이 학 박 사 학 위 논 문

고차원 자료에 대한 심층 신경망에서의 초매개변수 조율 및 영향점 분석

2020년 2월

부 경 대 학 교 대 학 원

통 계 학 과

이 재 은

이 학 박 사 학 위 논 문

고차원 자료에 대한 심층 신경망에서의 초매개변수 조율 및 영향점 분석

지도교수 장 대 흥

이 논문을 박사 학위논문으로 제출함.

2020년 2월

부 경 대 학 교 대 학 원

통 계 학 과

이 재 은

이재은의 이학박사 학위논문을 인준함.

2020년 2월 21일

위 원 장 이학박사 이 성 백 (인)

위 원 이학박사 박 인 호 (인)

위 원 이학박사 문 승 호 (인)

위 원 공학박사 이 창 준 (인)

위 원 이학박사 장 대 흥 (인)

목 차

표 차례

그림 차례

Hyper-parameter Tuning and Influence Diagnostics in Deep Neural Network for High-dimensional Data

Jae Eun Lee

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

In deep neural network(DNN) hyper-parameters, set directly by the user, have significant effect on model performance and their optimization problem becomes more important. There have been many studies on hyper-parameter selection, but still limitations.

In this study, we propose a hybrid-search that combines the design of experiment(DOE) methodology and the grid search in order to overcome the limitations of the hyper-parameter optimization problem. The method presented in this study can avoid the time-consuming problem through screening in advance the hyper-parameter values using the DOE methodology. We then consider many combinations of the screened values using the grid search. Furthermore, we investigate the optimal number of hidden layers and hidden nodes in terms of the error rate for the high-dimensional data, and compare the case of all the input variables

with that of the selected input variables, changing the number of hidden layers and hidden nodes.

Finally, the impact of influential observations can be very significant for high-dimensional data where the number of variables is even larger than that of observations. In this study, we propose a deep neural network-influence box plot(DNN-influence box plot) as a graphical method for diagnosing the influential observations. It is shown that the influential observations can be diagnosed through three high-dimensional examples. Also, we compare the influence plots and the HIM plot which are useful graphical tools for diagnosing the influential observations in high-dimensional data.

Ⅰ. 서 론

심층 신경망(deep neural network, DNN)에 기반하여 만들어진 딥러닝(deep learning)은 기계가 스스로 학습할 수 있도록 만들어진 기술을 뜻한다. 비록 몇 년 전까지만 해도 영상 인식, 음성 인식, 자연어 인식 등은 해결하기 어려운 분야였다. 하지만 딥러닝을 통하여 패턴 인식이 가능하게 되면서 의료, 금융, 제조, 국방 등 다양한 분야에서 지속적으로 적용 및 발전을 해오고 있다. 딥러닝을 포함한 모든 기계학습(machine learning)에서는 초매개변수(hyper-parameter)의 값에 따라 성능이 크게 좌우된다. 초매개변수는 모델이 스스로 설정하거나 갱신하지 않고 사람이 직접 설정을 해줘야 하는 변수이며, 초매개변수의 종류로는 은닉층의 개수(hidden layer), 은닉층 내 노드의 개수(hidden node), 배치(batch)의 크기, 학습률(learning rate), 모멘텀(momentum), 활성화 함수(activation function)의 선택, 가중치 감소 시 규제 강도(regularization strength) 등이 있다. 높은 성능을 얻기 위해서는 초매개변수를 설정하는 것이 필수적인 작업이다. 이러한 작업을 초매개변수 최적화(hyper-parameter optimization) 또는 초매개변수 조율(hyper-parameter tuning)이라고 한다. 초매개변수 조율

시 아직까지 어떤 초매개변수의 값을 사용해야 하는지는 이론적으로 정해진 바가 없다. 이 문제를 해결하기 위하여 초매개변수들을 최적화 시키기 위한 다양한 연구들이 계속해서 진행되어오고 있다. 그중 매뉴얼 탐색(manual search)은 기존에 초매개변수의 조율을 통하여 얻은 경험과 지식을 활용하여 초매개변수 값을 결정하는 방법이고, 격자 탐색(grid search)은 기존 지식을 이용하여 초매개변수의 범위를 설정한 후 일정한 간격으로 값을 변경하면서 범위 내 모든 조합을 검토하는 방법이다. 임의 탐색(random search)은 범위 내 모든 조합 대신 임의로 초매개변수를 선정하는 방법, 베이지안 최적화(bayesian optimization)는 베이지안 이론과 가우시안 프로세스(gaussian process)를 이용하여 최적의 초매개변수를 결정하는 방법, 자동 선택 방법(automatic selection method)은 자동으로 최적의 초매개변수 값을 찾아주는 방법이다 (Bergstra과 Bengio, 2012; Snoek 등, 2012; Thornton 등, 2013). 그리고 Lujan-Moreno 등(2018)은 랜덤포레스트(random forest)에서 초매개변수의 조율을 위한 방법으로 실험계획법(design of experiment)을 사용하는 것을 새로이 제시하였다. 하지만 격자 탐색은 간격이 조밀하거나 탐색을 해야 할 초매개변수가 많을 경우 시간이 매우 오래 걸린다는 단점이 있다. 임의 탐색은 임의로 값을 정하고, 베이지안 최적화는 통계적 방법을 사용하기 때문에 최적의 값이라고 장담하기는 어렵다. 또한, 자동 선택 방법은 전체 자료에서 기계학습

알고리즘과 초매개변수 값의 조합을 시험하는 데 긴 시간이 걸리기 때문에 특히 유전자 자료와 같은 의학 자료를 이용할 때에는 효율성의 한계가 있다 (Luo, 2016).

최근 들어 컴퓨터 기술의 발달로 엄청난 양의 자료가 축적되면서 빅데이터 분석(big data analytics)은 오늘날 없어서는 안 될 핵심 분야 중 하나가 되었다. 빅데이터는 크게 대용량 자료(massive data)와 고차원 자료(high-dimensional data)로 나눌 수 있다. 대용량 자료는 자료의 크기 n이 차원의 크기 p 보다 훨씬 더 큰 반면 고차원 자료는 자료의 크기 n 보다 차원의 크기 p가 더 크다. 그리고 자료의 크기 n 보다 차원의 크기 p가 매우 큰 고차원 자료를 우리는 초고차원 자료(ultra-high dimensional data)라고 부른다. 이러한 고차원 자료의 특징 때문에 Bellman(1961)이나 Hastie 등(2009)은 고차원 자료에 대하여 "차원의 저주"라고 언급하며 분석의 어려움을 강조하였다. 딥러닝에서도 고차원 자료는 모델 학습을 위한 데이터가 충분하지 않고, 입력층의 개수가 많기 때문에 학습 알고리즘의 복잡도가 크다. 하지만 LeCun 등(2015)은 딥러닝이 고차원 데이터에서 복잡한 구조를 발견하는데 매우 좋은 모델이며 앞으로 많은 양의 계산과 데이터를 쉽게 이용할 수 있기 때문에 가까운 미래에 더 많은 성공을 거둘 것이라고 언급하였다. 이처럼 고차원 자료에서 딥러닝은 중요한 이슈 중 하나이며, 많은 연구가들이 이 문제에 대하여 계속해서 연구해오

고 있다.

본 논문에서는 딥러닝을 이용한 고차원 자료 분석에 대하여 크게 세 가지 주제로 나누어서 다룬다. 첫 번째 주제는 고차원 자료에서 딥러닝의 초매개변수를 탐색하는 방법이며, Lujan-Moreno 등(2018)이 제안한 실험계획법과 격자 탐색 방법을 혼합한 하이브리드 탐색(hybrid-search)을 제안하였다.

두 번째 주제는 초매개변수의 종류 중 가장 중요한 두 가지 변수인 은닉층의 수와 각 은닉층의 노드 수 탐색이다. 고차원 자료에 대하여 탐색적 자료 분석 방법을 이용하였다. 기존에는 통상적인 경험에 비추어 최적의 은닉층 수와 각 은닉층의 노드 수를 결정하였지만 그 이후 수학적인 근거에 의해 값을 찾기 위하여 Huang(2003), Shen 등(2008) 등 많은 연구들이 진행되고 있다. 하지만 아직까지 명확한 해답을 찾지는 못하였다. 특히 고차원 자료의 경우 입력층의 수가 매우 많기 때문에 앞의 방법들을 적용하기가 매우 힘들다. 이러한 점을 고려하여 탐색적 자료 분석 방법을 새로이 제시하였으며, 고차원 자료를 대상으로 심층 신경망 모델 적용 시 모든 입력 변수들을 사용하였을 때와 입력 변수 선택을 한 경우를 구분하여 두 가지 결과를 서로 비교해보았다. 입력 변수 선택의 방법으로는 Fan과 Lv(2008)이 초고차원자료에서 변수 선택을 할 때 위 상관관계(spurious correlation)로 인하여 잘못된 모형을 구축할 수 있다는 문제점을 제기하

면서 제안한 sure independence screening(SIS) 방법을 사용하였다 (Fan 등, 2016).

마지막 세 번째 주제는 딥러닝에서 고차원 자료 분석 시 영향점 (influential observation)을 확인하는 그래픽 방법을 제안한 것이다. 고차원 자료에서는 자료의 크기 n 보다 차원의 크기 p가 더 크기 때문에 딥러닝에서 자료 각각이 주는 영향이 매우 심각할 수 있다. 따라서 고차원 자료에서 회귀 계수를 추정할 때 영향점의 영향이 매우 클 수 있으므로 영향점을 진단하는 절차가 필수적으로 필요하다. Cook(1977)은 영향점에 대한 Cook의 거리통계량을 제안하였으며 Kim 등(2015)은 Cook의 거리 통계량 개념을 확장하여 LASSO 추정량 사용시 영향점을 확인할 수 있는 방법을 제시하였다. Zhao 등(2013)은 고차원 자료에서 주변상관(marginal correlation)에 기초하여 영향점을 진단하는 방법을 제시하였다. Jang과 Anderson-Cook(2017)은 LASSO 추정량 사용 시 그래픽 방법을 이용하여 영향점을 찾을 수 있는 방법을 제안하였다. 그리고 Lee와 Jang(2018a)은 고차원 자료에서 반응 변수와 모든 입력 변수들의 상관관계를 삼차원 그림으로 확인할 수 있는 삼차원 waterfall firework 그림과 LASSO 추정량을 이용하여 영향점을 찾을 수 있는 그래픽 방법인 LASSO-influence plot 및 Zhao 등(2013)이 제안한 방법을 그래픽 방법으로 제시한 HIM 그림을 제안하였다. 본 논문에서는 딥러닝에서 고차원

자료 분석 시 영향점을 확인하는 그래픽 방법으로서 심층 신경망-영향 상자그림(deep neural network-influence box plot, DNN-influence box plot)을 새로이 제안하였으며, influence plot과 HIM 그림의 분석 결과를 서로 비교해보았다.

본 논문의 구성은 다음과 같다. 2장에서는 고차원 자료, 초매개변수 탐색 방법 그리고 영향점에 대한 개념과 이론을 정리한다. 3장에서는 고차원 자료에서 초매개변수 최적화를 위한 새로운 방법으로서 하이브리드 탐색(hybrid search) 방법을 제안하였으며, 하이브리드 탐색 방법과 Lujan-Moreno 등(2018)이 제안한 방법을 서로 비교한다. 4장에서는 고차원 자료를 대상으로 심층 신경망 모델 적용 시 입력 변수를 사용하였을 때와 변수 선택을 하였을 때의 적절한 은닉층의 수와 각 은닉층의 노드 수를 탐색적 자료 분석 방법을 이용하여 확인한다. 그리고 이 두 가지 결과를 서로 비교하여 본다. 5장에서는 고차원 자료에서 영향점을 탐지하기 위한 그래픽 방법으로서 심층 신경망-영향 상자그림을 제시한다. 또한 심층 신경망-영향 상자그림을 이용하여 고차원 자료에서 영향점을 찾아보고, 영향점을 진단하는 다른 그래픽 방법들과 서로 비교해본다. 마지막으로 6장에서는 본 연구에 대한 내용을 요약하고 결론을 맺는다.

Ⅱ. 배경지식

2.1 고차원 자료

고차원 자료는 빅데이터(big data)의 한 종류이며, 변수의 수 p가 자료의 수 n 보다 큰 경우를 뜻한다. 유전자발현자료, 종양분류, 이미지 분석, 신호처리, 금융 등 다양한 분야에서 적용되고 있다(Fan과 Lv, 2008). 고차원 자료는 변수의 수 p가 자료의 수 n 보다 많기 때문에 관측값들이 퍼지게 되며 분포의 꼬리가 두꺼워지고 이상치(outlier)도 많아지게 된다. 따라서 대부분의 변수들은 소음(noise)이며 유의미한 변수들이 많지 않다고 하는 성김현상이 발생한다. 또한, 서로 관련이 없는 변수와 중요한 변수끼리 거짓 상관관계를 갖게 되는 위 상관관계(spurious correlation)의 특징도 갖게 된다(Jang 등, 2016). 이러한 문제를 해결하기 위하여 가장 많이 사용되는 분석 방법은 차원축소와 성긴 벌점함수가 있다. 대표적인 방법으로 차원축소에서는 주성분 분석(principle component analysis, PCA), 성긴 벌점함수를 이용한 회귀계수 추정에는 LASSO(Tibshirani, 1996), SCAD(Fan과 Li, 2001), MCP(Zhang, 2010), adaptive lasso(Zou, 2006), elastic net(Zou와 Hastie, 2005), dantzig selector

(Candes와 Tao, 2007), fused lasso(Tibshirani 등, 2005) 등이 있다. 그리고 Fan과 Lv(2008)는 고차원 자료에서 위 상관관계로 인하여 잘못된 변수를 선택할 수 있는 위험을 언급하며 각각의 입력 변수와 반응 변수의 상관관계를 고려한 SIS를 제안하였다.

고차원 자료에서는 여러 가지 문제점들이 발생하게 된다. 첫 번째로 고차원 자료는 대부분의 전통적인 통계기법들이 적용되기가 어렵다. 그 이유는 변수의 수 p가 자료의 수 n 보다 크기 때문에 회귀분석 시 모형행렬에 대한 적률행렬의 역행렬을 구할 수가 없어서 최소제곱 방법을 사용할 수가 없기 때문이다. 두 번째로 고차원 자료에서는 오차제곱합(SSE)이 매우 작아지게 된다. 따라서 변수 선택 및 모형선택의 기준이 되는 맬로우즈의 C_p, akaike information criterion(AIC), bayesian information criterion(BIC), R^2 통계량 기법들을 사용하는 것은 부적절하다. 세 번째로 James 등(2013)은 고차원 자료에서 과적합(over-fitting) 문제가 발생할 수 있음을 보였다. 예를 들어 관측값의 수 22개 중 2개의 이상값(outlier)이 있다고 하자. 그림 2.1.1을 보면, n=20을 사용했을 때에는 점선과 같은 추정 회귀식을 얻게 될 것이다. 하지만 n=2를 사용할 경우 실선과 같은 추정 회귀식들을 얻게 될 것이다. 즉, 두 개의 점을 이용하여 회귀식을 추정한다고 했을 때 실제로 추정해야 하는 식과는 다른 식이 추정될 것이다. 고차원 자료에서도 마찬가지로 관측값의 수보다 입력 변수의 수가 과도하게 적기 때문에 비록 훈련 자료에서는 완벽하게 적합이 될 수

있더라도 검증자료에서는 성능이 좋지 않을 것이다. 이러한 고차원 자료 분석의 어려움 때문에 현재 많은 연구가 진행되고 있지만 아직까지 고차원 자료의 구조를 밝히는 문제는 난제 중 하나이다.

