 경로 최적화, 순회 세일즈맨 문제, 금융거래 의사결정, 유전자 정보 해석, 그래프 분할 문제 등

④ 제어 문제 : 프로세스 제어, 로봇 제어 등

⑤ 기 타 : 유전자 프로그래밍(프로그램의 자동 생성), 화상 복원, 목표물 검출, 시스템의 동정, 단백질의 구조 추정, 데이터베이스 검색, 작곡 등[4,5,6,7,8]

2. Neural Network (신경회로망)

Neural Network(신경회로망, 이하 NN)란 생물의 신경계의 기본단위인 neuron을 모델화한 인공적인 신경망이다. NN는 복잡한 유형을 찾아주는 컴퓨터 프로그램이자, 대용량의 데이터로부터 예측모델을 만들어주는 기계적 학습 알고리즘이다.

가. 생물학적 신경망과 인공신경망

신경망에 관한 연구는 인간의 두뇌와 신경 세포 모델에 대한 연구에서 시작되었다. 신경 시스템에서 가장 기본적인 단위는 뉴런이라는 세포이다. 각각의 뉴런은 신경 시스템에서 여러 가지 기능적 역할을 담당한다.

(1) 생물학적 신경망

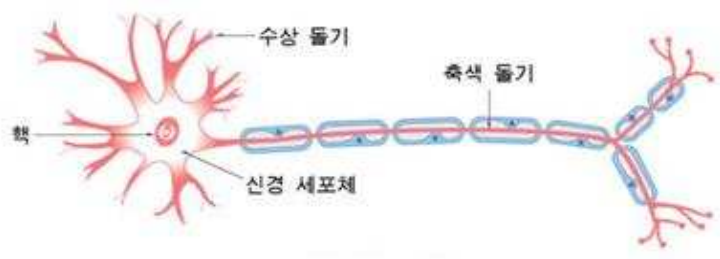


Figure 3. Neuron : 신경계의 기본단위

뇌는 다수의 뉴런이 서로 연결된 신경망으로 구성되어 있다. 대뇌나 소뇌의 피질에는 이러한 뉴런이 1㎣안에 수 만개가 있다. 뉴런의 기본적인 기능은 정보의 수용, 연산처리, 출력의 전송 등의 3가지이고, 뉴런의 형태나 크기는 여러 가지가 있으며, 기본적으로는 체세포(soma)와 체세포로부터 나온 많은 돌기로 구성되어 있다. 체세포는 신경의 중앙에 위치하며 직경 5~100㎛의 크기이다. 체세포로부터 나온 돌기는 Axon(축색돌기)라 불리는 한 개의 가느다란 섬유와 Dendrite(수상돌기)라 불리는 나무의 가지처럼 넓게 펴져 있는 비교적 두껍고 짧은 다수의 돌기로 나뉜다. Axon은 체세포에 붙어있으며 전기적으로 활성화되고, 뉴런에 의해 발생되는 펄스를 다른 뉴런들에게 전달하는 기능을 한다. Dendrite는 다른 뉴런과 연결되어 입력 신호를 받아 연산을 수행한 후 체세포에 보내는 역할을 한다. 즉, 체세포는 Dendrite로부터 받은 신호 및 체세포 내에 직접 입력되는 신호를 펄스 신호로 변환되는데, Axon은 펄스 정보를 다른 뉴런에 전달하는 케이블 역할을 하는 것이다. Axon의 끝부분은 가느다란 가지로 나뉘어져 있으며, 다른 뉴런의 Dendrite와 체세포에 접속하는 Synapse(시냅스)라 부르는 특별한 연결체를 가지고 있다. 뉴런 간의 정보 교환은 모두 시냅스를 통하여 이루어지며, 정보의 전달 방향은 항상 단방향으로 이루어진다.

(2) 인공 신경망

인공 신경망의 뉴런과 생물학적인 뉴런과는 매우 큰 차이를 가지고 있다. 우선 생물학적인 뉴런은 주위의 뉴런들과 고밀도의 연결을 가지고 있다. 즉, 두뇌에서 뉴런의 경우 1000개에서 100,000개의 다른 뉴런들과 연결되어 있는데, 이에 반해 인공 신경망에서는 이것의 1% 가량의 연결성도 현재 수준으로는 원활하게 처리하기 어렵다. 또한 생물학적 뉴런들은 본래부터 전기 화학적(electrochemical), 즉 뉴런들 사이의 연결 강도는 전기적인 신호에만 의존하는 것이 아니고, 전기

적이고 화학적인 신호들에 의해 조정되지만, 현재까지 이러한 전기화학적인 신경망의 모델링은 이루어지지 못했다.

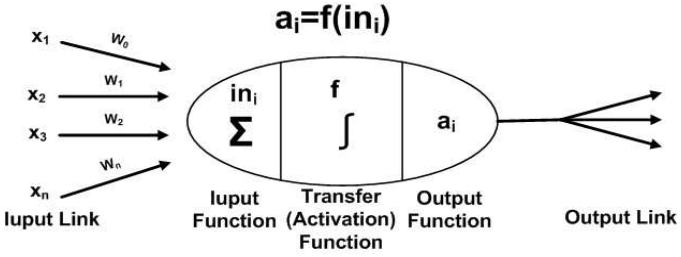


Figure 4. McCulloch-Pitts Neuron Model

이러한 인공신경망의 성질을 바탕으로 추상화된 뉴런 모델이 제시되었다. Figure 4는 인공신경망의 현대적 기원이 된 'McCulloch-Pitts Neuron(맥클로크-피츠 뉴런, 이하 MPN)' 모델이다.

MPN 모델은 N개의 입력을 갖고 각 입력의 신호를 x_1, x_2, x_3 … x_n , 출력을 y라 하면, 뉴런의 동작은 아래의 식과 같이 이산 시간의 차분 방정식으로 기술할 수 있다.

$$a_i = f(\sum_{i=1}^{N} w_i * x_i - \theta)$$

이 때, 함수 f는 계단함수(Step Function)로서, 다음의 조건을 만족한다.

$$f(x) = \begin{cases} f(x)=1 & (x>0) \\ f(x)=0 & (x \le 0) \end{cases}$$

MPN 모델의 각 입력 및 출력은 1 또는 0의 값을 가지며, 1은 뉴런이 흥분 상태(활동 전위 펄스를 생성한 상태), 0은 뉴런이 흥분하지 않은 상태(정지상태)에 각각 대응된다.

위의 식에서 w_i는 i번째 입력의 시냅스 연결 강도를 나타낸다. w_i가 양(+)이면 흥분성 시냅스, 음(-)이면 억제성 시냅스를 나타내며, 결합이 없으면 w_i=0이다. Θ는 임계값을 나타내며 $\sum_{i=1}^{N} w_i * x_i$의 값이 Θ보다 클 때만 뉴런이 흥분하여 펄스를 출력한다.

MPN 모델은 일종의 다수결로 출력을 결정하는 소자로, 다입력-1출력의 비선형 소자이다. 또한 MPN 모델은 입력 부위, 가합 기능 부위(입력 조합 및 가중치 부여), 임계값 기능 부위, 출력 부위 등 네 개의 기능 부위로 나뉜다. 입력 부위는 뉴런의 수상 돌기에 해당되며, 다른 뉴런의 신호를 시냅스로부터 받는 기능을 하고, 가합 기능 부위는 뉴런의 체세포처럼 활성적 정보를 가진 입력 신호와 억제적 정보를 가진 입력 신호를 조합하고 가중치(weight)를 부여한다. 임계값 기능 부위는 뉴런의 활동 전위가 임계값을 상회할 때 뉴런이 활성화되어 점화하고, 가합된 신호가 임계값에 미달되면 아무런 반응도 일어나지 않게 된다. 출력 부위는 뉴런의 축색 돌기에 해당되며, 체세포의 점화에 의해 발생되는 전기적 에너지를 통해 다른 뉴런으로 전달하는 기능을 한다.

이와 같은 수학적 모델로서의 뉴런 모델이 상호 연결되어 네트워크를 형성할 때 이를 Neural Network(신경회로망)이라고 하며, 이를 생물학적 신경망과 구별하여 특히 Artificial Neural Network(인공신경망)이라고도 한다.

나. Neural Network의 기본 구성 요소

Neural Network(이하 NN)은 연결성 모델(Connectionist model), PDP(Parallel Distributed Processing), 또는 뉴로모르픽 시스템(Neuromorphic system)이라고 불리는 것으로서, 그 기본 단위는 뉴런이 된다. NN 모델은 모두 단순한 계산소자의 연결을 통해 좋은 성을 늘 나타낸다는 것을 기본 가정으로 하고 있다. 따라서 NN은 음성, 이미지분석 등 계산량이 많고, 병렬성을 요구하는 문제에 적합한 모델이다.

Figure 5에서 처리기는 원으로 표시되었다. 각 시점에서 처리기 Ui는 $a_i(t)$라는 활성화값을 갖는다. 이 값은 출력값 $o_i(t)$을 생성하기위해 f_i를 통해 전달된다. 이 출력값 $o_i(t)$는 단방향으로 전달된다. 이때 각

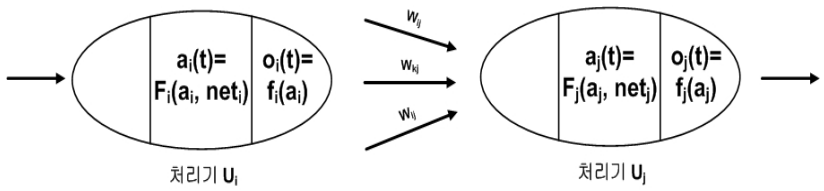


Figure 5. Neural Network 기본 구성 요소

연결선에는 첫 번째 처리기값 U_i가 두 번째 처리기는 U_j에 미치는 영향을 의미하는 가중치 w_{ij}를 갖게 된다. 이와 같이 한 처리기에 연결된 모든 입력값을 어떤 연산에 의해 결합하여, 현재의 활성화값에 따라 새로운 활성화 상태 F를 생성한다. 이 때, f와 F 함수의 예로서, 임계치 계단함수(Threshold Step Function)과 시그모이드 함수(Sigmoid Function)이 있다.