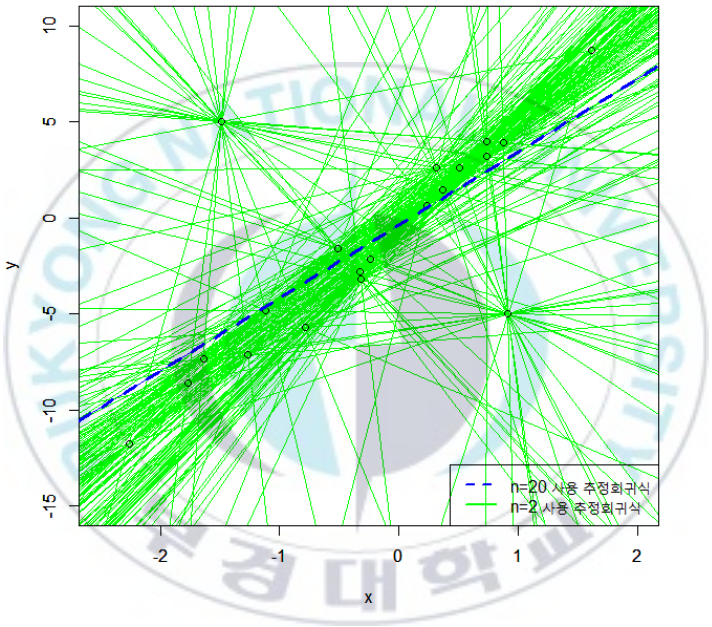

그림 2.1.1. n=2일 때와 n=20일 때 추정 회귀식

2.2 심층 신경망

2.2.1 심층 신경망의 이론적 고찰

심층 신경망은 머신러닝 알고리즘 중 하나이며, 생물학적 신경망을 모방하여 만든 인공 신경망(artificial neural network)이 여러 단계로 이루어진 구조이다. 인공 신경망은 하나 이상의 은닉층(hidden layer)을 포함하는데, 층이 많은 신경망 학습을 딥러닝이라 부른다.

단층 신경망의 구조는 d개의 입력 $\boldsymbol{x}=(x_1, x_2, \cdots, x_d)'$에 대하여 다음과 같이 나타낼 수 있다.

$$y=f\left(\sum_{i=1}^{d} w_i x_i + b\right) \tag{2.2.1}$$

여기서 x_i는 입력층의 i번째 노드의 입력, w_i는 입력층의 i번째 노드의 가중치, b는 편향(bias), f는 활성화 함수, y는 출력값이다. 그림 2.2.1은 단층 신경망의 구조를 보여준다. 그림 2.2.2는 여러 개의 중간층이 있는 심층 신경망의 구조를 나타낸다. 연결 강도는 가중치 $\boldsymbol{w}$에 의하여 결정되며,

가중치 $\boldsymbol{w}$와 절편 b는 학습을 통하여 오차제곱합이 최소가 되는 방향으로 갱신(update)이 된다. 최종 목표값(target value) y는 비선형 활성화 함수를 적용하여 구해진다.

신경망은 연속형, 범주형 변수에 상관없이 분석이 가능하며, 입력 변수들 간의 비선형 조합이 가능하다. 특히 입력 변수와 출력 변수 간에 복잡한 비선형 관계가 존재하거나 변수의 수가 많은 경우 유용하며, 의사결정나무와 회귀분석보다 분류 및 예측력 면에서 뛰어나다고 평가받고 있다. 하지만 신경망은 비선형성과 가중치에 대한 정확한 해석이 어렵기 때문에 입력 정보와 출력 정보 사이에 완벽한 관계를 설명하기가 거의 불가능하다. 또한, 최적의 초매개변수를 결정하기가 쉽지 않고, 초깃값에 따라 전역해가 아닌 지역해로 수렴할 수 있으며 모형이 복잡할 경우 계산시간이 길어지게 된다.

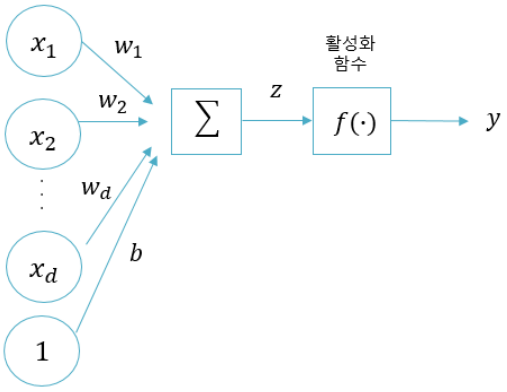

그림 2.2.1. 단층 신경망의 구조

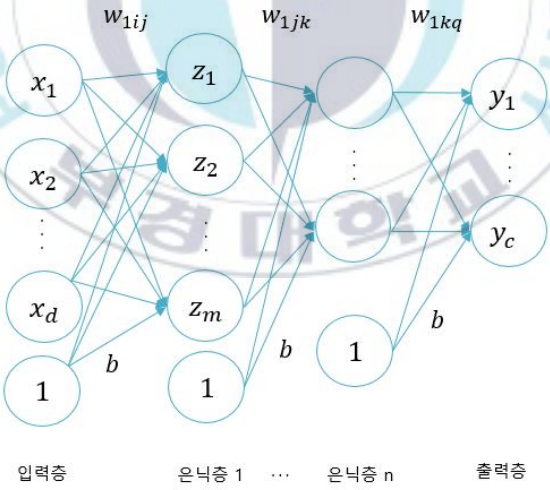

그림 2.2.2. 심층 신경망의 구조

2.2.2 심층 신경망에서의 초매개변수

머신러닝 학습을 할 때 효과가 더 좋도록 하기 위하여 사람이 직접 설정해줘야 하는 변수를 초매개변수라고 한다. 초매개변수 값은 이론적으로 정해진 것이 아니라서 최적값을 찾는 것이 중요하다.

심층 신경망에서 초매개변수들의 종류로는 은닉층의 개수, 은닉층 내 노드의 개수, 학습률, 모멘텀, 배치의 크기, epoch 횟수, 활성화 함수의 선택, 가중치 감소 시 규제 강도, 역전파 알고리즘의 종류 등이 있다.

앞먹임 신경망(feed-forward neural network)에서 학습 자료가 다음과 같이 N개가 주어졌다고 하자.

$$D_N = \{(\boldsymbol{x}^{(d)}, y^{(d)})\}_1^N \tag{2.2.2}$$

여기서 $\boldsymbol{x_i}$는 i번째 학습 자료에서 입력 벡터이고 y_i는 i번째 학습 자료에서 목표 출력이다. 그러면 오차함수 $E(\boldsymbol{w})$는 다음과 같이 정의된다.

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{d=1}^{N}\left(y^{(d)} - f^{(d)}\right)^2 \tag{2.2.3}$$

여기서 $f^{(d)} = f(\boldsymbol{x}^{(d)};\boldsymbol{w})$는 활성화 함수이고 $\boldsymbol{w}$는 가중치 벡터이다. 우리의 목표는 이 오차함수의 최솟값을 구하는 것인데 복수 개의 국소 극소점(local minimum) 중 하나의 점 $\boldsymbol{w}$에서의 오차함수 $E(\boldsymbol{w})$의 값이 충분히 작으면 분류(classification)문제나 예측(prediction)문제를 해결할 수 있다. 하나의 국소 극소점은 $\boldsymbol{w}$의 초기점을 출발점으로 하여 경사하강법(gradient descent method)을 사용하여 다음과 같이 반복적으로 갱신한다.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \triangle \boldsymbol{w}_t \qquad (2.2.4)$$

$$\triangle \boldsymbol{w}_t = -\eta \frac{\partial E}{\partial \boldsymbol{w}_t}$$

여기서 $\boldsymbol{w}_t$는 현재의 가중치 벡터, $\boldsymbol{w}_{t+1}$은 현재의 $\boldsymbol{w}_t$를 음의 기울기 방향으로 조금씩 움직인 후의 가중치 벡터이고 η는 $\boldsymbol{w}_t$의 갱신량의 크기를 결정하는 값으로 학습률이라고 부르는 초매개변수이다. 학습률은 0~ 1 사이 값을 가지나 통상 작은 값을 이용한다.

앞먹임 신경망에서 하나의 학습 자료만을 사용하여 가중치를 갱신하는 방법이 확률적 경사 하강법(stochastic gradient descent, SGD)이다. 큰 규모의 신경망 학습은 대규모의 계산 비용이 들기 때문에 확률적 경사 하

강법을 바로 적용할 수 없으므로 효율적인 수치계산을 위하여 복수 개의 학습 자료를 묶어 확률적 경사 하강법을 적용하는 데 이러한 복수 개의 학습 자료에 대한 묶음 단위를 미니배치(mini-batch)라 한다. 미니배치의 크기는 확률적 경사 하강법의 장점과 병렬 계산 자원의 유효한 이용을 고려하여 결정한다. 이러한 미니배치의 크기가 하나의 초매개변수가 된다.

하나의 미니배치를 D_t라 하자. 여기서 t는 미니배치가 갱신되는 첨자를 가리킨다. t번째 갱신마다 D_t에 포함되어 있는 모든 학습 자료에 대한 오차 E_t를 다음과 같이 정규화하여 계산하고 오차 E_t의 기울기 방향으로 가중치를 갱신한다.

$$E_t(\boldsymbol{w}) = \frac{1}{n_t} \sum_{i \in D_t} E_i(\boldsymbol{w}) \tag{2.2.5}$$

여기서 $n_t = |D_t|$는 미니배치에 포함되는 학습 자료의 개수이고 $\boldsymbol{w}$는 가중치 벡터를 나타낸다.

우리는 경사하강법의 수렴 성능을 향상시키는 초매개변수로서 모멘텀을 사용할 수 있다. 이 방법은 다음과 같이 가중치의 갱신값에 이전 갱신값의 일정비율을 더하는 방법이다.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \triangle \boldsymbol{w}_t + \mu \triangle \boldsymbol{w}_{t-1}$$

$$\triangle \boldsymbol{w}_t = -\eta \frac{\partial E}{\partial \boldsymbol{w}_t} \tag{2.2.6}$$

$$\triangle \boldsymbol{w}_{t-1} = \boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2}$$

여기서 μ는 모멘텀이고 0~ 1 사이 값을 가지지만 통상 $\mu = 0.5 \sim 0.9$ 사이의 값을 주로 사용한다.

오차함수가 깊은 골짜기와 같은 모습을 띄는 경우 경사하강법을 사용할 때 골짜기가 깊어 골짜기 바닥을 조금 빠져나오면 골짜기와 직교하는 방향으로 큰 기울기가 형성되고 가중치는 매번 골짜기와 직교하는 방향으로 갱신되어 결로가 지그재그(zigzag) 모양을 형성하여 골짜기 바닥을 정상적으로 탐색하지 못한다. 이때 우리는 모멘텀을 사용하면 골짜기 방향을 따라 골짜기 바닥을 효과적으로 탐색할 수 있다.

과적합(overfitting)을 완화시키는 초매개변수는 규제화, 가중치 상한 그리고 드롭아웃(dropout)이 있다. 규제화는 학습 시 과적합을 방지하기 위하여 가중치의 자유도를 제약하는 방법으로써 오차함수에 다음과 같은 능형(ridge) 타입의 L^2-norm 벌점(penalty)을 부여한 뒤 이를 최소화하는 방법이다.

$$E_t(\boldsymbol{w}) = \frac{1}{n_t}\sum_{i \in D_t} E_i(\boldsymbol{w}) + \frac{\lambda}{2} \| \boldsymbol{w} \|^2 \qquad (2.2.7)$$

여기서 λ는 규제화의 정도를 나타내는 초매개변수이다. 일반적으로 λ =0.00001 ~ 0.01 범위 내에서 선택한다. 이 항으로 인하여 학습 시 더 작은 가중치가 쓰이게 된다. 우리는 능형(ridge) 타입의 L^2-norm 벌점 대신 라쏘(lasso) 타입의 L^1-norm 벌점을 사용할 수 있다. 가중치는 자신의 크기에 비례하는 속도로 항상 감쇠 하도록 갱신되므로 가중치 감쇠(weight decay)라 부른다.

가중치 상한은 각 노드의 입력층 결합의 가중치에 대하여 최대놈 제약(max−norm constraint)을 사용하는 방법이다. 층 l의 j번째 노드가 층 $l-1$의 I개의 노드들의 출력을 입력으로 받을 때 그 사이의 가중치 w_{ij}가 다음 조건을 만족하도록 가중치를 제약한다.

$$\sqrt{\sum_{i=1}^{I} w_{ij}^2} \; < c \qquad (2.2.8)$$

이 방법은 비교적 최근에 제안된 방법으로서 가중치 감쇠보다 뛰어난 효과를 가지며 드롭아웃과 함께 사용하면 높은 효과를 얻는다.

드롭아웃은 학습 과정에서 신경망의 노드 중 일부를 일정 비율로 배제시킨다. 드롭아웃된 노드들은 일시적으로 다음 층 노드 출력을 보내지 않거나 역전파 단계에서 가중치 수정의 대상에서 제외된다. 학습 단계에서 학습 세대당 층별 드롭 아웃 비율을 p라 하자. 이 비율 p가 하나의 초매개변수가 된다. 이 비율 p만큼 노드의 연결을 끊는다. 검증 단계에서 노드의 활성값을 $q=1-p$ 비율만큼 재조정한다.

신경망에서 입력층으로부터 출력층까지 뉴런이 순차적으로 활성화되며 뉴런의 가중치와 활성화 함수가 적용되는 단계를 순방향 단계(forward phase)라 하고, 순방향 단계에서 생성된 네트워크 출력을 훈련 자료의 실제 목표값과 비교하여 오차를 계산하면 네트워크에서 역방향으로 전파되어(backpropagation) 뉴런 사이의 가중치를 수정하고 오차를 줄이는 단계를 역방향 단계(backward phase)라 한다. 이러한 두 과정을 여러 번 순환하여 반복하는 데 각 순환(cycle)을 주기(epoch)라 부른다. 이러한 주기의 반복 횟수가 하나의 초매개변수가 된다.