① 처리기 (Processing Units)

처리기는 NN 모델에서 매우 중요하고, 기본적인 단위이다. 처리기는 element 등의 보통 의미를 가진 작은 엔티티를 나타내는 패턴이다. N을 처리기의 개수라고 가정하면, i번째 처리기 U_i는 입력함수를 통해 다른 이웃 처리기에 보낼 출력값을 계산하는 일을 수행한다.

② 활성화 상태 (Activation State)

각 처리기들이 시간 t에서 시스템에서의 상태를 의미한다. 즉, 시스템이 시간 t에서 나타내는 처리기들의 활성화 패턴을 말하는 것이다.

③ 각 처리기에 대한 출력 함수 (Output Function)

처리기들은 이웃 처리기들에게 신호를 전달함으로써 서로 정보를 전달하게 된다. 이 때 신호의 강도는 이웃 처리기에 영향을 미치는 정도, 즉 활성화 정도에 의해 결정된다. 각 처리가 U_i와 연결된 출력함수 $f_i(a_i(t))$와 현재 활성화 상태 $a_i(t)$에 의해 출력신호 $o_i(t)$가 결정된다.

④ 각 처리기간의 연결패턴 (Connectivity Pattern)

연결패턴이란 시스템이 임의의 처리기에 대해 어떻게 반응하느냐를

의미하는 것으로, 일반적으로 모든 처리기로부터의 입력은 단순히 가중치 합에 의해 구한다. 이때 양의 가중치는 흥분성 연결을, 음의 가중치는 억제성 연결을 의미한다. w_{ij} 처리기 U_i로부터 처리기 U_j로의 연결강도를 의미하는 가중치값(Weight)이며 연결패턴이다. 이 때 절대값 w_{ij}는 연결강도(Connection strength)를 의미한다.

⑤ 전파 규칙 (Propagation Rule)

전파규칙은 한 처리기의 출력벡터 o(t)와 처리기로의 각 입력 형태에 대해 순수 입력값(net input)을 생성하기 위해 가중치와 결합화할 규칙을 말한다. net_{ij}를 처리기 i에서 처리기 j로의 순수 입력값이라 하고, net_i를 입력 패턴 I에 대한 입력벡터라 하면, net=Wo(t)로서 쉽게 구한다. 만약 복잡한 연결패턴을 갖는다면, 다음과 같은 복잡한 전파 규칙이 필요할 것이다.

$$net_j = \sum_{i=0}^{n} w_{ij} o_i$$

⑥ 활성화 규칙 (Activation Rule)

특정 처리기에 들어오는 각 순수 입력값들을 조합하여 그 처리기의 현재 상태로부터 새로운 상태를 구할 수 있는 규칙을 말한다. 즉, a(t)와 net_j로부터 활성화 규칙 F를 이용하여 새로운 활성화 상태를 구하는 규칙을 말한다. 만약 F가 일대일 함수라면,

$$a(t+1) = W_o(t) = net(t)$$

가 된다.

⑦ 학습 규칙 (Learning Rule)

한 처리기의 지식 변화는 인접된 다른 처리기에도 변형을 주는데, 보통 세 가지 형태로 영향을 주게 된다. ⅰ)새로운 연결 생성, ⅱ) 기존 연결의 상실, ⅲ) 기존 연결의 강도 수정 중에서 ⅲ) 기존 연결의 강도 수정이 많이 쓰인다. 이 연결강도는 경험적으로 변형되는데, 이와 같이 연결강도의 변화를 학습 규칙이라고 한다. 신경망 연구에 있

어 가장 중요한 것 중의 하나가 바로 학습 방법의 개발이었다. 지금까지 많은 학습 규칙들이 만들어져 왔으나, 그 기본은 언제나 주어진 입력에 대해 연결 가중치를 변화시키는 것이라 할 수 있다.

목적 패턴이란 주어진 입력 패턴에 대해 신경망이 출력해 주기 원하는 출력 패턴으로, 외부로부터 주어진 패턴을 말하는데, 이러한 목적 패턴을 사용하는 학습 방법을 감독 학습 방법이라 하고, 목적 패턴을 사용하지 않는 학습 방법을 비감독 학습 방법이라고 한다.

⑧ 환경 (Environments)

NN 모델에서의 환경은 입력 패턴에 대해 시간이 변화하는 확률론적 함수라고 본다.

다. Perceptron

Perceptron은 1957년 미국의 Rosenblatt가 단순히 패턴을 인식하기 위해 제안한 모델로 당시에는 상당한 주목을 받은 NN 모델이다. 즉, Perceptron은 입력 패턴이 두 개의 클래스 중 어느 하나에 속함을 결정할 때 주로 사용되는 모델이다.

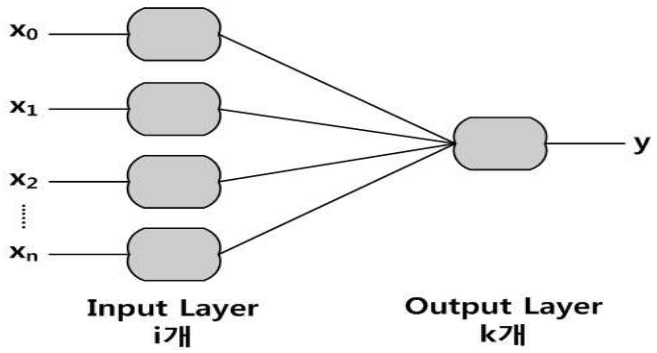


Figure 6. Perceptron Model

Figure 6에서 단일 노드는 입력값의 가중치합에서 임계값 Θ를 뺀 다음 Hard-limit Transfer Function에 의해 출력값을 산출한다. 이때 출력값은 +1이거나 -1인데, A 클래스에 속하면 +1을, 반대로 B 클래스에 속하면 -1을 출력한다.

(1) Perceptron Learning Rule

오늘날 대부분의 학습 규칙들은 Hebb의 규칙에서 발전된 것들이다.

(가) 헤브의 학습규칙(Hebb's Rule, Hebb. 1949)

헤브의 학습규칙은 "만일 어떤 신경세포의 활성이 다른 신경세포가 활성화되는데 계속적으로 공헌한다면, 두 신경세포 간의 연결가중치를 증가시켜 주어야한다."라는 가설에 입각한 학습 규칙으로서, 뉴런 간의 연결강도를 학습에 의해 최적화시키기 위한 규칙이다. 헤브의 규칙을 수식화하면 다음과 같다.

$$W(new)_{ij} = W(old)_{ij} + \alpha a_i b_j$$

$$\begin{cases} W(new)_{ij} : \text{신경세포} i, j \text{사이의 조절후 } Weight \\ W(old)_{ij} : \text{신경세포} i, j \text{사이의 조절전 } Weight \\ \alpha : \text{학습률} (0 < \alpha \leq 1) \\ a_i : \text{신경세포} i \text{의 활성값} \\ b_i : \text{신경세포} j \text{의 활성값} \end{cases}$$

이 때, 학습률 α는 Weight의 조절량을 결정하는 0과 1사이의 상수로서, 즉 그 값이 크면 Weight가 많이 조절되고 그 값이 작으면 Weight는 조금만 조절된다.

이 헤브의 규칙은 목적패턴이 필요치 않으며, 학습은 단지 연결된 두 개의 신경세포와 그 Weight에 의해서만 이루어지는 목적 없는 무감독 학습 규칙이다.

(나) 델타 규칙 (Delta Rule, Rosenblatt. 1958)

헤브의 학습규칙을 기초로 하여 컴퓨터 과학자 Rosenblatt는 뉴런 모델을 사용하여 패턴인식을 하는 Perceptron을 발표하였는데, 여기서 사용한 학습 규칙이 델타 규칙이다. 델타 규칙의 기본 가설은 "만일, 어떤 신경세포의 활성이 다른 신경세포가 잘못된 출력을 내는데 공헌하였다면, 두 신경세포 간의 Weight를 그것에 비례하여 조절해주어야 한다."이다. 델타 규칙을 수식화하면 다음과 같다.

$$W(new)_{ij} = W(old)_{ij} + \alpha e_i a_i \qquad e_j = t_j - b_j$$

$W(new)_{ij}$: 신경세포 i, j 사이의 조절후 $Weight$
$W(old)_{ij}$: 신경세포 i, j 사이의 조절 전 $Weight$
α : 학습률 $(0 < \alpha \leq 1)$
t_j : 목적패턴의 출력층 신경세포 j에 대응하는 성분값(목표출력값)
b_j : 출력층 신경세포 j의 활성값(실제출력값)
a_i : 신경세포 i의 활성값

(2) Perceptron Algorithm Procedure

① 가중치 $w_i(0)$ $(0 \leq i \leq n-1)$와 임계값을 양수인 임의의 작은 수로 초기화한다. 이때 $w_i(t)$는 시간 t에서 입력 i의 가중치이고, Θ는 출력 노드에 있는 값을 의미한다.

② 원하는 출력값 d(t)을 갖는 새로운 연속적인 입력값 $x_0, x_1, x_2, \cdots x_{n-1}$을 입력한다.

③ 다음 식에 따라 실제의 출력값을 계산한다.

$$y(t) = f_h\left(\sum_{i=0}^{n-1} w_{ij}(t)x_i(t) - \theta\right)$$

④ 다음 식에 따라 가중치를 조절한다. (학습과정)

$$w_i(t+1) = w_i(t) + \eta[d(t) - y(t)]x_i(t),\ (0 \leq i \leq n-1, 0 < \eta < 1)$$

$$d(t) = \begin{cases} +1 & A\text{클래스로부터의 입력일 경우} \\ -1 & B\text{클래스로부터의 입력일 경우} \end{cases}$$

η : 학습을 조절하는 임의의 이득률(gain fraction)

만약, 정확히 결정 경계선에 의해 패턴 결정이 되면 가중치는 불변하는데, 이 경우를 수렴(Convergence)한다고 한다.