활성화 함수는 입력 신호의 총합이 활성화를 일으킬지 정하는 역할이며, 역전파 알고리즘에서는 미분이 필요하기 때문에 연속함수이어야 한다. 로지스틱 함수로도 불리는 시그모이드(sigmoid) 함수는 다음과 같으며 결과는 $0 \leq y \leq 1$인 연속형이다.

$$y = \frac{1}{1+\exp(-z)} \tag{2.2.10}$$

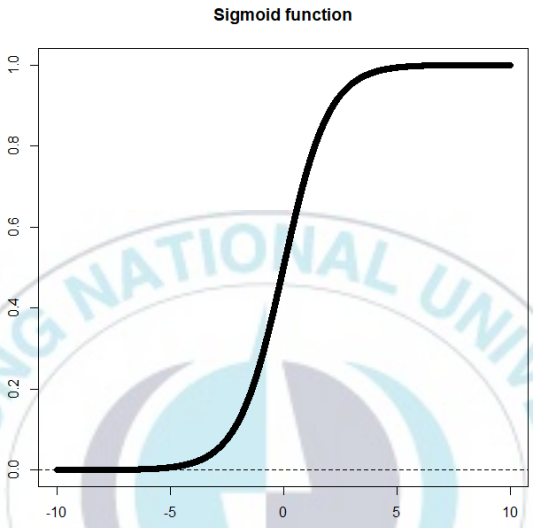

그림 2.2.3. 시그모이드 함수

하이퍼볼릭 탄젠트(hyperbolic tangent, tanh) 함수는 다음과 같으며 $-1 \leq y \leq 1$인 연속형이다.

$$y = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)} \tag{2.2.11}$$

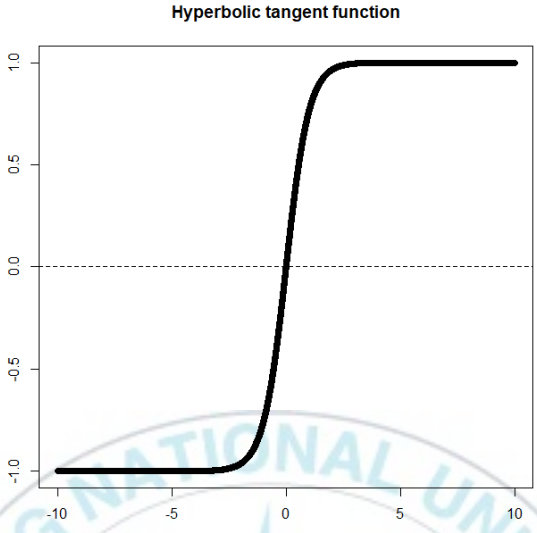

그림 2.2.4. 하이퍼볼릭 탄젠트 함수

시그모이드 함수와 하이퍼볼릭 탄젠트 함수는 성능이 어느 정도 보장되었기 때문에 오랫동안 많이 사용되어왔다. 하지만 은닉층의 수가 많아질수록 기울기가 점점 작아져서 사라지는 vanishing gradient 문제점이 있다. 이를 보완하기 위하여 렐루(rectified linear unit, ReLU) 함수가 제시되었으며 다음과 같이 정의한다.

$$y = \max(z,0) = \begin{cases} z\,(z \geq 0) \\ 0\,(z < 0) \end{cases} \qquad (2.2.12)$$

$z \leq 0$일 때 출력값은 항상 0이고, $z > 0$일 때에는 1이다. 이러한 특성 때문에 신경망의 깊이가 깊어지더라도 기울기가 사라지지 않게 되며, 시그모이드, 탄젠트 함수보다 학습이 훨씬 빠르고 극솟값으로 수렴되는 문제가 발생하지 않는다. 하지만 입력 변수가 0 이하일 때 출력값이 0이 되어 해당 노드의 weight 업데이트가 되지 않는 dying ReLU 문제가 발생하게 된다. 이러한 ReLU의 문제점을 극복하기 위하여 리키 렐루(leaky rectified linear unit, Leaky ReLU), PReLU(parametric ReLU), SELU(scaled exponential linear unit), ELU(exponential linear unit), GELU(gaussian error linear unit) 등이 제안되었다.

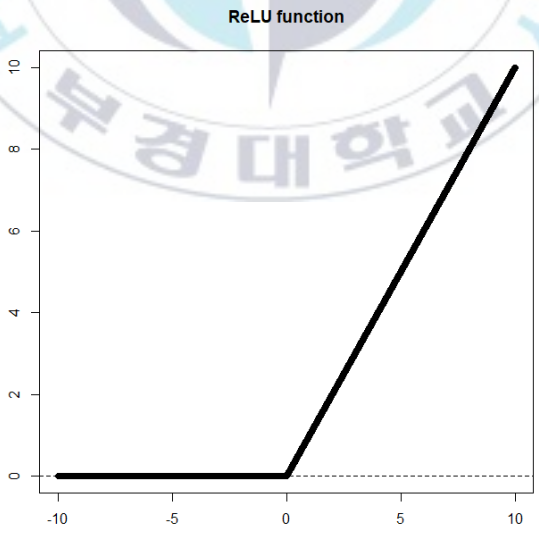

그림 2.2.5. 렐루 함수

소프트맥스(softmax) 함수는 출력값 z가 여러 개(L개)로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후 확률이 제공되며 형태는 다음과 같다.

$$y = \frac{\exp(z_j)}{\sum_{i=1}^{L}\exp(z_i)}, \quad j = 1,...,L \tag{2.2.14}$$

2.2.3 초매개변수 조율을 위한 방법들

머신러닝에서 모델을 학습시키기 위해서는 복수의 초매개변수들을 설정해주어야 한다. 이때 모델의 성능은 초매개변수의 값에 따라 크게 좌우되기 때문에 최적의 값을 찾기 위한 작업이 필수적이다. 이러한 최적의 초매개변수들을 찾기 위하여 다양한 연구들이 진행되어왔다.

격자 탐색(grid search)은 먼저 알고 있는 지식을 이용하여 초매개변수의 범위를 설정한 뒤, 일정한 간격으로 값을 변경하면서 CV(cross validation)를 이용하여 평균 오차를 계산한다. 모든 경우 중 평균오차가 최소가 되는 조합을 찾는 방법이다. 임의 탐색은 격자탐색과 마찬가지로 먼저 알고 있는 지식을 이용하여 초매개변수의 범위를 설정한 후, 난수를

이용하여 복수 개의 초매개변수 조합을 골라서 평균오차가 최소가 되는 초매개변수 조합을 찾아보는 방법이다. 베이지안 최적화는 베이지안 이론과 가우시안 프로세스(gaussin process)를 이용하여 초매개변수를 결정하는 방법이며, 자동선택 방법은 자동으로 다양한 초매개변수 조건에서 교차검증을 이용하여 최적의 값을 찾아주는 방법이다. 또한, 통계적 방법으로써 Lujan-Moreno 등(2018)은 실험계획법을 이용하여 초매개변수 값을 먼저 선별(screening)한 뒤, 반응표면분석(response surface methodology, RSM)을 이용하여 초매개변수 값을 조율하는 방법을 제안하였다.

2.3 영향점

영향점이란 입력변수의 관측값 범위를 벗어났으며, 제거 시 회귀계수 변화에 큰 영향을 미치는 점을 뜻한다. 영향점은 회귀분석에서 추정량 $\hat{\beta}$, s^2 등에 영향을 미치기 때문에 회귀계수 유의성을 매우 높이고 결정계수 값을 과하게 높일 수 있어서 회귀진단 단계에서 반드시 확인을 해야 한다. 이러한 관측치의 영향력을 판단하기 위한 방법으로는 크게 소거법(deletion method), 무한소 교란법(infinitesimal perturbation method), 국소 영향력

(local influence), 대치법(replacement method) 네 가지가 있다. 그중 본 연구에서 적용하고자 하는 소거법은 i번째 관측치의 영향력을 판단하기 위하여 $n-1$개의 관측치로 추정한 추정치와 전체 n개의 관측치로 추정한 추정치의 차이로 영향점을 판단한다(김과 강, 2010). 영향점을 진단하기 위한 대표적인 방법으로 수치적인 측도를 제안한 Andrews와 Pregibon(1978), Belsley 등(1980), 소거법을 이용하여 쿡 통계량(Cook's distance)을 제안한 Cook(1977)이 있으며, Kim 등(2015)은 LASSO 회귀에서, Zhao 등(2013)은 고차원 자료에서 주변상관 소거법을 이용하여 영향점을 진단하는 방법을 제안하였다. Jang과 Anderson-Cook(2017)은 LASSO 회귀에서 영향점의 영향을 평가하기 위한 방법으로 소거법을 이용한 그래픽 방법을 제안하였다. 즉, 전체 관측값을 이용하여 그린 그림과 관측값을 하나씩 제거한 뒤 그린 그림을 비교하였을 때 유의한 변수가 다르거나 추정된 회귀계수 값에 대한 그림이 큰 차이가 있을 때 그 관측값은 영향점이라고 판단할 수 있다. 또한 Lee와 Jang(2018a)은 Zhao 등(2013)이 주변 상관(marginal correlation)에 기초하여 제안한 고차원적 영향력 측도(high-dimensional influence measure, HIM)를 이용하여 그림으로 영향점을 진단할 수 있는 그래픽 방법인 HIM plot을 제시하였다.

Ⅲ. 고차원 자료에서 초매개변수 조율을 위한 하이브리드 탐색

3.1 하이브리드 탐색

초매개변수는 경험적으로 설정되는 값으로써 높은 모델 성능을 위하여 최적의 초매개변수 값을 설정하는 것이 매우 중요하다. 또한, 초매개변수는 오차함수와 역전파를 이용하여 수정할 수 있는 weight, bias와 달리 수정을 할 수가 없기 때문에 초매개변수의 최적값을 찾기 위한 연구가 활발하게 이루어지고 있다. 초매개변수를 최적화하기 위한 방법들 중 하나인 격자 탐색은 범위 내의 모든 조합을 고려하기 때문에 전체적인 탐색이 가능하다는 장점이 있다. 하지만 간격이 너무 조밀하거나 탐색을 해야 할 초매개변수가 많다면 탐색 시간이 기하급수적으로 증가하는 단점이 있다. Lujan-Moreno 등(2018)은 통계적 방법론을 이용한 초매개변수 최적화 방법으로써 실험계획법을 이용하여 초매개변수를 미리 선별하는 방법을 제시하였다.

본 연구에서 제안하는 하이브리드 탐색은 실험계획법과 격자탐색법을 혼합

한 방법을 이용하여 최적의 초매개변수를 찾기 위한 방법이다. 분석 과정은 algorithm 3.1에서 설명된다.

Algorithm 3.1. 하이브리드 탐색(hybrid search)

1. 각 초매개변수들의 낮은 수준과 높은 수준을 정하여 2^n 완전 요인설계 (n=초매개변수의 수)를 한다.
2. (1) 효과의 반정규 확률 플롯 및 Pareto 차트를 평가하여 유의한 초매개변수를 확인한다.
   (2) 주효과 그림을 통하여 두 수준 사이의 오류율 (또는 평균제곱오차) 차이를 확인한다.
   (3) 유의한 초매개변수들을 대상으로 낮은 오류율 (또는 평균제곱오차)을 갖는 수준 조합을 찾는다.
3. 선별된 수준 조합을 중심으로 격자 탐색을 한다.
4. 격자탐색의 결과를 각종 그래프 방법들 (평행좌표그림, 산점도 행렬, 히스토그램 등)을 통하여 그려보고 오류율 (또는 평균제곱오차)이 작은 결과 값들을 찾아 최적의 초매개변수 값을 결정한다.

이때 오류율(error rate)은 아래의 표 3.1.1과 같이 혼동행렬(confusion matrix)로부터 구할 수 있다. 여기서 y_i는 실제 값, $\hat{y}_i$는 예측 값을 뜻한다. 오류율은 식 3.1.1과 같다.

$$error\ rate = \frac{FP + FN}{TP + TN + FN + FP} \tag{3.1.1}$$

표 3.1.1. 혼동행렬

	$y_i = 1$(positive)	$y_i = 0$(negative)
$\hat{y}_i = 1$(positive)	TP(true positive)	FP(false positive)
$\hat{y}_i = 0$(negative)	FN(false negative)	TN(true negative)

3.2 심층 신경망의 초매개변수 탐색 및 분석

본 연구에서는 세 가지 고차원 자료를 이용하여 분석을 시행하였다. 첫 번째 예제는 Golub 등(1999)이 소개한 백혈병 자료(leukemia cancer data)이며 오픈 소스 소프트웨어인 R의 SIS 패키지에 내장되어있는 자료이다. 훈련 자료와 테스트 자료에 대하여 각각 leukemia.train과 leukemia.test 자료로 제공된다. leukemia train 자료에는 27명의 급성 림프 모구 백혈병과 11명의 급성 골수성 백혈병, leukemia.train 자료에는 20명의 급성 림프 모구 백혈병과 14명의 급성 골수성 백혈병으로 구성되어 있다. 설명변수는 7,129 유전자이고 반응 변수는 $y=0$ 또는 $y=1$이다(급성 림프 모구 백혈병 / 급성 골수성 백혈병). 일반적으로 학습 자료와 테스트 자료를 세분화 할 때 파레토 법칙(pareto principle)에 따라 8:2로 분류한다. 하지만 고차원 자료에 이 법칙을 적용할 경우 테스트 자료가 매우 적을 수 있다. 본 연구에서는 leukemia.train 자료와

leukemia.test 자료를 합친 후, 72개의 관측값에 대하여 학습 자료 40개와 테스트 자료 32개를 랜덤하게 선택하여 분석하였다.

두 번째 예제는 Singh 등(2002)이 소개한 전립선 암 자료(prostate cancer data)이며, R의 SIS 패키지에 내장되어 있는 자료이다. 학습 자료와 테스트 자료가 각각 prostate.train과 prostate.test로 제공되어 있다. prostate.train 자료에는 52명의 전립선 종양환자와 50명의 전립선 비 종양환자가 있으며, prostate.test 자료에는 25명의 전립선 종양환자와 9명의 전립선 비 종양환자로 구성되어 있다. 설명변수는 12,600 유전자이며, 반응 변수는 $y=0$ 또는 $y=1$이다(전립선 종양 환자 / 전립선 비 종양 환자). 본 연구에서는 prostate.train과 prostate.test 자료를 합친 후, 136개의 관측값 중 학습 자료 100개와 테스트 자료 36개를 랜덤하게 선택하여 분석하였다.

세 번째 예제는 Alon 등(1999)이 제공한 대장암 자료(colon cancer data)이다. R의 datamicroarray 패키지에 alon 자료로 내장되어있다. 종양 조직 40명, 정상 조직 22명이고 2,000 유전자의 설명변수가 있으며 반응 변수는 $y=0$ 또는 $y=1$이다(종양 조직 / 정상 조직). 62개의 관측값에 대하여 학습 자료와 테스트 자료를 7:3으로 세분화시켜 학습 자료 43개, 테스트 자료 19개를 랜덤하게 선택하여 분석하였다.

본 연구에서 사용한 컴퓨터 언어는 오픈 소스 소프트웨어인 R을 이용하여 분석하였다. 백혈병 자료와 전립선 암 자료는 deepnet 패키지 함수를

표 3.2.1. (백혈병 자료) 2^5 완전 요인 설계(DNN)

요인	낮은 수준(−)	높은 수준(+)
A(첫 번째 은닉층의 노드 수)	20	30
B(두 번째 은닉층의 노드 수)	20	30
C(학습률)	0.01	0.1
D(모멘텀)	0.4	0.74
E(활성화 함수)	sigm	tanh

사용하였고, 대장암 자료는 MXNet 패키지 함수를 사용하였다.

백혈병 자료에 대하여 은닉층의 수가 2개일 때 5개의 초매개변수인 첫 번째 은닉층의 노드 수, 두 번째 은닉층의 노드 수, 학습률, 모멘텀, 활성화 함수를 고려하였으며 표 3.2.1과 같이 2^5 완전 요인 설계를 하였다. 분산분석 결과, R^2=0.993이며 그림 3.2.1의 표준화된 효과의 1/2정규 확률도와 Pareto 차트로부터 첫 번째 은닉층의 노드 수($p=0.00173$)와 활성화 함수($p=0.00154$)가 유의한 변수로 판단된다. 주효과도에서는 첫 번째 은닉층의 노드 수 30, 두 번째 은닉층의 노드 수 20, 학습률 0.1, 모멘텀 0.4, 활성화 함수 tanh일 때 낮은 오류율을 갖는다. 이 결과를 바탕으로 표 3.2.2와 같이 각 초매개변수 별 범위를 정한 후 격자탐색을 하였다.

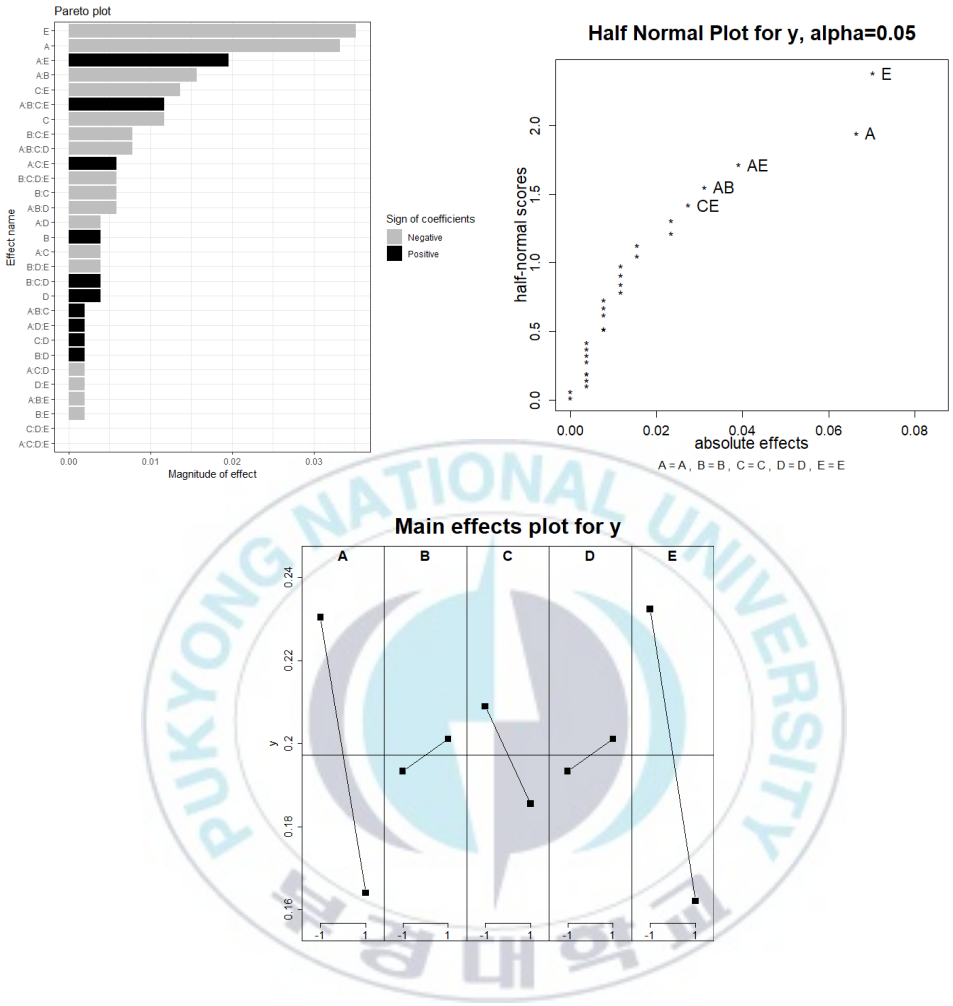

그림 3.2.1. (백혈병 자료) Pareto 차트, 표준화된 효과의 1/2정규 확률도, 주효과도(DNN)

표 3.2.2. (백혈병 자료) 격자 탐색으로 살펴볼 최적 조건 범위(DNN)

요인	실험계획법 이용한 최적 조건	격자 탐색으로 살펴볼 최적 조건 범위
A(첫 번째 은닉층의 노드 수)	30	(25, 40), 5씩 증가
B(두 번째 은닉층의 노드 수)	20	(5, 20), 5씩 증가
C(학습률)	0.1	(0.05, 0.15), 0.01씩 증가
D(모멘텀)	0.4	(0.2, 0.6), 0.1씩 증가
E(활성화 함수)	tanh	

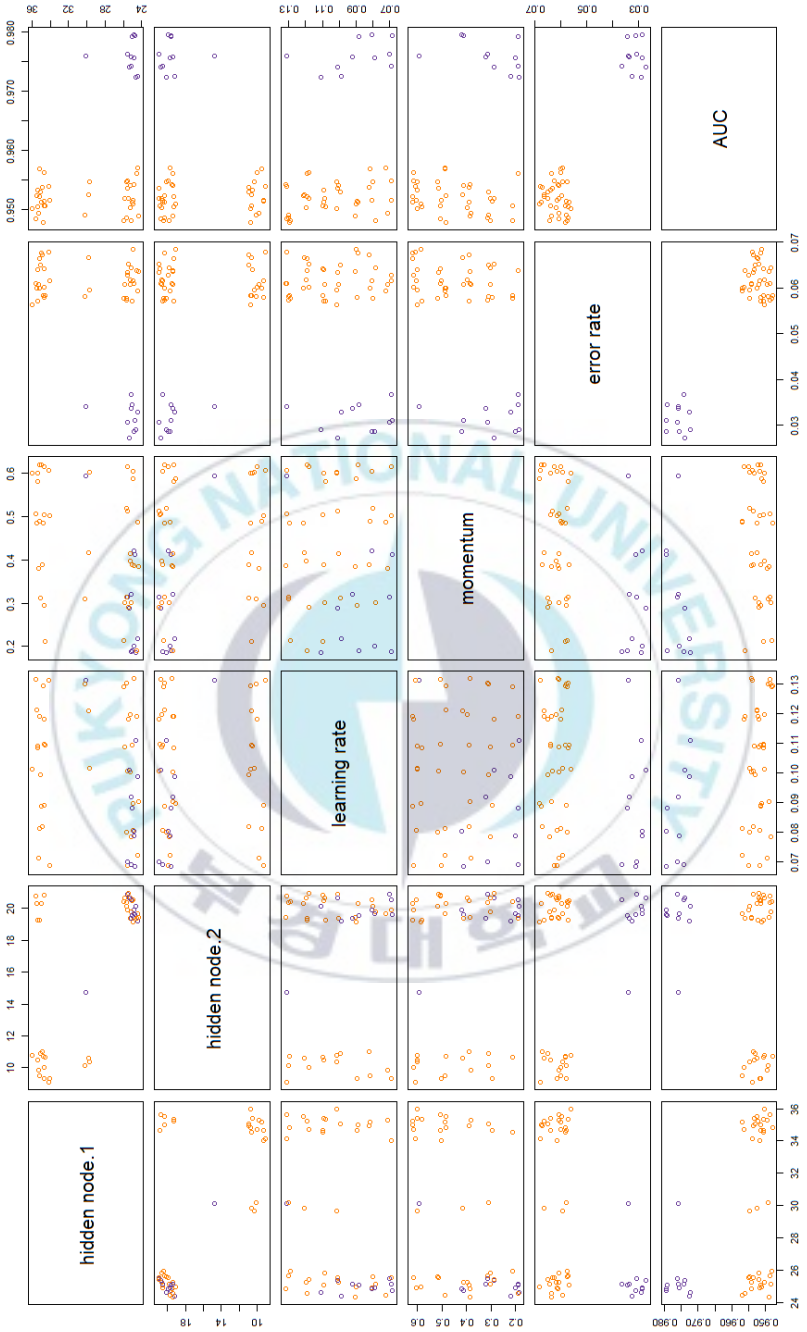

그림 3.2.2. (백혈병 자료) 격자 탐색 결과에 대한 산점도 행렬(DNN)

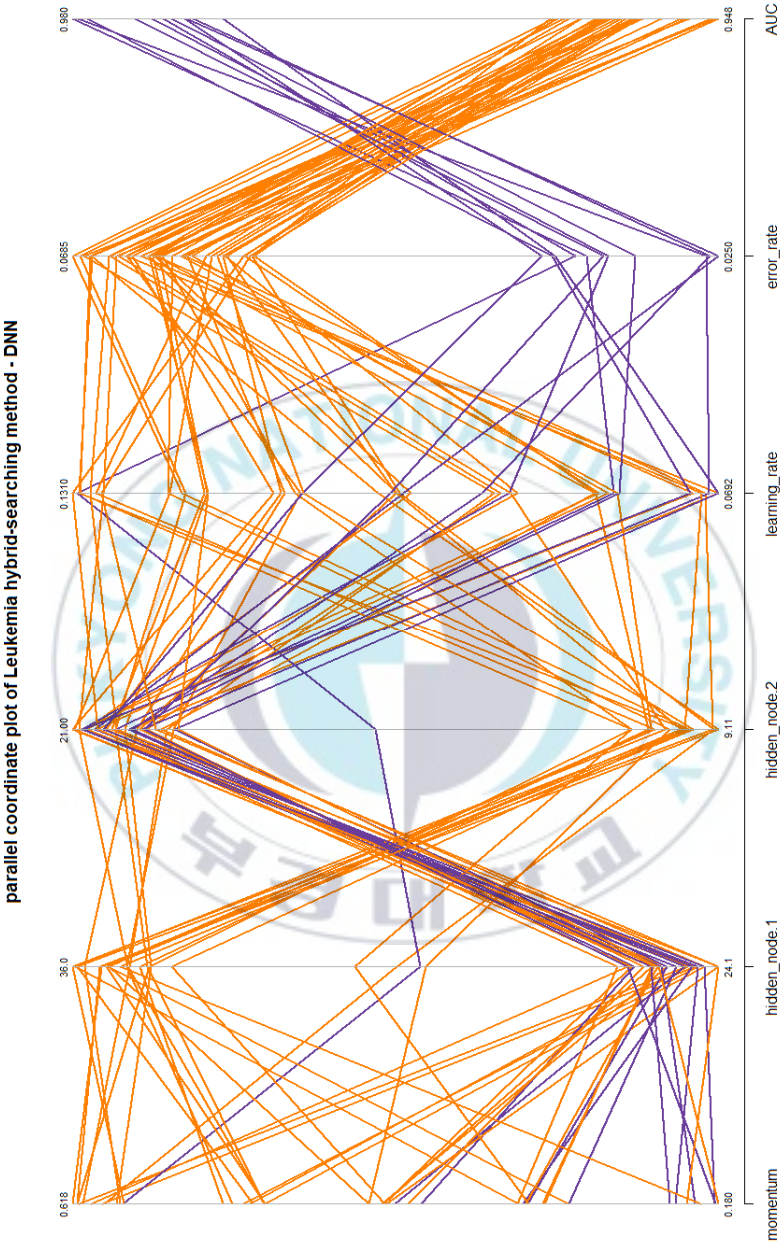

그림 3.2.3. (백혈병 자료) 격자 탐색 결과에 대한 평행좌표그림(DNN)

격자탐색의 결과 오류율의 범위는 0.03125~0.21875이었으며, 그 중 오류율이 작은 값들인 0.03125 ~ 0.0625에 대하여 분석하였다. 그림 3.2.2~3.2.3은 오류율이 작은 값들에 대한 산점도 행렬과 평행좌표그림이다. 그림에서 모여 있는 값이 구분되도록 jitter함수를 사용하였다. 그림 3.2.2~ 3.2.3을 보면 첫 번째 은닉층의 노드 수가 25, 두 번째 은닉층의 노드 수가 20일 때 오류율이 가장 많이 분포되어 있다. 학습률과 모멘텀은 전반적으로 고르게 분포되어 있지만 각각 0.07과 0.2일 때 오류율이 0.03125(보라색)인 경우가 조금 더 많다. 따라서 하이브리드 탐색을 이용한 최적 조건은 첫 번째 은닉층의 노드 수 25, 두 번째 은닉층의 노드 수 20, 학습률 0.07, 모멘텀 0.2로 나타났다.

하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건을 적용한 모형에 대하여 100번 반복 분석한 결과를 비교 분석하였다. 각 방법에 대한 검증 지표로 오류율, 정확도, AUC, 계산시간을 사용하였다. 실험계획법으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.1256, 평균 정확도 0.8744, 평균 AUC 0.8950, 평균 계산시간 14.17초이었다. 반면, 하이브리드 탐색으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.1156, 평균 정확도 0.8844, 평균 AUC 0.9036, 평균 계산시간 12.15초로 나타났다. 그림 3.2.4를 통하여 하이브리드 탐색으로 찾은 최적 조건들이 모형 성능을 개선시키는데 효과가 있다는 것을 확인할 수 있다.

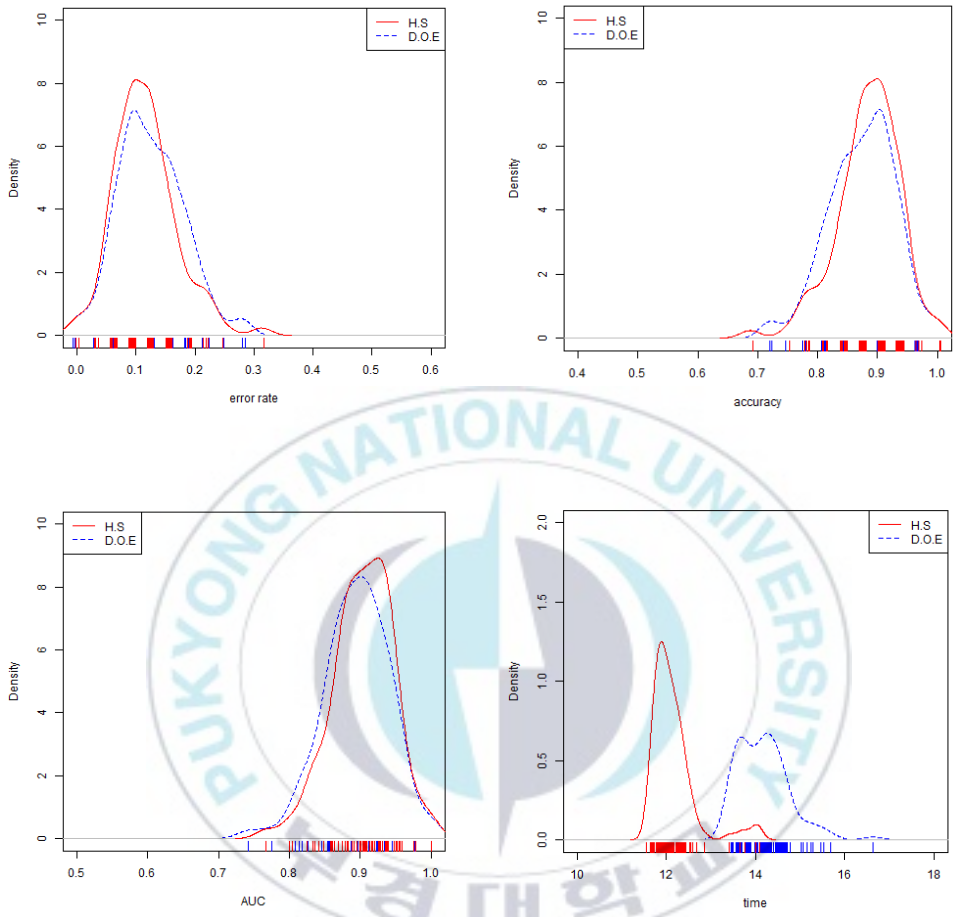

그림 3.2.4. (백혈병 자료) 하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건 결과 비교(DNN)

전립선 암 자료에 대하여 은닉층의 수가 2개일 때 5개의 초매개변수인 은닉층의 노드 수, 학습률, 모멘텀, 활성화 함수를 고려하였으며 표 3.2.3과 같이 2^5 완전 요인 설계를 하였다. 분산분석 결과, R^2=0.9924이며 그림 3.2.5의 표준화된 효과의 1/2정규 확률도와 Pareto 차트로부터 활성화 함수($p=0.013666$)가 유의한 변수로 판단된다. 주효과도에서는 첫 번째 은닉층의 노드 수 30, 두 번째 은닉층의 노드 수 20, 학습률 0.1, 활성화 함수 tanh일 때 낮은 오류율을 갖는다. 이 결과를 바탕으로 표 3.2.4와 같이 각 초매개변수 별 범위를 정한 후 격자탐색을 하였다.