⑤ 단계 ②로 이동하여 반복한다.

(3) Perceptron Transfer Function

Perceptron에서 사용된 전달함수는 Hard-Limit Transfer Function으로서, 그 출력값은 0과 1이다.

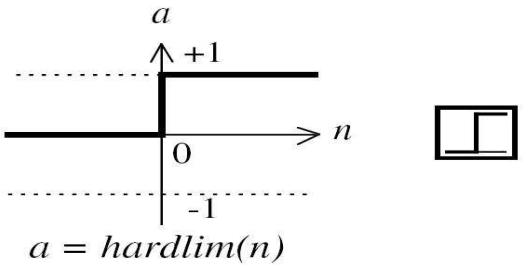


Figure 7. Hard-Limit Transfer Function

라. Multi-Layer Perceptron

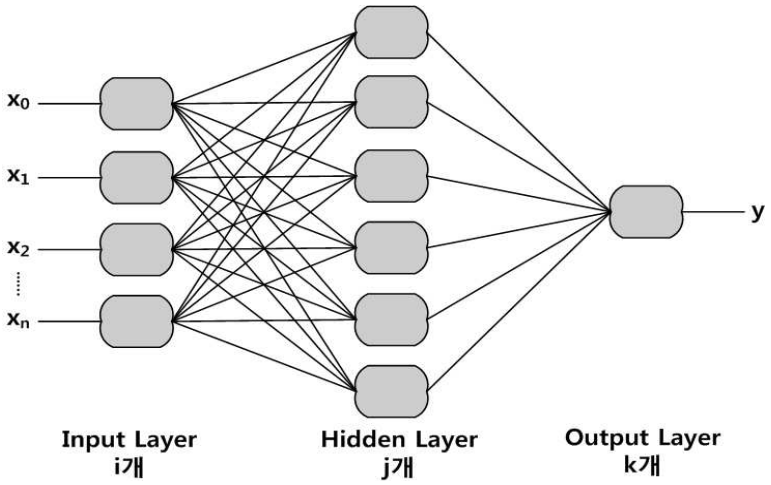


Figure 8. Multi-Layer Perceptron Model

Multi-Layer Perceptron(이하 MLP)은 입력층과 출력층 사이에 하나 이상의 계층을 갖는, Perceptron보다 진보된 NN 모델이다. MLP는 Perceptron의 단점을 해결하기 위해 제안된 모델이다. 층의 개수가 증가할수록 Perceptron이 형성하는 결정구역의 특성은 세밀해진다. 단층일 경우에는 패턴 공간을 두 영역으로 나누어주고, 2층인 경우에는 오목(convex) 개구역 또는 폐구역을 형성하며, 3층인 경우에는 이론상 어떠한 형태의 구역도 형성할 수 있다.

MLP에서 각 처리 인자를 PE(Perceptron Element)라고 하는데, PE의 기능은 활성화 함수가 갖는 비선형성에 따른다. 활성화 함수를 비

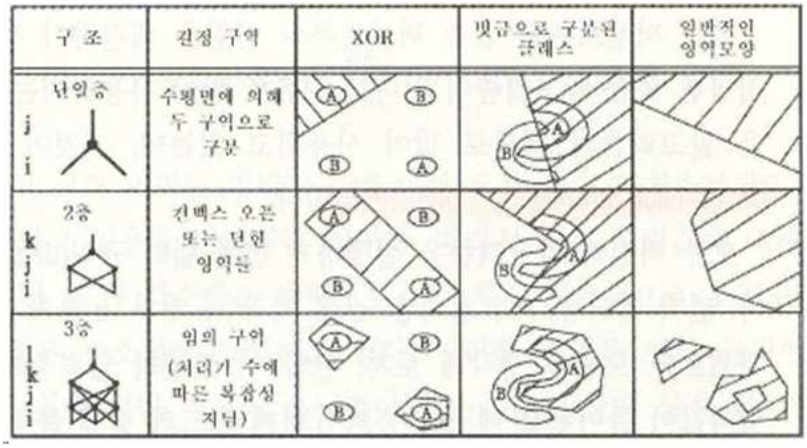


Figure 9. Neural Network 층의 수에 따른 결정 영역

선형으로 하고 층의 개수가 증가함에 따라 나눌 수 있는 영역이 복잡해 질 때 층의 구조에 따른 결정 경계의 차이를 Figure 9에서 보여주고 있다. 하지만 MLP는 적당한 학습 알고리즘이 없다는 이유로 많이 사용되지는 못하다가, 오류 역전파 알고리즘(Error Back-Propagation Algorithm)의 개발로 많이 사용되기 시작하였다. 역전파 알고리즘의 학습은 전향(forward) 단계와 후향(backward) 단계로 이루어진다.

① 전향 단계 : NN에 입력 패턴을 제시하고, 각 노드에 대해서 입력 함수와 활성화 함수를 이용하여 출력을 산출하는 단계이다. 이 때 입력 신호는 전방향, 즉 출력측 방향으로만 전달된다.

② 후향 단계 : 목표 출력과 실제 출력의 차이를 계산하여 오차를 구하고, 이를 역방향, 즉 출력층에서 입력층으로 순서대로 층과 층 사이의 Weight를 갱신하는 단계이다. Weight를 조절한 후, 다시 입력을 넣어 계산하면 출력은 처음 시도 때보다 오차가 작은 값을 얻게 된다. 이를 반복하여 시스템이 안정될 때까지, 즉 총 오차의 합이 정해진 오차의 기준치에 도달할 때까지 실행시켜서 원하는 값에 수렴시키는 것이다.

Weight를 조절할 때, 그 보정 값을 오차에 비례하게 해주는 것은 델타규칙에 LMS 학습 규칙을 적용한 일반화한 델타 규칙

(Generalized Delta Rule)으로 행하게 된다.

(1) Multi-Layer Perceptron Learning Rule

(가) LMS 학습 법칙 (Widrow & Hoff, 1960)

LMS(Least Mean Square) 학습 알고리즘은 Perceptron 학습 규칙보다 더 일반화된 학습규칙으로, 전체 학습 데이터에 대한 전체 오차를 최소화하는 방향으로 Weight 갱신시키는 방법이다. 입력 x에 대한 목표 출력이 y인 경우, 자승 오차는

$$E = \frac{1}{2} Err^2 = \frac{1}{2} (y - h_w(x))^2$$

이다. 이 함수식은 $h_w(\cdot)$에 따라서 결정된다. 만약, $h_w(\cdot)$가 일차식이라면 자승 오차는 2차 함수가 되므로 포물선 모양이 될 것이다.

$$w(n+1) = w(n) - \mu \frac{dE}{dW}$$

이 때 최소 오차점을 찾기 위해서는 W에 관하여 오차 함수식을 편미분한 결과가 기울기의 크기를 의미하므로, 그 값이 음이면 그 기울기만큼 ΔW 이전 연결강도에 가산하고, 그 값이 양이면 그 기울기만큼 ΔW를 빼주는 과정 반복하여 최소점에 도달할 수 있다. 이것이 경사강하법(Gradient Descent)이라 하고, 이를 수식으로 표현하면 다음과 같다.

$$\therefore \frac{dE}{dW_j} = Err * \frac{dErr}{dW_j} = Err * \frac{d}{dW}(y - g(\sum_{j=0}^{n} W_j x_j)) = -Err * g'(\in) * x_j$$

이 식을 weight 갱신 규칙에 적용하면 다음과 같이 된다.

$$W(new)_j \rightarrow W(old)_j + \alpha * Err * g'(\in) * x_j$$

LMS 학습법칙은 오차를 줄이기 위하여 학습할 때마다 현재의 Weight에 의한 오차, 즉 이상적인 Weight와의 거리를 계산하여 그 방향으로 weight의 값 조정하는 것이다. 이 때 α(학습률, Learning

Rate)는 매우 중요한 상수로서, 이 값이 너무 작으면 최소점 도달 시간이 길어지고, 국소 최소(local minima)에서 벗어나지 못한다. 반대로 이 값이 너무 크면 전역 최소(global minima)에 수렴하지 않을 수도 있고, 진동할 가능성도 있다.

그리고 델타 규칙에 LMS 알고리즘을 적용하기 위해서는 각 Unit에서 미분 가능한 활성화 함수를 사용하여야 한다. 즉, Hard-Limit Transfer Function은 미분이 불가능하므로, Sigmoid Function과 같은 미분 가능한 함수를 활성화 함수로 사용하여야 한다.

(나) Back-Propagation Algorithm (Rumelhart, 1986)

입력층이 i개, 은닉층이 j개, 출력층이 k개로 구성된 MLP가 있다고 가정하면, 은닉층 j번째의 Unit에 대한 입력의 합은

$$net_{pj} = \sum W_{ji} O_{pi}$$

이고, 은닉층 j번째 Unit의 출력은

$$O_{pj} = f(net_{pj})$$

이다. 여기서 f는 Sigmoid Transfer Function이므로

$$O_{pj} = \frac{1}{1+\exp(-net_{pj})}$$

가 된다. 마찬가지로 출력층 k번째 Unit에 대한 입력의 합은

$$net_{pk} = \sum W_{kj} O_{\pi}$$

이고, 출력층 k번째 Unit의 출력은

$$O_{pk} = f(net_{pk})$$

이다. 일반적으로 출력 O_{pk}는 원하는 값 t_{pk}와 같지 않기 때문에 각각의 패턴에 대한 오차는

$$E = \frac{1}{2}\sum_{k}(t_{pk} - O_{pk})^2$$

이다. 일반화한 델타 규칙은 LMS 학습 규칙을 이용하여 이 오차 E를

최소로 하는 Weight를 구한다. 이 방법은 Weight의 미분 값을 구하여, 이 값에 비례해서 weight를 변화시킬 양을 산출하는 방법이다.