표 3.2.3. (전립선 암 자료) 2^5 완전 요인 설계(DNN)

요인	낮은 수준(−)	높은 수준(+)
A(첫 번째 은닉층의 노드 수)	20	30
B(두 번째 은닉층의 노드 수)	20	30
C(학습률)	0.01	0.1
D(모멘텀)	0.4	0.74
E(활성화 함수)	sigm	tanh

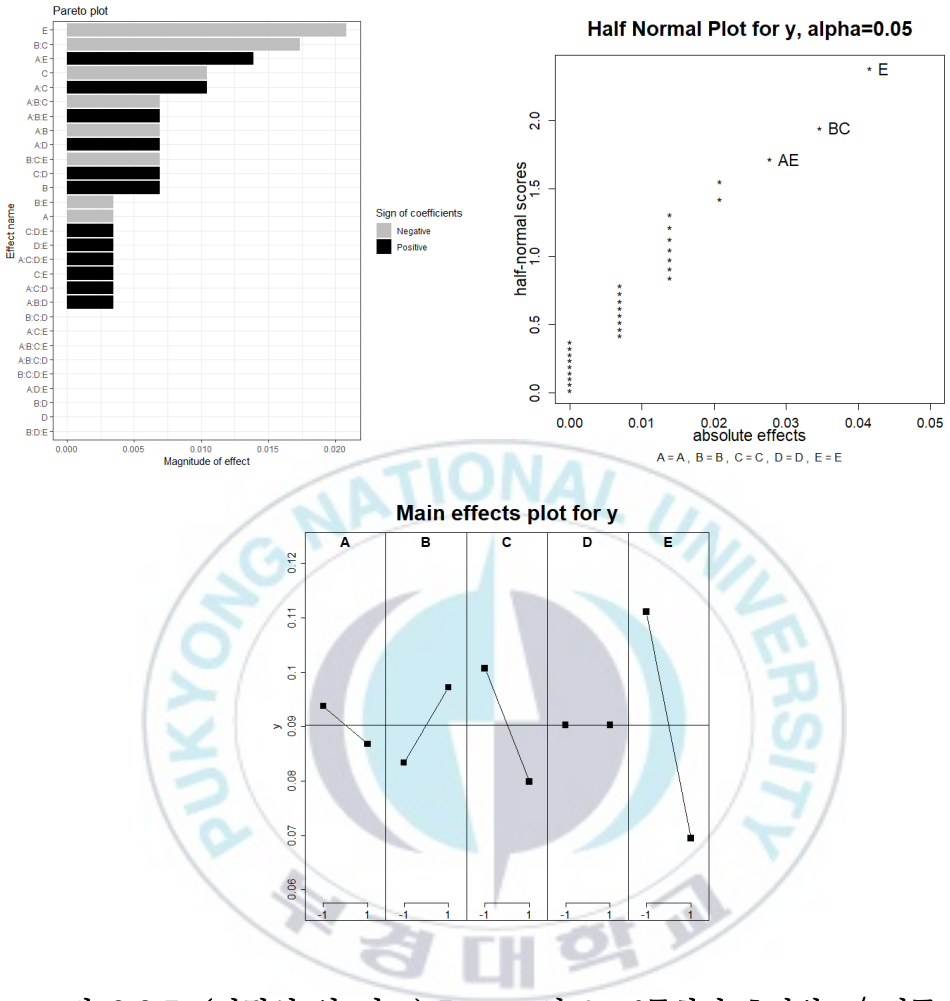

그림 3.2.5. (전립선 암 자료) Pareto 차트, 표준화된 효과의 1/2정규 확률도, 주효과도(DNN)

표 3.2.4. (전립선 암 자료) 격자 탐색으로 살펴볼 최적 조건 범위(DNN)

요인	실험계획법 이용한 최적 조건	격자 탐색으로 살펴볼 최적 조건 범위
A(첫 번째 은닉층의 노드 수)	30	(25, 35), 5씩 증가
B(두 번째 은닉층의 노드 수)	20	(10, 20), 5씩 증가
C(학습률)	0.1	(0.07, 0.13), 0.01씩 증가
D(모멘텀)	0.4	(0.2, 0.6), 0.1씩 증가
E(활성화 함수)	tanh	

그림 3.2.6. (전립선 암 자료) 격자 탐색 결과에 대한 산점도 행렬(DNN)

격자탐색의 결과 오류율의 범위는 0.05556~0.19444이며, 그 중 오류율이 작은 값들인 0.05556 ~ 0.08333에 대하여 분석하였다. 그림 3.2.6은 오류율이 작은 값들에 대한 산점도 행렬이다. 그림 3.2.6을 보면 첫 번째 은닉층의 노드 수가 35, 두 번째 은닉층의 노드 수가 10과 20일 때 오류율이 가장 많이 분포되어 있으며 학습률과 모멘텀은 전반적으로 고르게 분포되어있다. 두 번째 은닉층의 수, 학습률, 모멘텀에 대한 빈도분석을 해본 결과, 두 번째 은닉층의 수는 20, 학습률은 0.13, 모멘텀은 0.4일 때 빈도수가 가장 많았다. 따라서 하이브리드 탐색을 이용한 최적 조건은 첫 번째 은닉층의 노드 수 35, 두 번째 은닉층의 노드 수 20, 학습률 0.13, 모멘텀 0.4로 나타났다.

하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건을 적용한 모형에 대하여 100번 반복 분석한 결과를 비교 분석해보았다. 실험계획법으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.1383, 평균 정확도 0.8617, 평균 AUC 0.8729, 평균 계산시간 42.64초이다. 반면, 하이브리드 탐색으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.1372, 평균 정확도 0.8628, 평균 AUC 0.8717, 평균 계산시간 47.25초로 나타났다. 하이브리드 탐색을 통하여 평균 계산시간은 느리지만 그림 3.2.4에서 하이브리드 탐색으로 찾은 최적 조건들로 모형 성능의 개선 효과가 조금 더 있음을 확인할 수 있다.

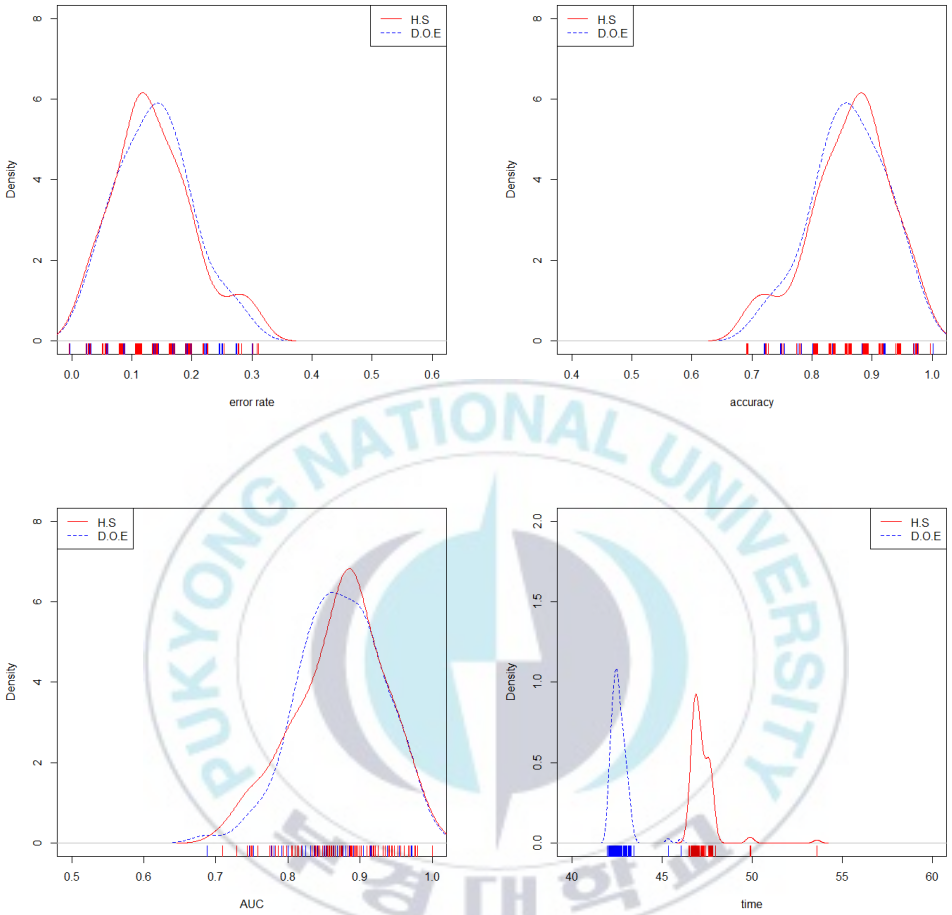

그림 3.2.7. (전립선 암 자료) 하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건 결과 비교(DNN)

표 3.2.5. (대장암 자료) 2^5 완전 요인 설계(DNN)

요인	낮은 수준(−)	높은 수준(+)
A(첫 번째 은닉층의 노드 수)	15	30
B(두 번째 은닉층의 노드 수)	15	30
C(활성화 함수)	relu	sigmoid
D(학습률)	0.01	0.1
E(모멘텀)	0.5	0.9

대장암 자료에 대하여 앞의 예제에서와 같이 은닉층의 수가 2개일 때 5개의 초매개변수인 첫 번째 은닉층의 노드 수, 두 번째 은닉층의 노드 수, 활성화 함수, 학습률, 모멘텀을 고려하였으며 표 3.2.5와 같이 2^5 완전 요인 설계를 하였다. 분산분석 결과, R^2=0.9895이며 그림 3.2.8의 표준화된 효과의 1/2정규 확률도와 Pareto 차트로부터 학습률(p=0.028336)과 모멘텀(p=0.028336)이 유의한 변수로 판단된다. 주효과도에서 첫 번째 은닉층의 노드 수 15, 두 번째 은닉층의 노드 수 15, 활성화 함수 relu, 학습률 0.01, 모멘텀 0.5일 때 낮은 오류율을 갖는 것으로 나타났다. 이 결과를 바탕으로 표 3.2.6과 같이 각 초매개변수 별 격자 탐색 범위에 따라 분석을 하였다.

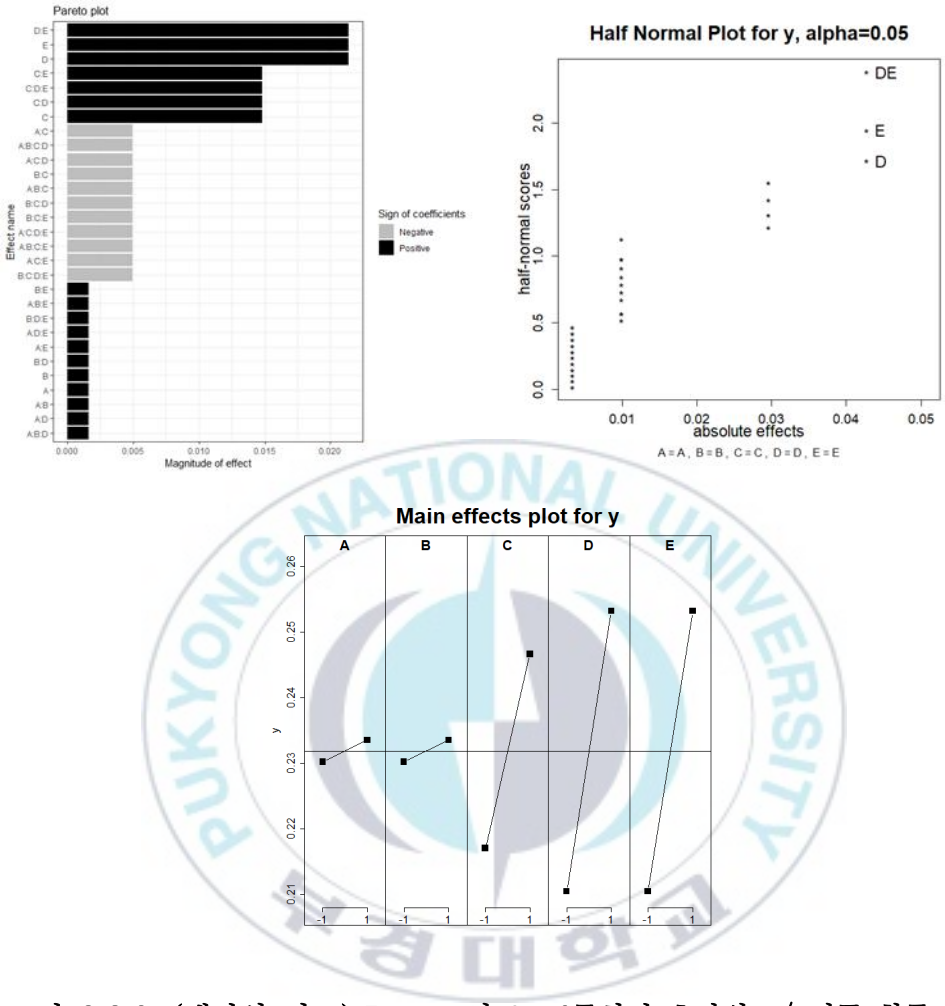

그림 3.2.8. (대장암 자료) Pareto 차트, 표준화된 효과의 1/2정규 확률도, 주효과도(DNN)

표 3.2.6. (대장암 자료) 격자 탐색으로 살펴볼 최적 조건 범위(DNN)

요인	실험계획법 이용한 최적 조건	격자 탐색으로 살펴볼 최적 조건 범위
A(첫 번째 은닉층의 노드 수)	15	(5, 20), 5씩 증가
B(두 번째 은닉층의 노드 수)	15	(5, 20), 5씩 증가
C(활성화 함수)	relu	
D(학습률)	0.01	(0.01, 0.05), 0.01씩 증가
E(모멘텀)	0.5	(0.7, 0.9), 0.1씩 증가

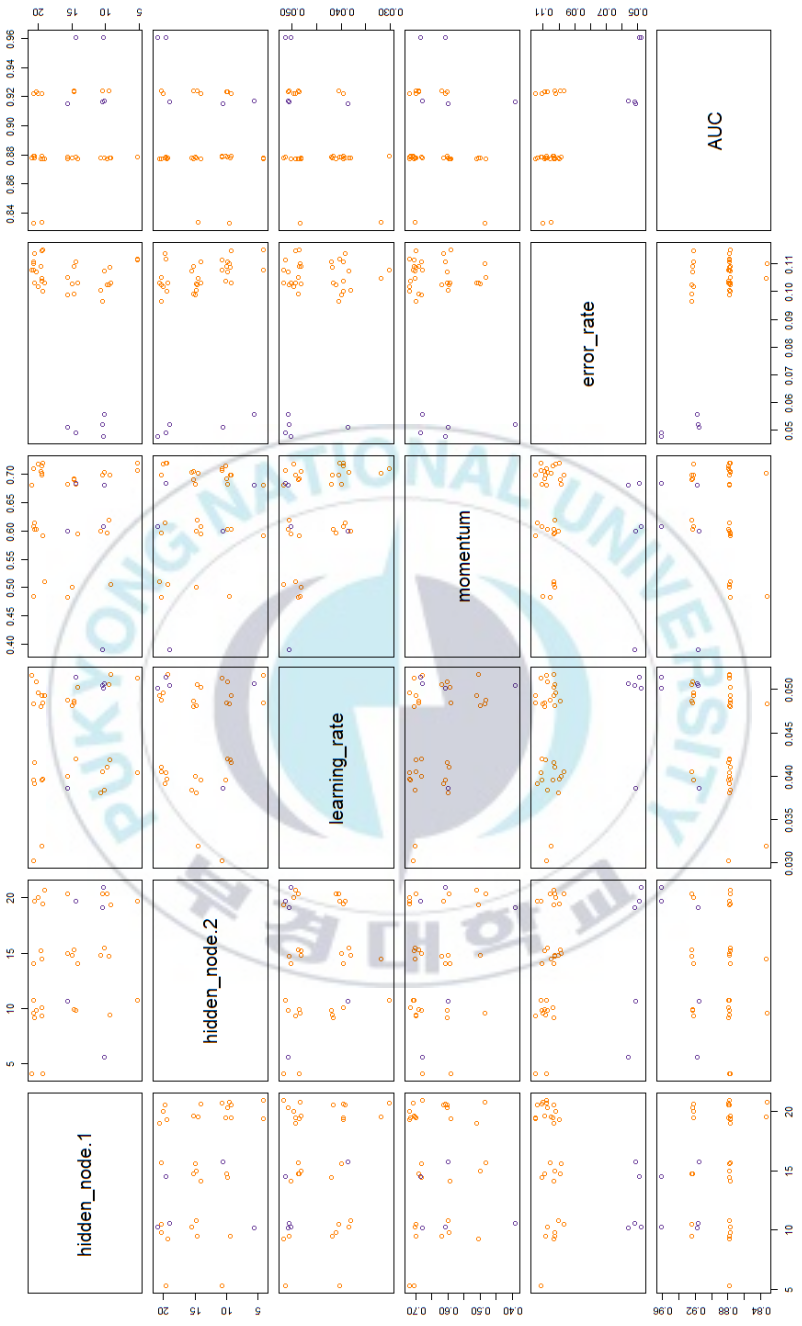

그림 3.2.9. (대장암 자료) 격자 탐색 결과에 대한 산점도 행렬(DNN)

격자탐색 결과 오류율의 범위는 0.0526~0.3158이며 0.0526 이하의 값들에 대하여 그림 3.2.9과 같이 산점도 행렬을 그린 후 분석을 하였다. 그 결과 하이브리드 탐색을 이용한 최적 조건은 첫 번째 은닉층의 노드 수 10, 두 번째 은닉층의 노드 수 20, 학습률 0.05, 모멘텀 0.7로 나타났다.

하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건을 적용한 모형에 대하여 100번 반복 분석한 결과를 비교 분석을 하였다. 실험계획법으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.2163, 평균 정확도 0.7837, 평균 AUC 0.7711, 평균 계산시간 0.3806초이다. 반면, 하이브리드 탐색으로 찾은 최적 조건들을 모형에 적용하여 분석한 결과, 평균 오류율 0.2111, 평균 정확도 0.7889, 평균 AUC 0.7754, 평균 계산시간 0.3707초로 나타났다. 그림 3.2.10에서 하이브리드 탐색으로 찾은 최적 조건들로 모형 성능의 개선 효과가 조금 더 있음을 확인할 수 있었다.

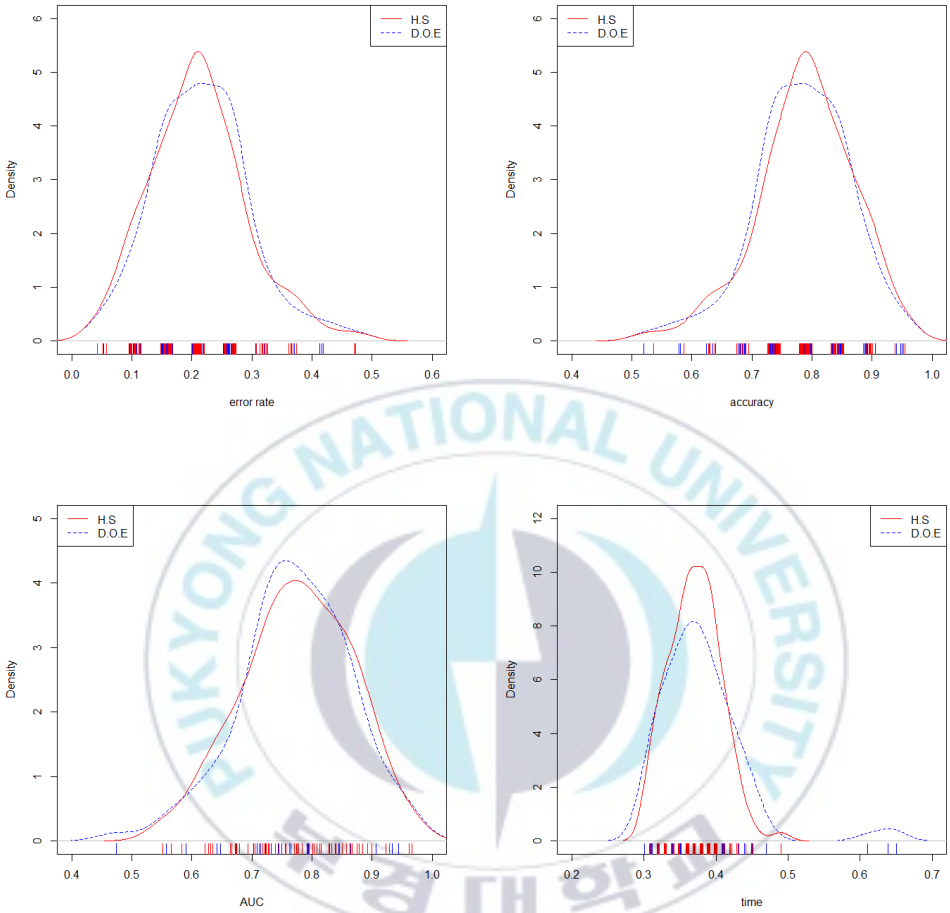

그림 3.2.10. (대장암 자료) 하이브리드 탐색으로 찾은 최적 조건과 실험계획법으로 찾은 최적 조건 결과 비교(DNN)

마지막으로 전통적 통계기법인 로지스틱 회귀분석과 신경망의 결과를 비교분석 하였다. 이때 은닉층의 수 1, 노드 수 1~30일 때의 신경망 오류율 결과와 비교하였다. 고차원 자료는 관측값의 수보다 입력 변수의 수가 더 많다는 특징 때문에 로지스틱 회귀분석이 실행되지 않는다. 따라서 본 연구에서는 Fan과 Lv(2008)이 제안한 방법을 이용하여 중요 입력 변수를 선택한 후 분석을 하였다. Fan과 Lv(2008)은 고차원 자료에서는 성김현상(sparsity)으로 인하여 서로 관련이 없는 확률변수들끼리 높은 상관관계를 가지는 위 상관관계를 일으켜서 잘못된 입력 변수를 선택할 수 있음을 언급하였다. 이러한 문제를 해결하기 위하여 고차원 자료에서 입력 변수를 선택하는 방법으로 SIS를 제안하였다. 이때 사용되는 패널티(penalty)는 LASSO, SCAD, MC+이다. 본 연구에서는 패널티가 SCAD인 SIS를 이용하여 중요 입력 변수를 선택하였으며, 총 100번을 반복 분석하였다.

백혈병 자료에서 SIS를 이용하여 선택된 입력 변수는 V3320, V4847이다. 그림 3.3.11에서 왼쪽은 로지스틱 회귀분석 결과이며, 오류율의 평균은 0.0788, 중앙값은 0.0938이다. 오른쪽 그림은 신경망을 이용하여 분석한 결과이며, 노드 수가 1일 때 오류율의 평균은 0.0668, 중앙값은 0.062이고 노드 수가 20일 때 오류율의 평균은 0.0582, 중앙값은 0.062이다. 그림 3.3.12는 전립선 암 자료 분석결과이며, SIS에 의해 입력변수 V4231, V6185, V8965가 선택되었다. 로지스틱 회귀분석 결과, 오류율의 평균은

0.1481, 중앙값은 0.1389이다. 반면, 은닉층이 1일 때 신경망을 이용하여 분석한 결과, 노드 수가 10일 때 오류율의 평균은 0.0558, 중앙값은 0.056이다. 따라서 두 고차원 자료 모두 신경망으로 분석하였을 때의 오류율이 더 낮게 나타났다.

본 연구에서는 최적의 초매개변수를 탐색하는 방법으로써 실험계획법과 격자탐색을 조합한 하이브리드 탐색을 제안하였으며, 심층 신경망에서 하이브리드 탐색을 이용하여 최적의 초매개변수를 탐색한 후 고차원 예제들을 분석해보았다. 본 연구에서는 하이브리드 탐색 결과와 실험계획법의 결과를 비교 분석하였지만 추후에는 임의탐색, 자동 선택 방법, 베이지안 최적화 등의 결과도 함께 비교분석을 하고자 한다.

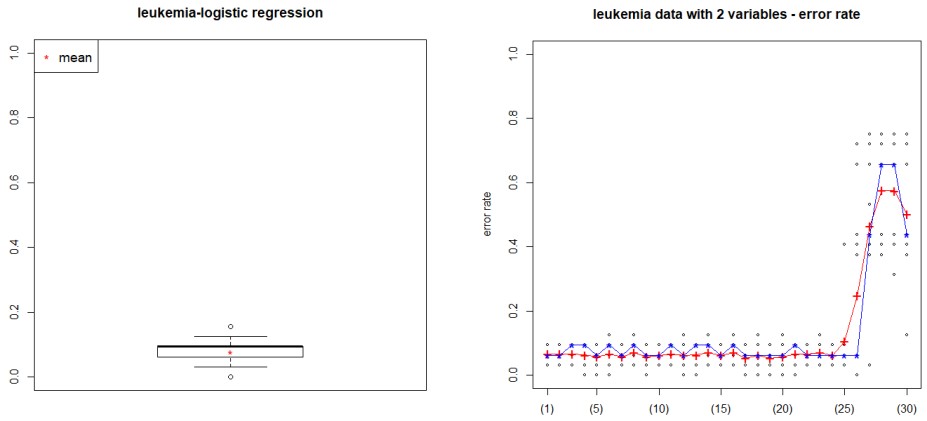

그림 3.3.11. (백혈병 자료) 로지스틱 회귀분석과 심층 신경망 결과

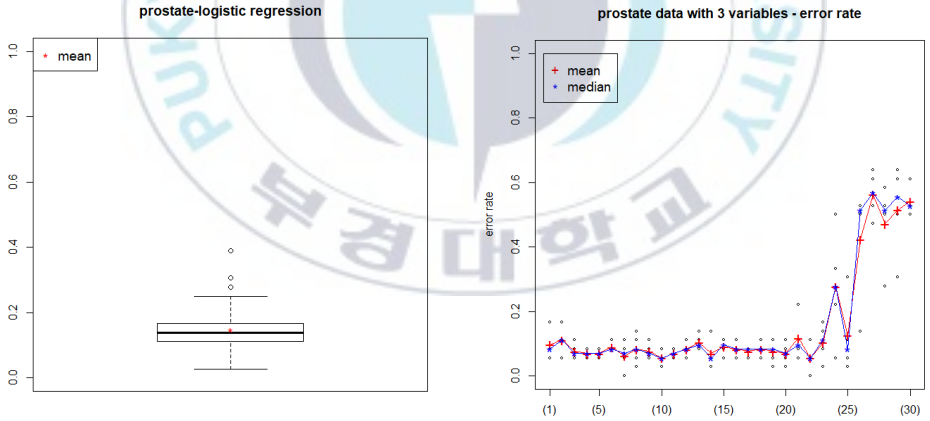

그림 3.3.12. (전립선 암 자료) 로지스틱 회귀분석과 심층 신경망 결과

Ⅳ. 고차원 자료에 대한 심층 신경망 최적의 은닉층 수와 노드 수 탐색

심층 신경망에서 초매개변수인 은닉층의 수와 노드 수의 선택은 출력 값과 계산시간에 큰 영향을 준다. 하나의 예로써 Lee와 Jang(2018b)은 오토인코더(sparse autoencoder)를 이용하여 R 로고 그림을 구현하는 실험을 하였다. R 로고 이미지는 회색 스케일로서 크기는 77×101를 대상으로 하였고 오토인코더 구성은 은닉층의 수 1개, 은닉층 노드 수는 1~100개까지 실험을 하였으며, 전체 계산과정을 20번 반복하였다. 그림 4.1을 보면, 은닉층의 노드 수가 대략 60일 때 평균제곱오차(MSE)와 계산시간의 변화시점이 발생한다. 이 예를 통하여 신경망에서는 은닉층의 수와 노드 수가 출력 값과 계산시간에 큰 영향을 주는 매우 중요한 초매개변수라는 것을 알 수 있었다. 본 절에서는 고차원 자료에서 최적의 은닉층 수와 노드 수를 확인하였다. 고차원 자료에서는 과적합 현상을 초래할 수 있으므로 모든 입력변수를 사용하는 경우와 선택된 입력변수를 사용하는 경우로 나누어서 살펴보았으며, 두 결과를 서로 비교 분석하였다. 이때 100번씩 반복 작업한 결과에 대하여 분석을 하였다.

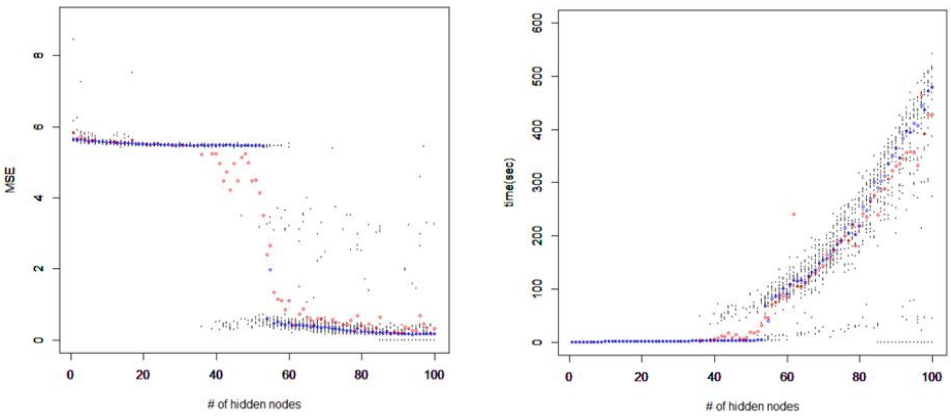

그림 4.1. 은닉층 수에 따른 평균제곱오차(MSE)와 계산시간 (단위: 초)

4.1 모든 입력 변수 사용의 경우 심층 신경망 최적의 은닉층 수와 노드 수 탐색

첫 번째 예제는 백혈병 자료이며, 은닉층이 1개일 때 노드 수를 1~30까지 증가시키면서 오류율과 계산시간을 측정하였다. 하지만 노드 수가 21~30일 때에는 계산이 되지 않는 경우가 종종 발생하였다. 따라서 1~20까지 측정한 자료에 대하여 분석을 하였다. 그림 4.1.1을 보면, 오류율의 평균이 가장 작은 경우는 노드 수가 15일 때이다. 이때 오류율의 평균은 0.1554, 중앙값은 0.1560이다. 하지만 오류율의 범위는 0.031~0.438

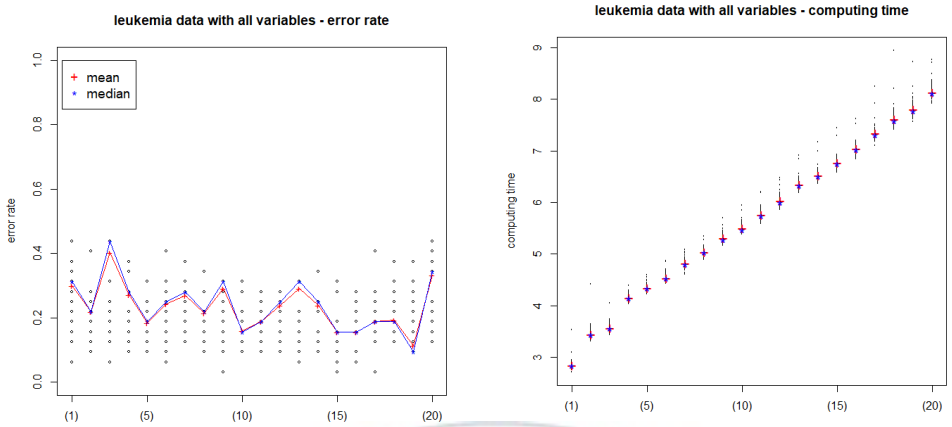

그림 4.1.1. (백혈병 자료) 은닉층의 수 1, 모든 입력 변수 사용의 경우 오류율과 계산시간 산점도

이며 오류율의 변동이 매우 크다. 계산시간의 범위는 2.71~8.95초이며, 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보인다.

은닉층의 수가 2개일 때, 첫 번째, 두 번째 은닉층의 노드 수를 각각 1~30까지 설정하여 실험하였다. 그림 4.1.2의 오류율 산점도를 보면, 첫 번째 은닉층의 수가 변화함에 따라 패턴이 반복되고 있다. 또한, 계산시간 산점도를 통하여 노드 수가 증가함에 따라 계산시간도 증가하며, 첫 번째 은닉층의 노드 수가 바뀔 때마다 계산시간의 폭이 바뀌는 것을 알 수 있다. 계산시간 범위는 2.54~15.42초이다.