① 출력층과 은닉층간의 Weight 변화

$$\Delta W_{kj} = -\eta \frac{dE}{dW_{kj}} \tag{1}$$

식 (1)에서 우측 항은

$$\frac{dE}{dW_{kj}} = \frac{dE}{dnet_{pk}} \frac{dnet_{pk}}{dW_{kj}} \tag{2}$$

로 두면, 식 (2)에서

$$\frac{dnet_{pk}}{dW_{kj}} = \frac{d\sum W_{kj} O_{pj}}{dW_{kj}} = O_{pj} \tag{3}$$

가 되고,

$$\delta_{pk} = -\frac{dE}{dnet_{pk}} \tag{4}$$

라고 두면, 식 (1)의 출력층과 은닉층간의 Weight의 변화량은

$$\Delta W_{kj} = \eta \delta_{pk} O_{pj} \tag{5}$$

로 나타낼 수 있다. 식 (4)에서

$$\delta_{pk} = -\frac{dE}{dnet_{pk}} = -\frac{dE}{dO_{pk}} \frac{dO_{pk}}{dnet_{pk}} \tag{6}$$

가 된다. 식 (6)의 우측 항에서

$$\frac{dE}{dO_{pk}} = -(t_{pk} - O_{pk}) \tag{7}$$

$$\frac{dO_{pk}}{dnet_{pk}} = f'_k(net_{pk}) \tag{8}$$

이고, 식 (7)과 식 (8)에서 최종 δ_{pk}값을 구할 수 있다.

$$\delta_{pk} = (t_{pk} - O_{pk}) f_k(net_{pk}) O_{pj} \tag{9}$$

따라서, 출력층 k번째 Unit에 대한 Weight의 변화량은 식 (10)과 같다.

$$\begin{aligned}\Delta W_{pj} &= \eta(t_{pk} - O_{pk}) f_k(net_{pk}) O_{pj} \\ &= \eta \delta_{pk} O_{pj}\end{aligned} \tag{10}$$

② 입력층과 은닉층 간의 Weight 변화

$$\begin{aligned}\Delta W_{ji} &= -\eta \frac{dE}{dW_{ji}} = = -\eta \frac{dE}{dnet_{pj}} \frac{dnet_{pj}}{dW_{ji}} \\ &= -\eta \frac{dE}{dnet_{pj}} O_{pi} \\ &= \eta(-\frac{dE}{dO_{pj}} \frac{dO_{pj}}{dnet_{pj}}) O_{pi} \\ &= \eta(-\frac{dE}{dO_{pj}}) f_j(net_{pj}) O_{pi} \\ &= \eta \delta_{pj} O_{pi}\end{aligned} \tag{11}$$

여기서, $-\frac{dE}{dO_{pj}}$는 직접 구할 수 없으므로,

$$\begin{aligned}-\frac{dE}{dO_{pj}} &= \sum_k -\frac{dE}{dnet_{pk}} \frac{dnet_{pk}}{dO_{pj}} \\ &= \sum_k (-\frac{dE}{dnet_{pk}}) \frac{d\sum_j W_{kj} O_{pj}}{dO_{pj}} \\ &= \sum_k (-\frac{dE}{dnet_{pk}}) W_{kj} \\ &= \sum_k \delta_{pk} W_{kj}\end{aligned} \tag{12}$$

와 같이 유도할 수 있다. 따라서 이 경우에

$$\delta_{pj} = f'_j(net_{pj}) \sum_k \delta_{pk} W_{kj} \tag{13}$$

가 된다. 여기서 활성화 함수가 Sigmoid Function이면

$$f(net_{pj}) = O_{pj} = \frac{1}{1+\exp(-net_{pj})}$$

이므로

$$f'(net_{pj}) = O_{pj}(1 - O_{pj}) \tag{14}$$

가 된다. 결국 출력층과 은닉층의 Unit에 대한 오차 δ를 구하기 위하여 출력층 k번째 Unit의 오차는

$$\delta_{pk} = (t_{pk} - O_{pk}) O_{pk}(1 - O_{pk}) \tag{15}$$

이고, 은닉층의 오차는

$$\delta_{pj} = O_{pj}(1 - O_{pj}) \sum_k \delta_{pk} W_{kj} \tag{16}$$

가 되어, 은닉층의 Unit에 대해서는 재귀적 연산이 이루어지게 된다.

따라서, 최종적인 출력층과 은닉층의 Weight의 갱신은

$$\Delta W_{kj}(n+1) = \eta \delta_{pk} O_{pj} + \alpha \Delta W_{kj}(n) \tag{17}$$

이고, 은닉층과 입력층의 Weight의 갱신은

$$\Delta W_{ji}(n+1) = \eta \delta_{pj} O_{pi} + \alpha \Delta W_{ji}(n) \tag{18}$$

이다. 여기서 η은 학습률이고, α는 모멘텀이다.

(2) Multi-Layer Perceptron Algorithm Procedure

① 모든 노드에 대한 가중치와 임계값을 임의의 수로 초기화한다.

② 모든 훈련 데이터를 반복적으로 입력한다. 연속적인 입력 벡터 x_0, x_1, …x_{n-1}과 차례로 대응되는 원하는 출력값 d_0, d_1, …d_{n-1}을 입력한다. 만약 패턴분류에만 사용할 경우 입력과 관련된 해당 클래스의 원하는 출력값만을 1로 하고, 이를 제외한 모든 출력층의 원하는 출력값을 모두 0으로 한다. 이 때 가중치가 불변할 때까지 계속적으로

반복한다.

③ 다음의 Sigmoid Transfor Function을 이용하여 실제 출력값 y_0, y_1, …y_{n-1}을 차례대로 구한다.

④ 출력층 노드의 값에서 시작해서 거꾸로 첫 번째 은닉층으로 전달하는 반복 알고리즘을 이용하여 가중치를 조절한다. 이 과정을 가중치가 거의 변화하지 않을 때까지, 즉 δ가 거의 0에 접근할 때까지 반복한다.

⑤ 단계 ②로 이동하여 반복한다.

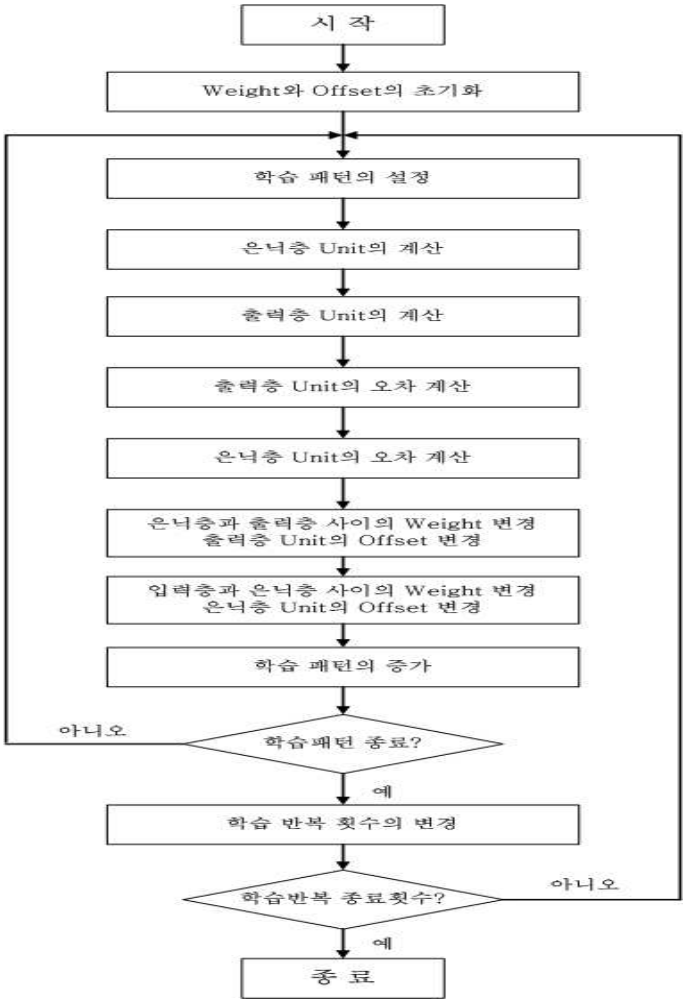
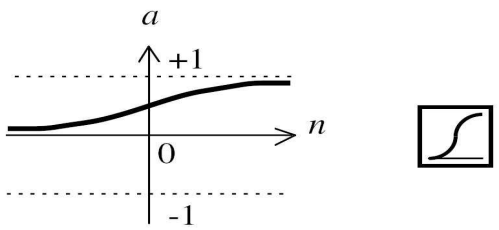


Figure 10. Back-Propagation Algorithm 순서도

(3) Multi-Layer Perceptron Transfer Function

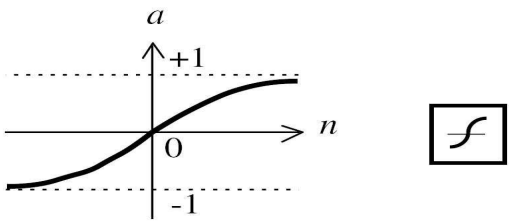
앞에서 언급했듯이, Multi-Layer Perceptron의 전달함수는 미분가능한, 즉 Sigmoid Function과 같은 함수를 사용해야한다. 대표적인 Sigmoid Function은 Log-Sigmoid Transfer Function과 Hyperbolic Tangent-Sigmoid Transfer Function이 있다.

Log-Sigmoid Transfer Function은 무한대의 입력값에 대하여 그 출력값을 0~1의 값으로 출력시키는 함수이고, Hyperbolic Tangent-Sigmoid Transfer Function은 그 출력값을 -1~1의 값으로 출력시키는 함수이다.