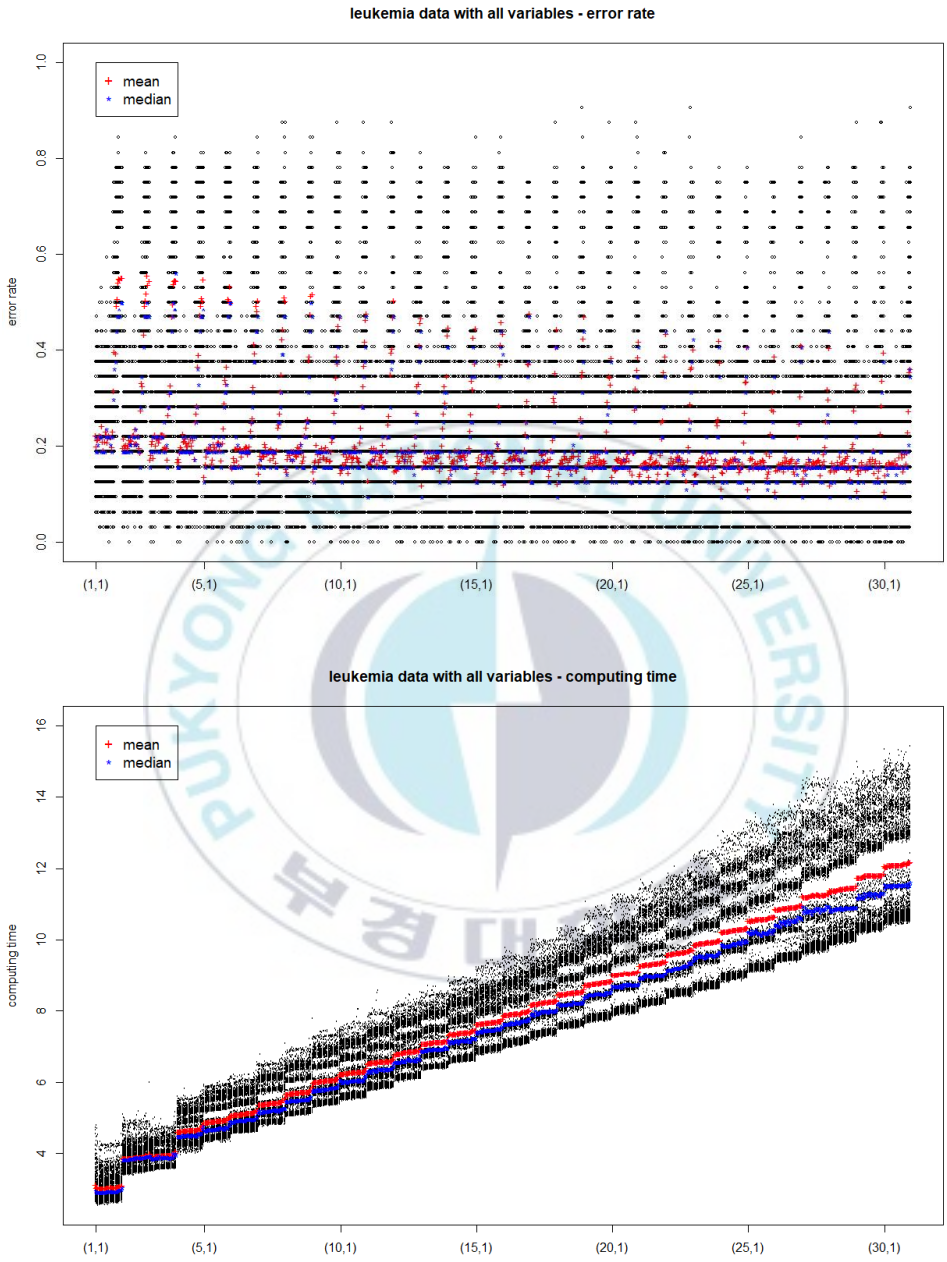

그림 4.1.2. (백혈병 자료) 은닉층의 수 2, 모든 입력 변수 사용의 경우 오류율과 계산시간 산점도

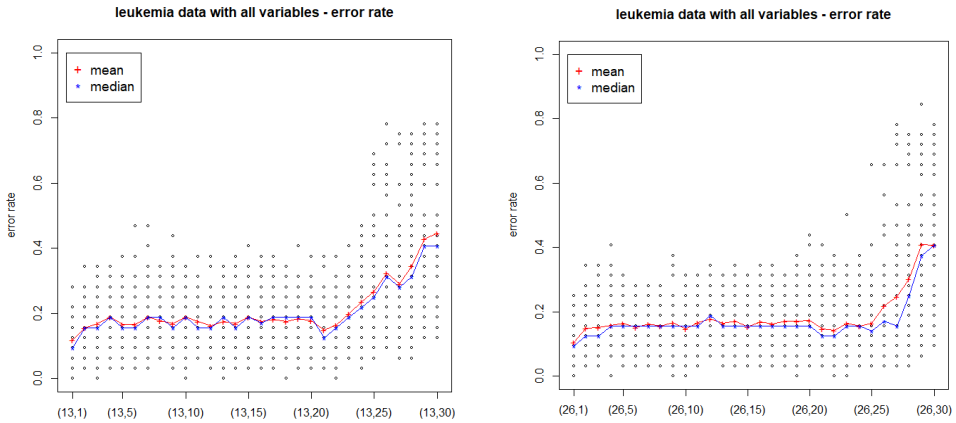

그림 4.1.3. (백혈병 자료) 은닉층의 수 2, 모든 입력 변수 사용의 경우 (13,1)~(13,30)과 (26,1)~(26,30)의 오류율 산점도

오류율의 패턴을 자세히 파악하기 위하여 오류율의 평균 및 중앙값이 작은 일부 은닉층의 노드인 (13,1)~(13,30)과 (26,1)~(26,30)을 분석하였다. 그림 4.1.3을 보면, (13,1)~(13,20)일 때 오류율의 평균 범위는 0.11872~0.18871, 중앙값은 0.094~0.188이며 (13, 1)일 때 오류율의 평균이 및 중앙값이 가장 작았다. (26,1)~(26,25)일 때 오류율의 평균 범위는 0.12499~0.18814, 중앙값은 0.125~0.188이며 (26, 1)일 때 오류율의 평균이 및 중앙값이 가장 작았다.

은닉층의 수가 3개일 때에도 첫 번째, 두 번째, 세 번째 은닉층의 수를 각각 1~30까지 설정하여 실험하였다. 그림 4.1.4의 (1,1,1)~(30,30,30) 오류율 산점도를 보면, 첫 번째 은닉층의 노드 수에 따른 오류율의 변동이

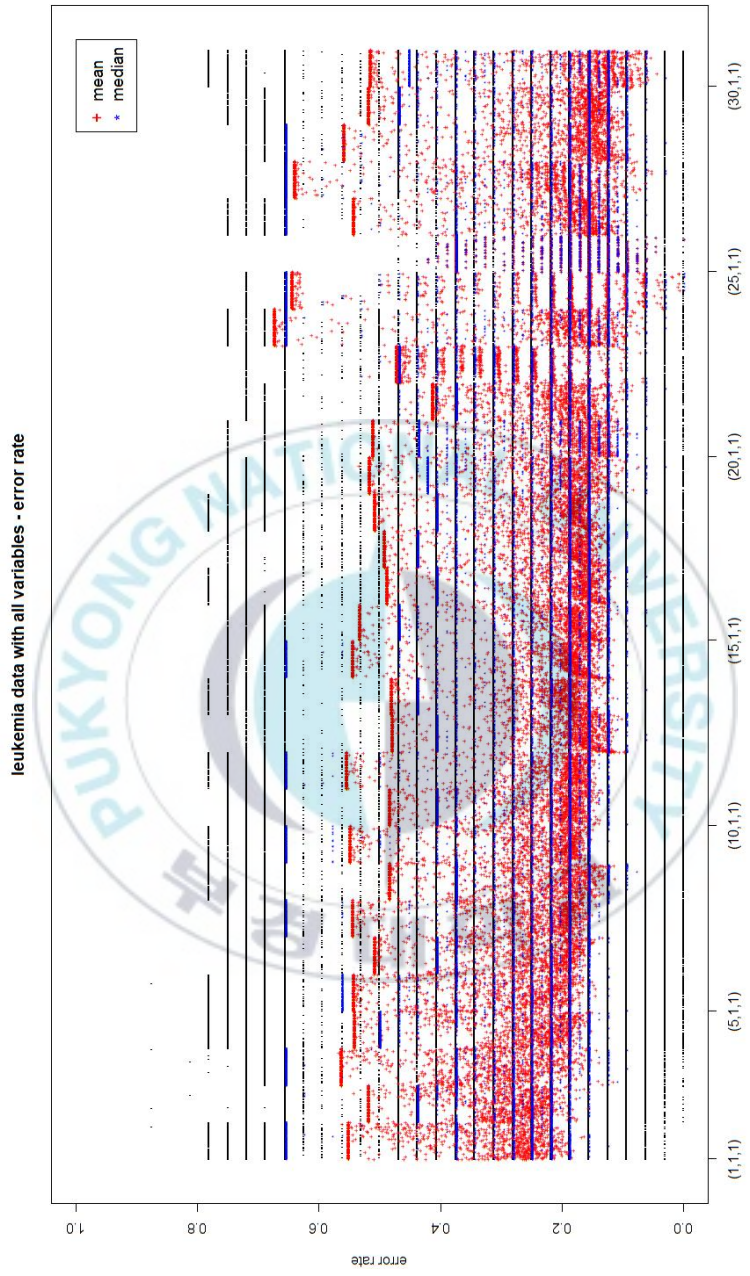

그림 4.1.4. (백혈병 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (1,1,1)~(30,30,30)의 오류율 산점도

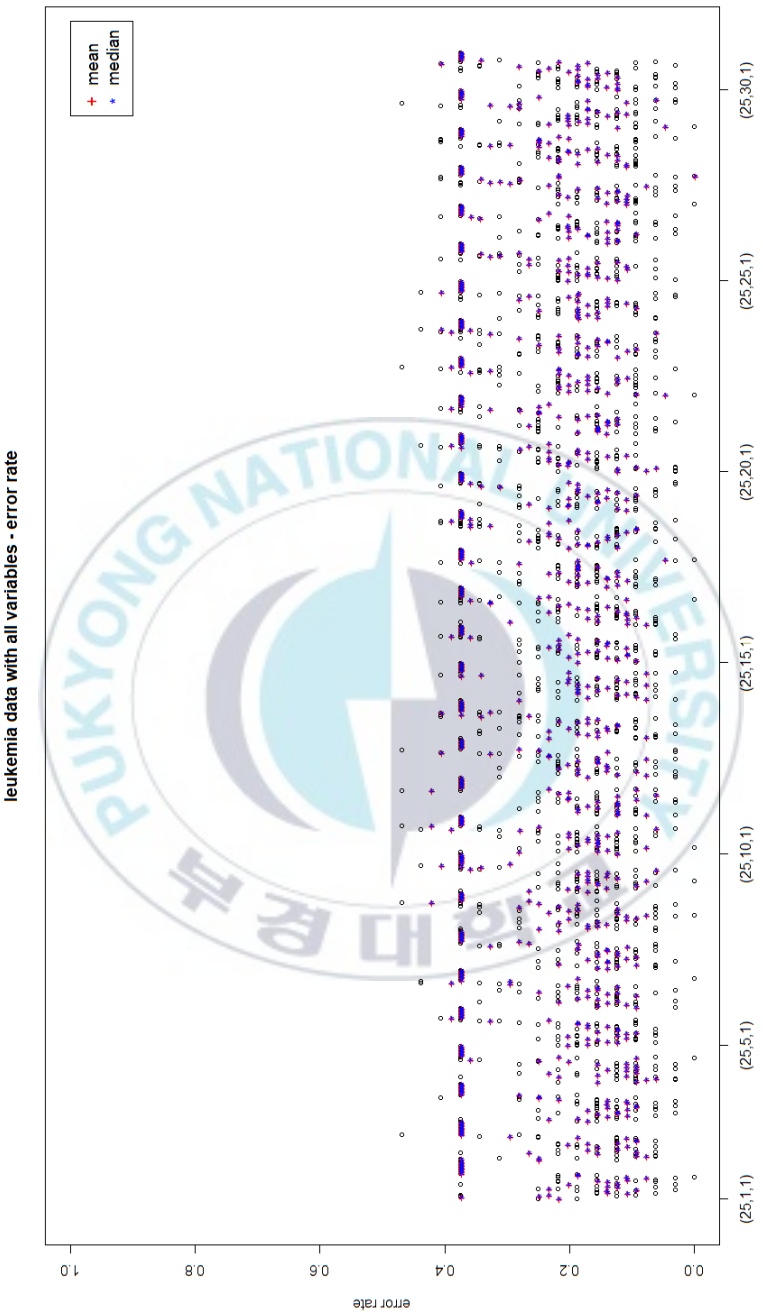

그림 4.1.5. (백혈병 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (25,1,1)~(25,30,30)의 오류율 산점도

매우 심하다는 것을 알 수 있다. 그림 4.1.5의 (25,1,1)~(25,30,30) 오류율 산점도를 보면, 전반적으로 두 번째 은닉층의 노드 수 변화에 따른 패턴이 반복되는 것을 확인할 수 있다. 오류율의 패턴을 좀 더 자세히 파악하기 위하여 그림 4.1.6과 같이 일부 은닉층의 노드를 분석하였다. (25,16,1)~(25,16,10)일 때 오류율의 평균값 및 중앙값의 범위는 모두 0.078~0.297이다. (25,25,1)~(25,25,12)일 때 오류율의 평균값 및 중앙값 범위는 모두 0.109~0.219이다. 두 경우의 모두 세 번째 은닉층의 노드수가 커질수록 오류율도 커진다는 것을 알 수 있다.

두 번째 예제는 전립선 암 자료이며, 은닉층이 1개일 때 노드 수를 1~30까지 증가시키면서 오류율과 계산시간을 측정하였다. 하지만 노드 수가 23~30일 때 100번 반복 분석 중 일부분은 계산이 되지 않아서 1~22까지 측정한 자료를 이용하여 분석하였다.

그림 4.1.7을 보면, 오류율의 평균이 가장 작은 경우는 노드 수가 10일 때이며 이때 오류율의 평균은 0.3724, 중앙값은 0.3610이다. 하지만 오류율의 범위는 0.083~0.722이며 오류율의 변동이 매우 크다. 계산시간 범위는 15.87~44.28초이며, 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보인다.