$a = logsig(n)$

Figure 11. Log-Sigmoid Transfer Function



$a = tansig(n)$

Figure 12. Hyperbolic Tangent-Sigmoid Transfer Function

마. Neural Network의 응용분야

Neural Network은 기존 컴퓨터의 한계를 극복할 수 있는 특징으로 인하여 기존의 인공지능 기법이나 계산이론 기법으로 해결하기 힘들었던 분야, 즉 패턴인식, 음석인식, 자연언어이해 등의 응용분야에 주로 많이 적용되었다.

① 음성 합성 및 인식 분야

존스 홉킨스 대학의 Sejnowski와 프리스턴 대학의 Rosenberg는 문장을 음성으로 변환하는 NN을 만들었고, NEC사에서도 동적 프로그래밍과 NN 기술을 결합하여 숫자를 인식하는 음성 인식 시스템을 만들어 오류율을 기존의 방법을 사용한 시스템의 1/3로 줄일 수 있다고 발표하였다.

② 언어 학습 분야

캘리포니아 대학의 Rumelhart와 카네기멜론 대학의 McClelland?는 영어 동사의 과거 시제를 배우는 NN을 개발하였다.

③ 문자 인식 분야

인쇄체 및 필기체 문자의 인식 능력을 바탕으로 이미 우편봉투 자동분류, 수표 및 지로용지의 인식, 인구센서스 결과의 통계, 세금보고서의 자동처리 등이 이루어지고 있는데, 현재 세계적으로 독일의 Siemens, 일본의 NEC, 미국의 CEDAR, SUNY at Buffalo 등이 우편물 분류를 중심으로 한 높은 기술 수준을 보유하고 있다.

④ 영상 처리 분야

미국 국방연구원(DARPA)의 지원에 의한 잠수함의 장애물 인식 신경망에 관한 연구를 진행하였다.

이외에도 NN은 주가 변동 예측, 항공사 좌석 예약 관리, 고객의 은행 신용도 판별, DNA 코드 분석 등에 응용되고 있는데, 특히 최근에는 대규모의 복잡한 데이터로부터 유용한 규칙이나 새로운 지식을 발견하기 위한 데이터마이닝(data mining)에 관한 연구에서도 NN이 활용되고 있다.

3. Experimental Data

연세대학교 대사증후군 연구 사업에 참여한 4254명을 실험대상자로 선정하였다. 여기서 실제 당뇨병을 앓고 있는 환자의 수는 241명이고, 정상인은 4013명이었다.

우선 대상자들의 기본 신체계측데이터(성, 연령, 키, 체중, 허리둘레, 수축기/이완기 혈압)와 혈액검사데이터(Adiponectin, Trigriceride, HDL 콜레스테롤), 흡연력, 음주력, 운동여부에 대한 데이터를 획득하였다. 그 중 수축기/이완기 혈압이 140/90㎜Hg이상일 경우 고혈압(Hypertension)으로 진단하고, BMI 수치는 $BMI = \frac{WT^2}{HT}$으로 산출할 수 있기 때문에 고혈압여부 및 BMI 수치데이터를 사용하였다.[18]

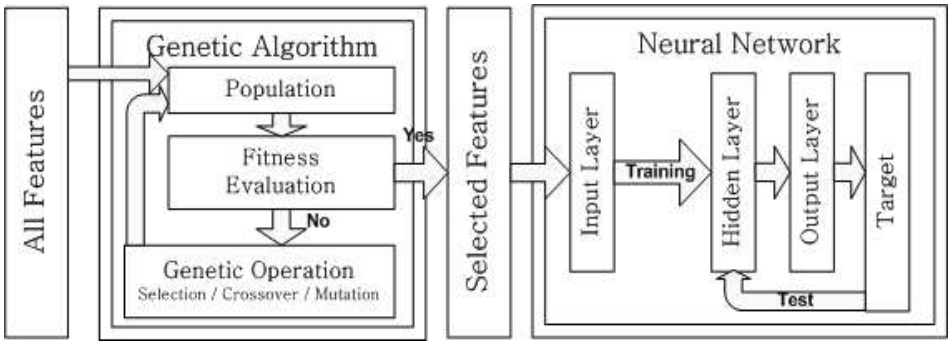
특히 최근 혈청 아디포넥틴은 당뇨병의 새로운 위험인자로 보고되고 있고, 허리둘레 역시 아디포넥틴과 함께 유의한 관련성을 보였다. 또한 아디포넥틴과 당뇨병 발생과의 관계에서 아디포넥틴은 높은 고밀도콜레스테롤과 관련이 있었고[19], 고밀도콜레스테롤은 당뇨병발생을 감소시키는 중요한 관련이 있다고 알려져 있다.[20]

따라서 실험에 사용한 데이터 변수는 Adiponectin, TG, HDL, Sex, Age, Waist, BMI, HTN, Smoke, Exercise, Alcol, 총 11개의 변수를 사용하였다.

Table 2. Experimental Data Analysis (Total / Normal / Patient)

Total : 4254				
Feature	Max	Min	Average	STD
Adipo	54.15	0.46	8.89	5.43
TG	1189	21	133.28	94.67
HDL	125	12	56.29	13.49
Sex	Male : 1690명, Female : 2564명			
Age	87	20	46.09	9.85
Waist	140	56	80.93	9.19
BMI	38.00	15.94	24.02	2.98
HTN	Patient : 943명, Normal : 3311명			
Smoke	Non : 2173명. Ex : 1014명, Current : 1068명			
Exer	운동 : 1625명, 비운동 : 2629명			
Alcol	음주 : 3099명, 비음주 : 1155명			
Normal : 4013				
Adipo	54.15	0.46	9.01	5.43
TG	1189	21	130.29	90.60
HDL	125	12	56.72	13.52
Sex	Male : 1649명, Female : 2364명			
Age	80	20	45.68	9.73
Waist	115	56	80.60	9.13
BMI	38.00	15.94	23.94	2.95
HTN	Patient : 838명, Normal : 3175명			
Smoke	Non : 2096명, Ex : 941명, Current : 977명			
Exer	운동 : 1563명, 비운동 : 2450명			
Alcol	음주 : 2910명, 비음주 : 1103명			
Patient : 241				
Adipo	41.40	0.62	6.88	4.91
TG	1104	31	183.14	137.70
HDL	88	28	49.13	10.68
Sex	Male : 41명, Female : 200명			
Age	87	32	52.76	9.54
Waist	140	66.5	86.39	8.48
BMI	37.70	17.98	25.35	3.02
HTN	Patient : 105명, Normal : 136명			
Smoke	Non : 77명. Ex : 73명, Current : 91명			
Exer	운동 : 62명, 비운동 : 179명			
Alcol	음주 : 189명, 비음주 : 52명			

4. Experimental Model



Feature 13. Experimental Model

Experimental Model은 위에서 설명한 Genetic Algorithm과 Neural Network을 동시에 적용하여 구성하였다. Experimental Dataset에서 구성한 11개의 데이터 변수는 Genetic Algorithm을 통해 당뇨병 발병 가능성을 예측하는데 가장 최적화된 변수만을 선택하게 된다. 그 선택된 변수들의 데이터를 기반으로 Neural Network Model을 구축하게 된다. 위에서 정의된 대로 Training Data를 통해 학습된 NN Model에 Test Data를 입력시켜 그 학습된 모델의 정확도를 민감도(Sensitivity)와 특이도(Specificity)를 이용하여 측정・비교한다.

이 때, 측정된 결과값은 Experimental Model에서 설정한 각종 파라미터와 사용한 알고리즘, 함수 등의 조합에 따라서 달라지게 된다. GA에서의 파라미터는 Selection / Crossover / Mutation Function 및 Coefficient, Population Size, Generation 등이 있으며, NN에서의 파라미터는 Training Algorithm, Transfer Function, Epoch, Learning Rate, Hidden layer의 수 및 그 layer의 노드 수 등이 있다.

Ⅲ. 결 과

1. Data Preprocessing

가. Outlier Rejection

LDL-Cholesterol(저밀도지단백콜레스테롤) 측정을 위해 가장 널리 쓰이는 방법은 Friededwald 공식에 의한 계산법으로서, Total Cholesterol, Triglyceride(TG), HDL-Cholesterol(고밀도지단백콜레스테롤)을 측정하여 아래의 식에 의하여 계산하는데, 이 공식은 TG가 400mg/dL 이상인 경우에는 사용할 수 없다.[21,22] 이 기준을 적용하여 데이터 중 TG의 데이터 값이 400mg/dL이상인 데이터를 Outlier로 간주하고 제거하였다.

(LDL-Cholesterol) = (Total Cholesterol)-[(HDL-Cholesterol)+TG]

따라서 전체 4254명의 데이터에서 Outlier 데이터인 79명의 데이터를 제외한 전체 4175명의 데이터를 이용하여 Dataset을 구성하였다.

나. Data Normalization

Dataset을 구성하고 있는 각각의 데이터들은 Numeric Data(Adipo, Tg, HDL, Age, Waist, BMI)와 Categorical Data(Sex, HTN, Smoke, Exercise, Alcol)로 구분할 수 있다. 따라서 이렇게 각각의 변화량이 서로 다른 데이터를 NN의 Input data로 설정하기 위해 각 데이터별로 정규화(Normalization) 과정을 통해 최소 / 최대값을 기준으로 아래의 식을 적용하여 0~1의 값으로 변환하였다.[16]

$$v_n = \frac{v_i - \min(v_1 \cdots v_n)}{\max(v_1 \cdots v_n) - \min(v_1 \cdots v_n)}$$

(v_n : normalized value, v_i : instance value)

다. Dataset Configuration

TG의 값이 400mg/dL 이상인 데이터인 Outlier를 제거한 4175명의

데이터는 230명의 환자데이터와 3945명의 정상데이터로 구성되어 있다. 이 데이터들을 Training Data와 Test Data로 구분함에 있어서 각각 3:1의 비율로 구분하였다. 따라서 3129명의 Training Data 중 정상데이터는 2958개, 환자데이터는 171개가 존재하고, 1046명의 Test Data 중 정상데이터는 987개, 환자데이터는 59개가 존재한다.

Table 3. Dataset Configuration

	Normal	Patient	Total
Training Data	2958	171	3129
Test Data	987	59	1046
Total	3945	230	4175

2. Optimization of Experimental Model

가. Parameters Setting

(1) Genetic Algorithm Parameter

① Selection Function : Roulette SelectionFcn

② Crossover Function : Two-point CrossoverFcn

③ Mutation Function : Uniform MutationFcn (계수 : 0.01)

④ Population Size : 20

⑤ Generation : 10

(2) Neural Network Parameter

① Training Algorithm : Resilient Backpropagation Algorithm

② Transfer Function : Hyper-Tangent Sigmoid TransferFcn

③ Training Epoch : 1000

④ Training Mode : Batch Mode

⑤ Learning Rate & the numbers of Hidden Nodes

: Selected Feature-set에 따라 개별적으로 최적화 진행

나. Procedure of Optimization

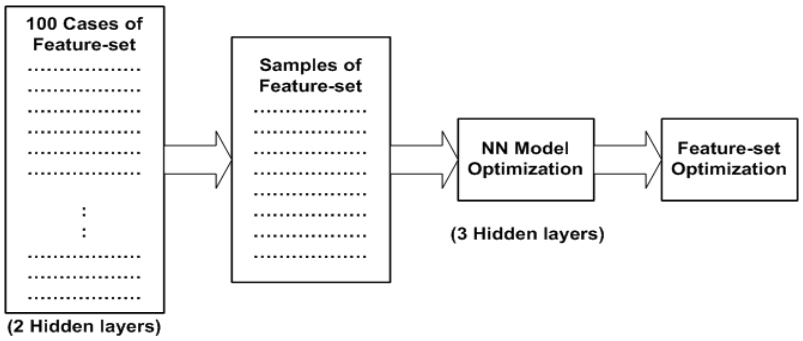


Figure 14. Procedure of Optimization Model

먼저 최적화 과정의 첫 단계로서, Hidden layer는 1 또는 2 layer로 설정한 상태에서 layer의 Node 수를 변경하였을 때, Sensitivity와 Specificity 정확도가 최상인 결과를 출력한 Feature-set을 추출하는 과정을 100 케이스 이상 진행하였다. 이런 과정을 통해 축적된 Feature-set들을 데이터베이스화한 후, 여기서 Feature-set의 샘플을 추출하였다. 샘플을 추출하는 기준은 Feature-set DB에서 중복 선택된 Feature-set과 우수한 결과를 보여준 Feature-sets, 그리고 각각의 Feature별 선택횟수를 파악하여 50회 이상 선택된 Feature들만으로 구성한 Feature-set 등이다. 이렇게 선정한 Feature-set 샘플들에 대하여 최상의 결과를 출력하는 3 Hidden Layer의 NN Model을 최적화 모델로 선정한 후, 이렇게 최적화된 NN Model에 대하여 Feature-set을 최적화시켰다.

3. Experimental Result

실험에 사용한 변수(Adipo, TG, HDL, Sex, Age, Waist, BMI, HTN, Smoke, Exercise, Alcol)에는 신체 계측 정보나 생활습관과 같이 단순 측정·문답과정을 통해 얻을 수 있는 데이터와 혈액검사를

통해 얻을 수 있는 데이터로 구분할 수 있다. 즉, 이 논문에서 연구하는 예측모델을 적용하고자 할 때에는 부가적인 혈액검사가 수반되어야 한다는 뜻이다.

만약 혈액검사를 시행한다고 하면 이를 통해 혈중 포도당 농도를 직접 구할 수 있기 때문에, 별도의 검사를 하지 않고 일상 환경에서 자신의 기본 계측정보만을 이용하여 당뇨병 발병을 예측할 수 있는 모델에 대한 연구도 필요하다고 판단하였다.

따라서 본 연구를 진행함에 있어서, 위의 11개의 변수를 모두 사용하여 결과를 구하는 경우와 혈액검사를 통해 얻을 수 있는 데이터를 제외한 환자 계측 데이터만 사용하여 결과를 구하는 경우, 두 가지로 구분하여 최적화 과정을 진행하였다.

GA-NN 모델의 결과에서 Table 5,6,7,8은 실험케이스별 Sensitivity와 Specificity 정확도 및 각종 Parameter, 선택된 Feature-set을 나타낸 것이고, Figure 15, 16, 18, 19는 최적화된 NN Model의 성능을 학습 횟수당 MSE(Mean Square Error)의 값으로 표현한 것이다. 그리고 Figure 17, 20은 최적화된 NN Model 사용시 Feature-set을 최적화시키는 GA의 성능을 표현한 것이다. 마지막으로 각각의 실험케이스별 결과를 Table 4에 정리하였다.

Table 4. Results of GA-NN Model

Feature	Sensitivity	Specificity	Percent Agreement
DM4175_11_All	72.9%	67.3%	67.6%
DM4175_11_Select	69.5%	67.7%	67.8%
DM4175_8_All	72.9%	61.3%	62.0%
DM4175_8_Select	72.9%	63.3%	63.9%

가. DM4175_11

(1) All Feature 적용시

Table 5. Result of DM4175_11_All Feature

	DM +	DM -	Total
Test +	43	323	366
Test -	16	664	680
Total	59	987	1,046
Sensitivity		43/59 * 100 = 72.9%	
Specificity		664/987 * 100 = 67.3%	
Percent Agreement		707/1046 *100 = 67.6%	
Hidden Nodes: 35-15-10		Learning Rate: 0.01	MSE: 0.0256

Adipo	TG	HDL	Sex	Age	Waist	BMI	HTN	Smoke	Exer	Alcol
●	●	●	●	●	●	●	●	●	●	●

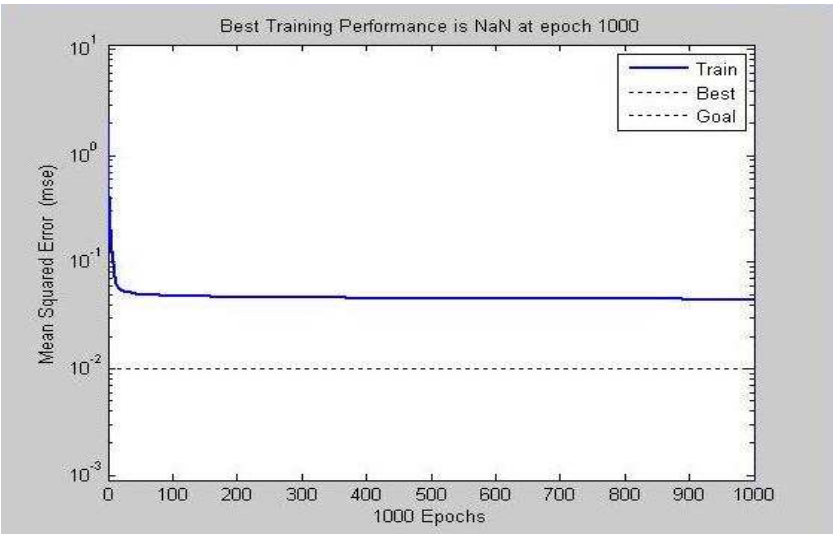


Figure 15. NN-Performance of DM4175_11_All Feature

(2) Selected Feature 적용시

Table 6. Result of DM4175_11_Selected Feature

	DM +	DM -	Total
Test +	41	319	360
Test -	18	668	686
Total	59	987	1,046
Sensitivity		41/59 * 100 = 69.5%	
Specificity		668/987 * 100 = 67.7%	
Percent Agreement		709/1046 *100 = 67.8%	
Hidden Nodes: 25-20-10		Learning Rate: 0.01	MSE: 0.0415

Adipo	TG	HDL	Sex	Age	Waist	BMI	HTN	Smoke	Exer	Alcol
●	○	●	○	●	●	●	●	●	●	○

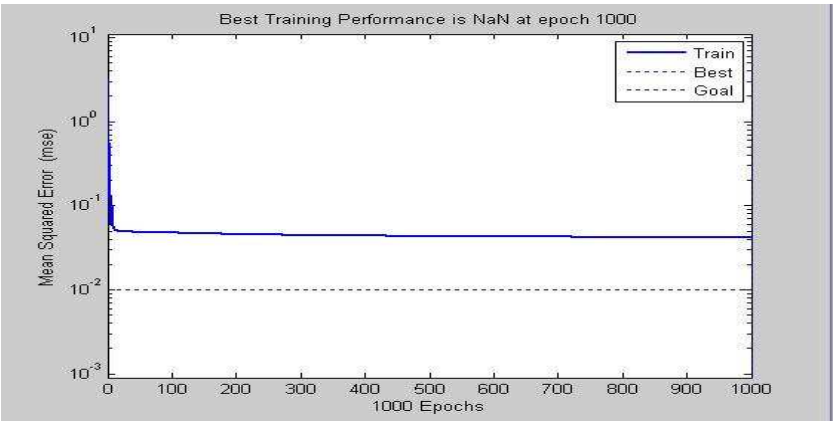


Figure 16. NN-Performance of DM4175_11_Selected Feature

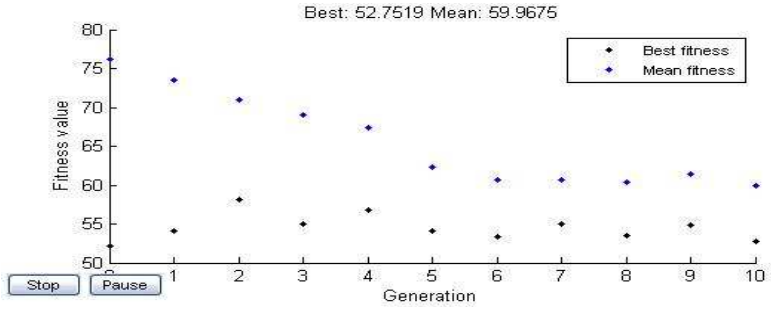


Figure 17. GA-Performance of DM4175_11_Selected Feature

나. DM4175_8

(1) All Feature 적용시

Table 7. Result of DM4175_8_All Feature

	DM +	DM -	Total
Test +	43	382	425
Test -	16	605	621
Total	59	987	1,046
Sensitivity		43/59 * 100 = 72.9%	
Specificity		605/987 * 100 = 61.3%	
Percent Agreement		648/1046 *100 = 62.0%	
Hidden Nodes: 50-80-80		Learning Rate: 0.05	MSE: 0.0317

Sex	Age	Waist	BMI	HTN	Smoke	Exer	Alcol
●	●	●	●	●	●	●	●

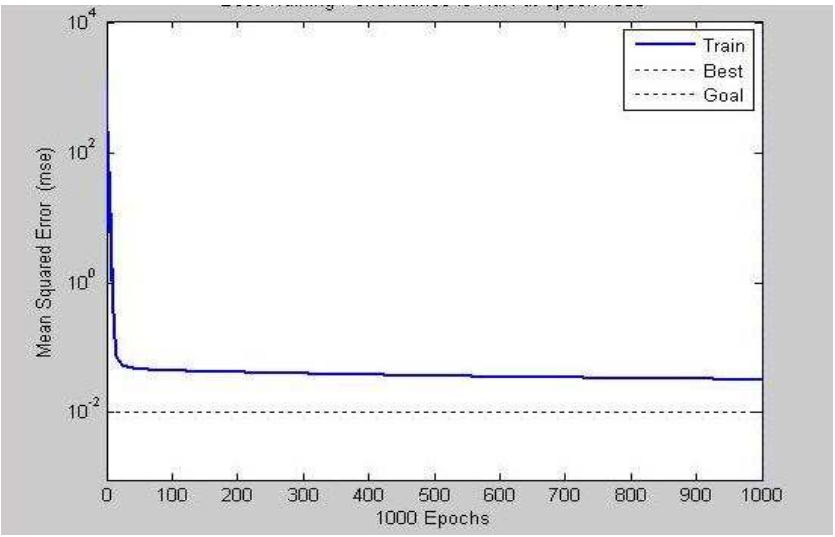


Figure 18. NN-Performance of DM4175_8_All Feature

(2) Selected Feature 적용시

Table 8. Result of DM4175_8_Selected Feature

	DM +	DM -	Total
Test +	43	362	405
Test -	16	625	641
Total	59	987	1,046
Sensitivity		43/59 * 100 = 72.9%	
Specificity		625/987 * 100 = 63.3%	
Percent Agreement		668/1046 *100 = 63.9%	

Hidden Nodes: 10-20-10	Learning Rate: 0.50	MSE: 0.0431

Sex	Age	Waist	BMI	HTN	Smoke	Exer	Alcol
○	●	●	○	●	●	●	○

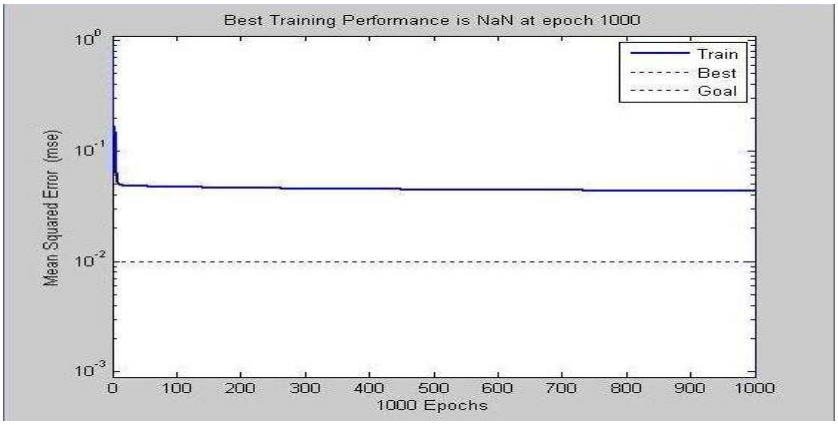


Figure 19. NN-Performance of DM4175_8_Selected Feature

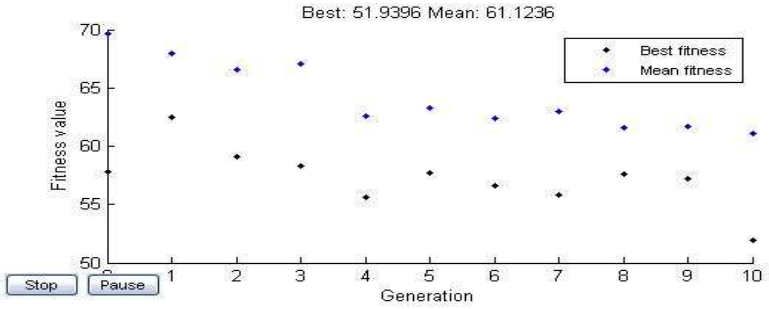


Figure 20. GA-Performance of DM4175_8_Selected Feature

4. Logistic Regression Model Result

위의 연구에서 사용한 동일한 Feature-set과 데이터셋을 이용하여 Logistic Regression Model을 사용한 결과를 GA-NN 모델의 대조군으로서 설정하여 비교·분석하였다.

Logistic Regression Analysis는 대수선형모형의 일종으로서, 로짓(logit) 회귀모형이라고도 불리며, 종속변수가 두 범주로 구성되어 있는 명목변수일 때 적용되는 통계적 기법이다.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x$$

위의 식이 로짓을 선형회귀식의 형태로 변형하는 로짓변환식이다. 종속변수가 0 또는 1만의 값을 갖는 가변수(dummy variable)인 경우에 y의 기댓값을 나타내는 반응함수의 모양이 S자형 곡선을 그리는 경우가 실제로 많이 나타난다. 이 반응함수는 x가 증가함에 따라 E(y)의 값이 1로 서서히 수렴하는 양상을 보인다. 이와 같은 함수를 로지스틱 함수(logistic function)라 한다. 즉, 로지스틱 회귀분석이란 단지 두 개의 값만을 가지는 종속변수와 독립변수들 간의 인과관계를 로지스틱 함수를 이용하여 추정하는 통계기법이다.[23]

로지스틱 회귀분석 모델을 사용하여 산출한 결과는 Table 9와 같다.

Table 9. Results of Logistic Regression Model

Feature	Sensitivity	Specificity	Percent Agreement
DM4175_11_All	79.7%	67.9%	68.5%
DM4175_11_Select	84.7%	65.9%	67.0%
DM4175_8_All	72.9%	64.2%	64.7%
DM4175_8_Select	69.5%	68.8%	68.8%

Ⅳ. 고 찰 및 결 론

인슐린 작용의 이상이나 부족에 의해 만성 고혈당증세를 보이는 당뇨병은 여러 특징적인 대사 이상을 수반하는 질환이다. 이러한 당뇨병의 예방을 위해 의료진은 식이요법, 지속적인 운동 등을 언급하고 있다. 이와 더불어 현재 자신의 체중, 신장, 혈압 등과 같은 신체 상태에 따른 질병 발병의 가능성을 상시적으로 관리하는 것 또한 중요하다. 이렇게 특정 질병의 발병가능성을 예측하거나 예측모델을 개발하는 연구는 이루어지고 있었다.

이러한 연구에는 주로 역학분야에서 통계적인 방법을 사용하여 예측모델을 개발하는 연구가 이루어져 왔지만, 최근에는 다양한 학문적 접근이 이루어지고 있다. 그 대표적인 것이 이 연구에서 사용한 Neural Network를 이용한 방법이다. Neural Network는 최근 금융업이나 증권업에서 방대한 데이터를 바탕으로 예측모델을 만들어내는데 많이 쓰이고 있는 인공지능시스템 이론 중 하나이다. 이 이론을 의료분야에서도 질병 발병가능성을 예측할 수 있는 예측모델 개발을 위해 많이 사용되고 있다. 그러나 이러한 연구의 대부분이 다양한 종류의 데이터를 모두 사용하기 때문에, 데이터 간의 Dimension reduction이 이루어지지 않아 오히려 결과에 좋지 않은 영향을 끼칠 가능성이 높다.

따라서 본 연구에서는 Neural Network를 이용하여 발병 예측 모델을 개발함에 있어서, Genetic Algorithm을 이용하여 개발한 모델의 정확도를 최대로 높여줄 수 있는 데이터만을 선별하도록 하여 최상의 정확도를 나타낼 수 있는 예측 모델을 최적화하고자 하였다.

연구에 사용한 총 4254명의 데이터 중에서 Friededwald 공식을 기반으로 TG의 값이 400이상인 데이터를 이상치(outlier)로 규정하여 데이터에서 제거한 후, 각 데이터별로 데이터 정규화(normalization)을 통해 0~1의 값으로 변화시켰다. 이렇게 데이터 전처리 과정을 거친

데이터를 training : test = 3 : 1의 비율로 구분하여 dataset을 구성한 후, 11개의 전체 데이터를 이용한 경우와 혈액검사를 통해 구할 수 있는 데이터를 제외한 8개의 데이터를 이용한 경우에 대하여 실험을 진행하였다. 각 실험케이스 별로 100여개의 Feature-set 데이터베이스에서 추출한 샘플 Feature-set에 대하여 GA 부분과 NN 부분의 각종 parameter와 function을 번갈아가면서 최적화되도록 하였다. 그 결과값은 dataset 중 Test Data의 민감도(sensitivity)와 특이도(specificity)를 기준으로 산출하였고, 그 결과를 Table 4에 정리하였으며, 동일한 dataset에 대하여 로지스틱 회귀모델을 이용하여 산출한 값을 실험군의 대조군으로 비교·분석하였다.

Table 4에서 먼저 Initial-Feature가 11개인 경우와 여기서 GA를 통해 Feature Selection을 한 경우를 비교해보면, GA를 이용하여 Feature를 선택한 경우의 Percent-Agreement가 67.8%로 약간 더 높은 것을 알 수 있었다. 또한 혈액검사를 통해 얻을 수 있는 데이터를 제거하여 Initial-Feature가 8개인 경우에도 GA를 통해 Feature를 선택한 경우의 Percent-Agreement가 63.9%로 약간 더 높은 것을 알 수 있었다. 즉, Initial-Feature의 모든 데이터를 사용한 경우보다는 Genetic Algorithm을 통해 Feature를 선택하였을 때에 더욱 좋은 결과를 보였다. 이는 발병 진단과 연관성이 낮은 데이터를 제거하였기 때문에 Dimension Reduction이 이루어졌기 때문이라고 판단된다. 또한 Initial-Feature에 혈액검사 데이터를 포함시켰을 경우에 더욱 좋은 결과를 보였는데, 혈액검사를 통해 얻을 수 있는 Adiponectin, TG, HDL-C와 같은 데이터가 당뇨병의 발병진단에 더욱 좋은 영향을 미치는 것을 알 수 있었다. 즉, 비록 혈액검사를 실시한다면 공복시 혈당치(FBS, Fasting Blood Sugar)를 이용하여 직접 당뇨병 여부를 판단할 수 있겠지만, Adiponection이나 TG, HDL-C와 같은 데이터가 있는 경우에는 발병 진단에 도움이 되는 것이다.

대조군인 Table 9의 Logistic Regression Model 결과와 비교해 보

면, 이 경우에도 Selection된 Feature만 이용한 결과가 Initial-Feature 전체를 사용하였을 경우보다 더 좋은 결과를 보였고, 혈액검사 데이터를 포함시킨 경우에 더 좋은 결과를 보였다. 실험군과 대조하여 볼 때, Logistic Regression Model을 사용하였을 때에 실험케이스별로 조금씩 변화는 있지만, 전반적으로 더 높은 결과를 보였다.

이러한 결과는 실험에 사용한 데이터에서 정상인의 데이터와 환자의 데이터간의 편중에 의한 것으로 분석된다. 일반적으로 Neural Network를 사용하여 어떠한 질병의 예측모델을 개발하는 경우에는 그 Training Data에서 정상데이터와 환자데이터를 1:1의 비율로 구성해야 한다. 네트워크 모델이 정상데이터와 환자데이터에 대하여 균등하게 training되지 못하고, 이 연구에 사용한 데이터와 같이 17:1의 비율로 정상데이터의 수가 훨씬 많은 데이터로 모델이 training된다면, 그 모델은 정상으로 bised된 모델이 되어버리기 때문에, Test Data에서 환자도 정상이라고 진단할 가능성이 높아지게 된다. 실례로 230명의 환자 데이터수와 동일하게 230명의 정상데이터를 random 선택한 후, Table 10과 같이 dataset을 구성한 다음 최적화 과정을 거친 결과 동일한 Feature-set에 대하여 다음과 같이 훨씬 높은 정확도를 나타내는 것을 확인할 수 있었다.

이와 같은 문제점을 해결할 수 있는 방안 중 현재 현실적인 방법은 Data Preprocessing 과정 중에 환자 데이터에 대해 Data Multiplication Method를 시행하는 것이다. 이는 편중된 데이터들 간의 균형을 맞추기 위해서 데이터를 복제하여 증식시키는 방법으로서, 이 방법에 시행하기 위한 가이드라인과 복제비율 등과 같은 사항에 대한 추가적인 연구가 이루어져야 할 것이다.

Table 10. DM460 - GA-NN Model Results

	Normal	Patient	Total
Training Data	180	180	360
Test Data	50	50	100
Total	230	230	460

Feature	Sensitivity	Specificity	Percent Agreement
DM460_11_All	80.0%	78.0%	79.0%
DM460_11_Select	82.0%	80.0%	81.0%
DM460_8_All	80.0%	72.0%	76.0%
DM460_8_Select	68.0%	86.0%	77.0%

그러나 문제점 해결을 위한 가장 이상적인 방법은 환자데이터를 추가적으로 획득하는 것일 것이다. 위의 Data Multiplication Method를 사용할 경우 환자데이터의 수가 증가하기 때문에 실험 결과에서 specificity는 상승할 수는 있겠지만, 네트워크 모델이 Training Data의 환자데이터에만 지나치게 overfitting되기 때문에 sensitivity는 오히려 감소할 가능성이 있다. 따라서 당뇨병 진단을 위한 방대한 데이터베이스를 기반으로 서로 다른 사람들의 데이터를 이용한 연구가 가장 최적화된 결과를 낼 것으로 판단된다. 정상인 3071명의 데이터와 환자 3071명의 데이터를 기반으로 Neural Network 기법을 사용하여 당뇨병을 예측한 선례 연구를 볼 때, Sensitivity 99.7%, Specificity 86.04%의 매우 높은 정확도를 나타내었다.[2]

환자 데이터의 추가 획득과 다양한 형태의 Neural Network 모델을 이용한 연구가 이루어져 더욱 높은 정확도를 보인다면, 실제 의료 환경에서 적용할 수 있는 예측 모델로서 정립될 수 있을 것이라 판단된다.

참 고 문 헌

1. Wichai A, Pongamorn B, Mark W, Piyamitr S, et al. A Risk Score for Predicting Incident Diabetes in the Thai Population. Diabetes Care Aug 2006;29:1872-7.

2. Jee SH, Park JW, Lee SY, Nam BH, Ryu HG, Kim SY, et al. Stroke risk prediction model: A risk profile from Korean study. Atherosclerosis 2008;197:318-25.

3. Jin Park, Dee W Edington. A sequantial neural network model for diabetes prediction. Artificial Intelligence in Medicine 2001;23:277-93.

4. 문병로. 쉽게 배우는 유전 알고리즘:진화적 접근법. 한빛미디어(주); 2008.

5. 진강규. 유전알고리즘과 그 응용. 2nd ed. 교우사; 2004.

6. 조영임. 인공지능시스템. 홍릉과학출판사; 2003.

7. 문경일,이현엽. MATLAB을 이용한 지능정보시스템. 도서출판 아진; 2003.

8. 한학용. 패턴인식 개론 - MATLAB 실습을 통한 입체적 학습. 한빛미디어; 2005.

9 변윤식,윤태성,김동준 외 2명 공역. 신경회로망 설계. 도서출판 인터비젼; 2008.

10. Simon Haykin. Neural Network - A Comprehensive Foundation. Macmillan Colledge Publishing Company; 1994.

11. 서혜숙, 최진욱, 이홍규, 민병구. 인슐린비의존형 당뇨병의 위험 인자 분석을 위한 신경망의 도입. 대한의료정보학회지 1998;4(2):127-31

12. Golnaz B, Ali MN. Controlling Blood Glucose Levels in Diabetics By Neural Network Predictor. Proceedings of the 29th Annual International Conference of the IEEE EMBS; 2007 August 23-26.

13. Stavroula GM, Aikaterini P, Dimitra I, Konstantina SN, Andriani V, Hristos SB. Neural Network based Glucose - Insulin Metabolism Models for Children with Type 1 Diabetes; Proceedings of the 28th IEEE EMBS Annual International Conference; 2006 Aug 30-Sept 3.

14. Alka MK, Christina LW, Nathalie DR, Ronald IS. Predicting the Development of Diabetes in Older Adults:The derivation and validation of a prediction rule. Diabetes Care Feb 2005;28(2):404

15. Muhammad AA, Ken W, Ralph AD, Michael S. What Is the Best Predictor of Future Type 2 Diabetes?; Diabetes Care Jun 2007;30(6):1544.

16. Dorian P. Data Preparation for Data Mining. Morgan Kaufmann Publishers; 1999.

17. 허미나,김창수,박민정,곽인숙,이규만. 균질법에 의한 저밀도지단백콜레스테롤의 측정 및 Friedewald 계산법과의 비교. 대한진단검사의학회지 2003;23(2):104-108.

18. 지선하,이희연,이선주,윤지은,지은정 외 4명. 한국인의 혈청 아디포넥틴과 당뇨병 진단기준 설정에 관한 연구. 한국역학회지 Dec 2007;29(2):176-86.

19. Schulze MB, Rimm EB, Shai I, Rifai N, Hu FB. Relationship between Adiponectin and glycemic control, blood

lipids, and inflammatory markers in men with type 2 diabetes. Diabetes Care 2004;27:1680-7.

20. Matsubara M, Maruoka S and Katayose S. Decreased plasma adiponectin oncentrations in women with dyslipidemia. J Clin Endocrinol Metab 2002;87(6):2764-9.

21. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin Chem 1972;18:499-502.

22. Smets EM, Pequeriaux NC, Blaton V, Goldschmidt HM. Analytical performance of a direct assay for LDL-cholesterol. Cin Chem Lab Med 2001;39:270-80.

23. 정광모,최용석. 로지스틱 회귀와 응용. 자유아카데미; 2003.

Abstract

A Study on Optimization Processing of Health-Data Using Neural Network

Do-Sung Kim

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sun-Kook Yoo)

Diabetes is a chronic disease, which occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. This leads to an increased concentration of glucose in the blood. In this study, we propose a prediction model that present the high accuracy of the outbreak of diabetes using genetic algorithm and neural network.

We used the data of 4175 persons(normal:3945, patient:230) that divided by training data and test data. And We used the data of Adiponectin, Triglyceride, HDL-Cholesterol, Sex, Age, Waist, BMI, HTN, Smoke, Exercise and Alcol. In these data, genetic algorithm was select the optimized features for prediction of disease, and neural network was predict the outbreak of diabetes. As a result, when use all features, prediction model was presented the accuracy of 67.8%. And when use the features except data what can get by a blood test, prediction model was presented the accuracy of 63.9%

As a further study, we expect that diabetes prediction model will use in medical environment, if additional patient data should get and a research of various neural network models should implement.

Key Words : Diabetes, Genetic Algorithm, Neural Network.