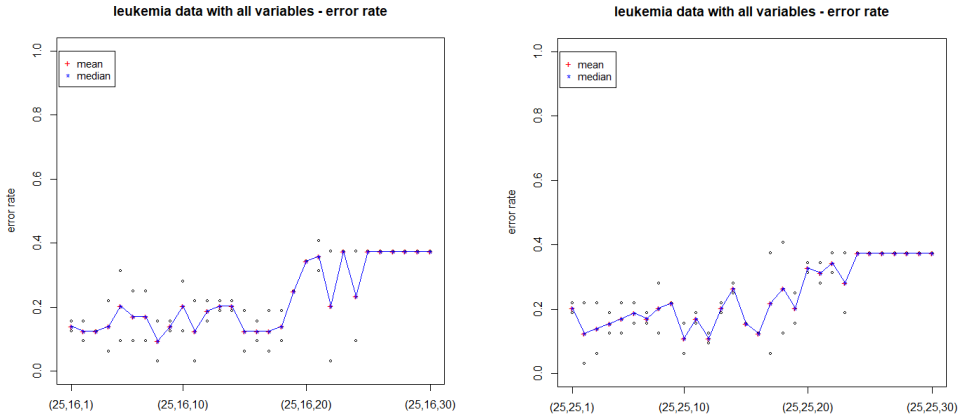

그림 4.1.6. (백혈병 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (25,16,1)~(25,16,30)과 (25,25,1)~(25,25,30)의 오류율 산점도

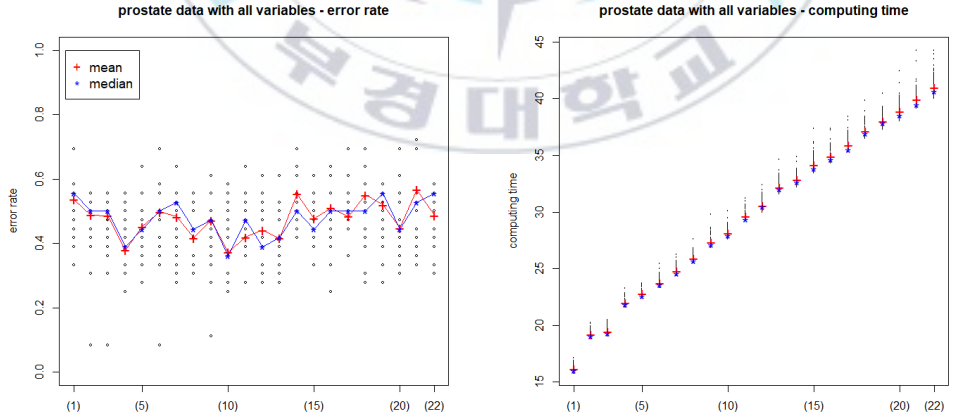

그림 4.1.7. (전립선 암 자료) 은닉층의 수 1, 모든 입력 변수 사용의 경우 오류율과 계산시간 산점도

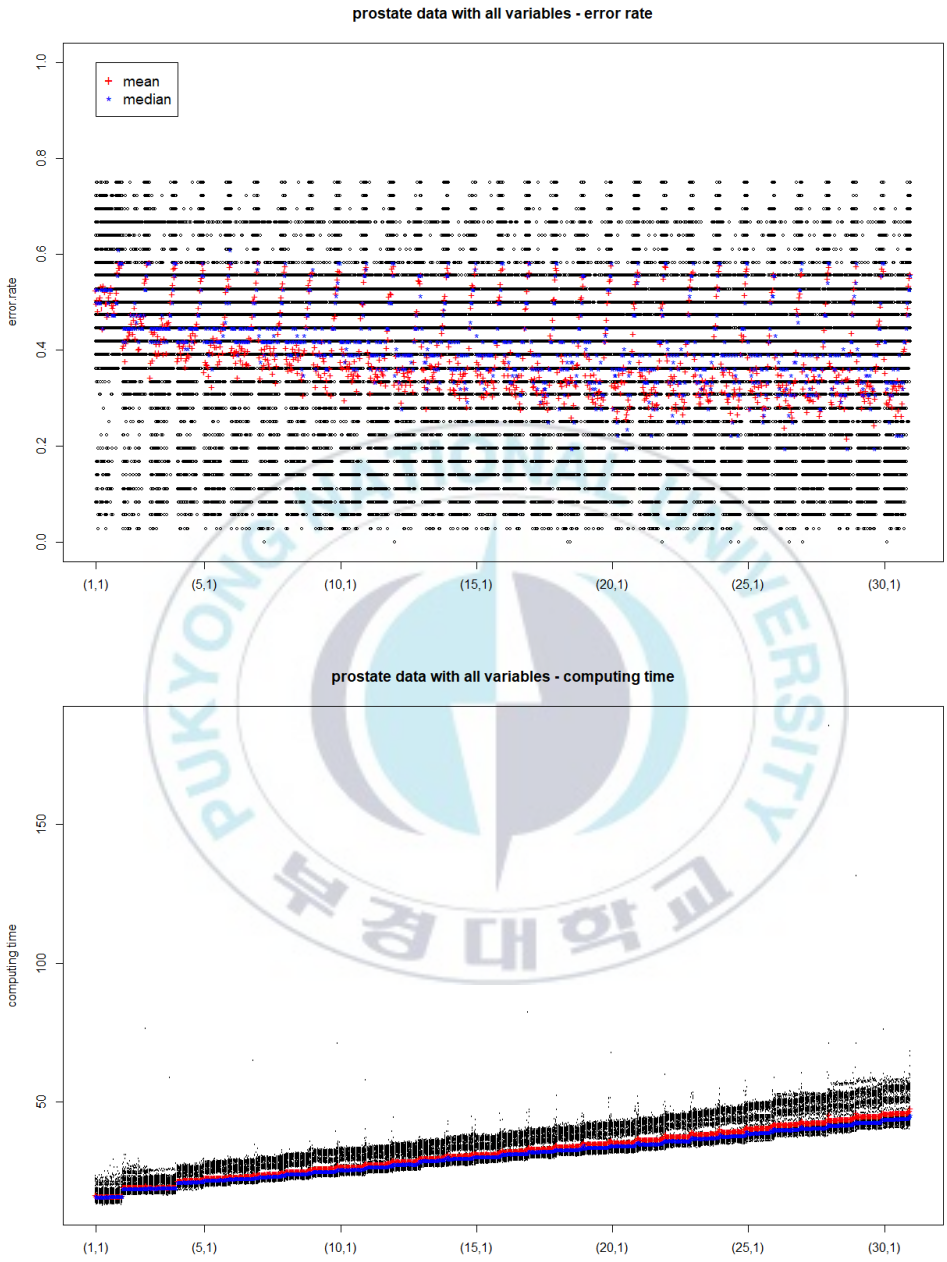

그림 4.1.8. (전립선 암 자료) 은닉층의 수 2, 모든 입력 변수 사용의 경우 오류율과 계산시간 산점도

은닉층의 수가 2개일 때, 분석 결과는 그림 4.1.8과 같다. 첫 번째와 두 번째 은닉층의 노드 수가 변화함에 따라 오류율의 평균 및 중앙값이 작아지고 있다. 하지만 오류율은 여전히 0.3 이상으로 높게 나타났다. 또한, 계산시간 산점도를 보면 전반적으로 50초 이하이지만 계산시간 범위는 12.76~185.19초이다. 따라서 노드 수가 증가함에 따라 계산시간이 매우 커진다는 것을 알 수 있다. 그림 4.1.9와 같이 일부 은닉층 노드의 오류율을 분석하였다. (9,1)~(9,20)일 때 오류율의 평균 범위는 0.3174~0.3968, 중앙값은 0.361~0.4305이며, (24,1)~(24,25)일 때 오류율의 평균 범위는 0.2551~0.4062, 중앙값은 0.194~0.417이다. 따라서 전반적으로 오류율 값 및 변동이 매우 크다는 것을 확인할 수 있다.

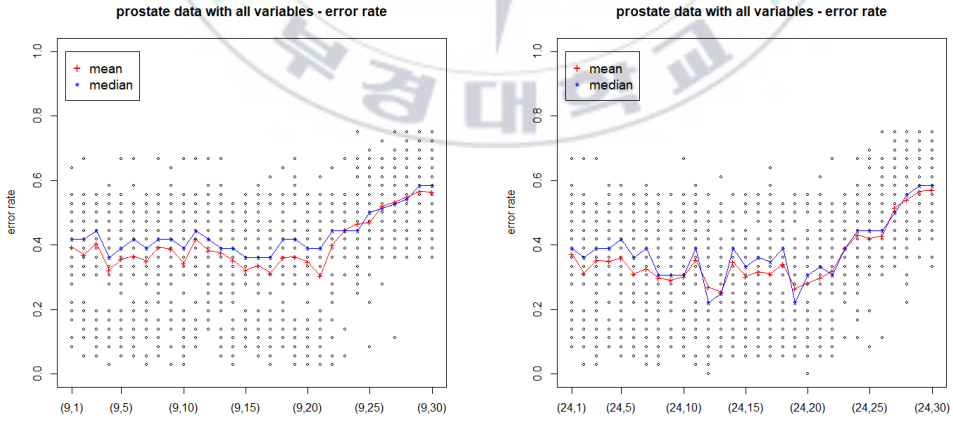

그림 4.1.9. (전립선 암 자료) 은닉층의 수 2, 모든 입력 변수 사용의 경우 (9,1)~(9,30)과 (24,1)~(24,30)의 오류율 산점도

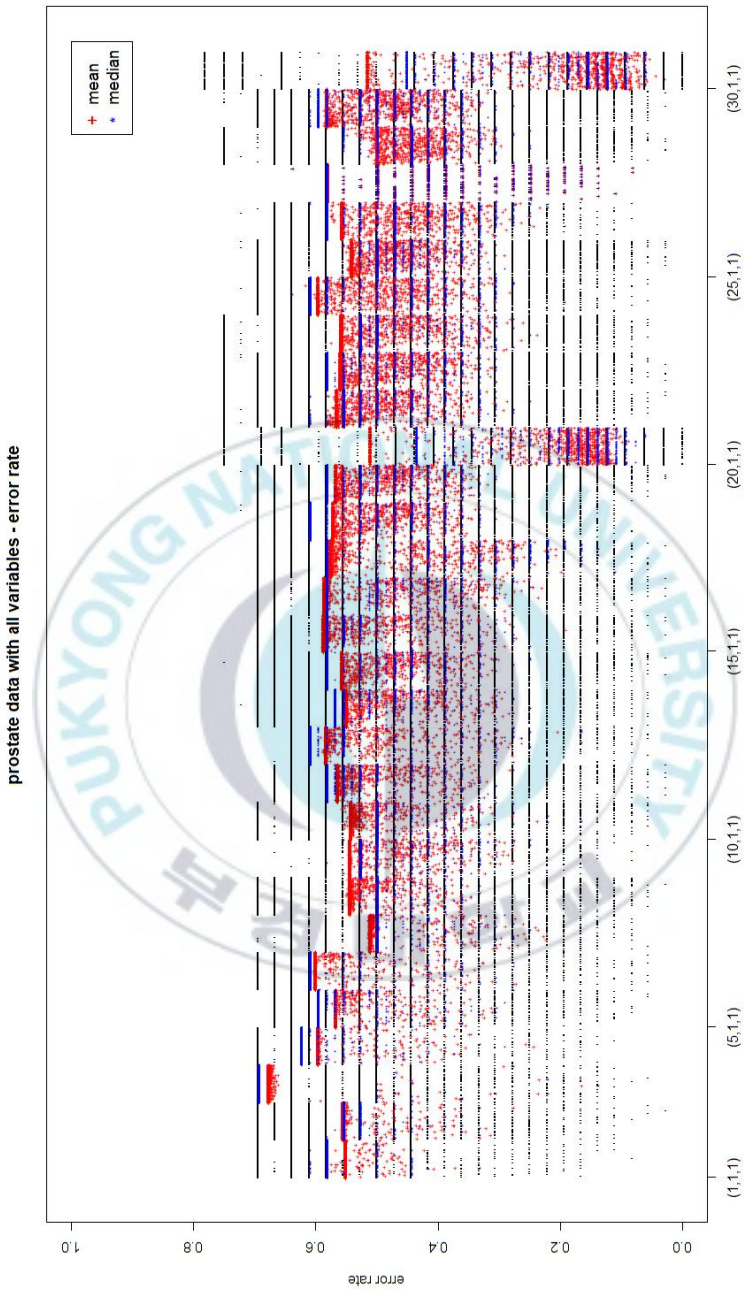

그림 4.1.10. (전립선 암 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (1,1,1)~(30,30,30)의 오류율 산점도

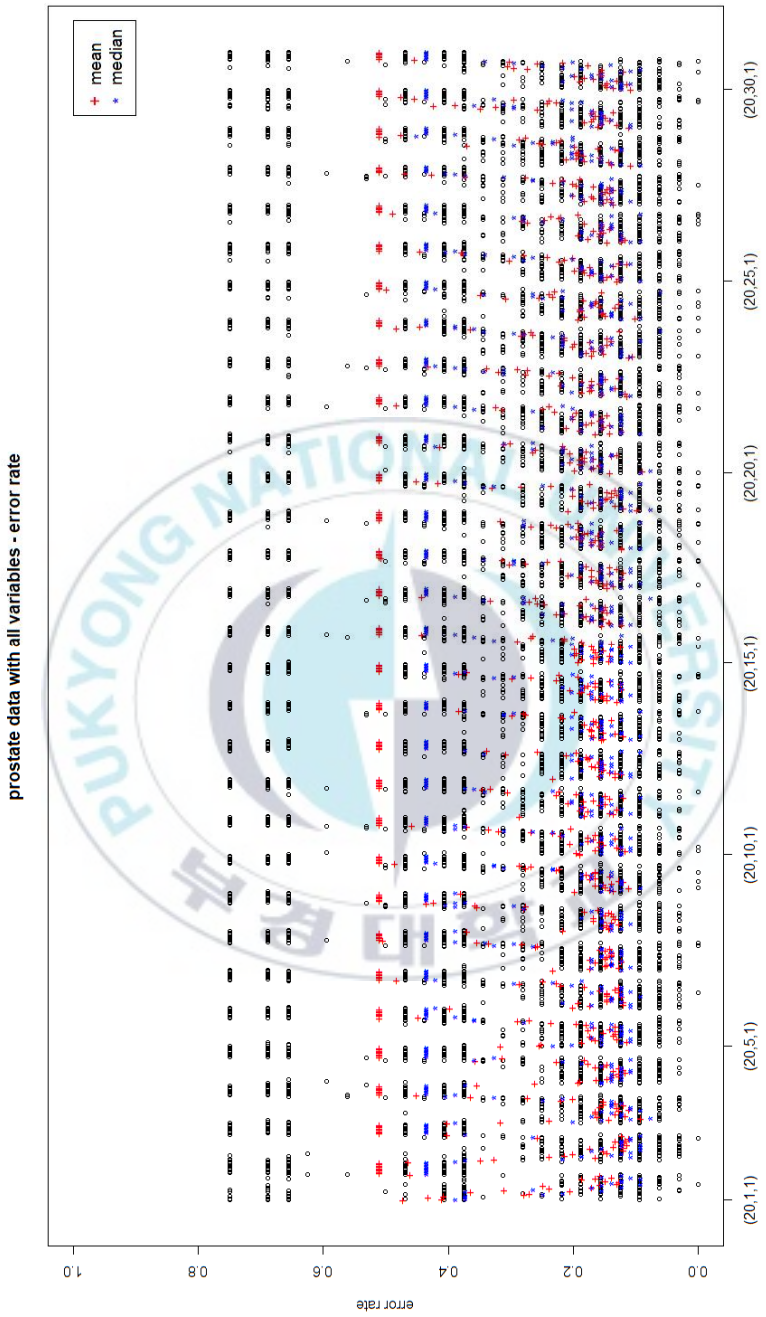

그림 4.1.11. (전립선 암 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (20,1,1)~(20,30,30)의 오류율 산점도

은닉층의 수가 3개일 때 (1,1,1)~(30,30,30) 오류율 산점도는 그림 4.1.10과 같으며, 오류율의 변동은 매우 크다. 그리고 그림 4.1.11의 (20,1,1)~(20,30,30) 오류율 산점도를 통하여 두 번째 은닉층 수 변화에 따른 패턴이 반복되는 것을 확인할 수 있다. 오류율의 패턴을 자세히 확인하기 위하여 그림 4.1.12와 같이 일부 은닉층의 노드를 확인해보았다. (20,3,1)~(20,3,20)일 때 오류율의 평균 범위는 0.0935~0.4031, 중앙값은 0.078~0.3905이며, 세 번째 노드 수가 5일 때 가장 작다. (20,19,1)~(20,19,6)일 때 오류율의 평균 범위는 0.1~0.2093, 중앙값은 0.078~0.204이며 세 번째 노드 수 2일 때 가장 작다.

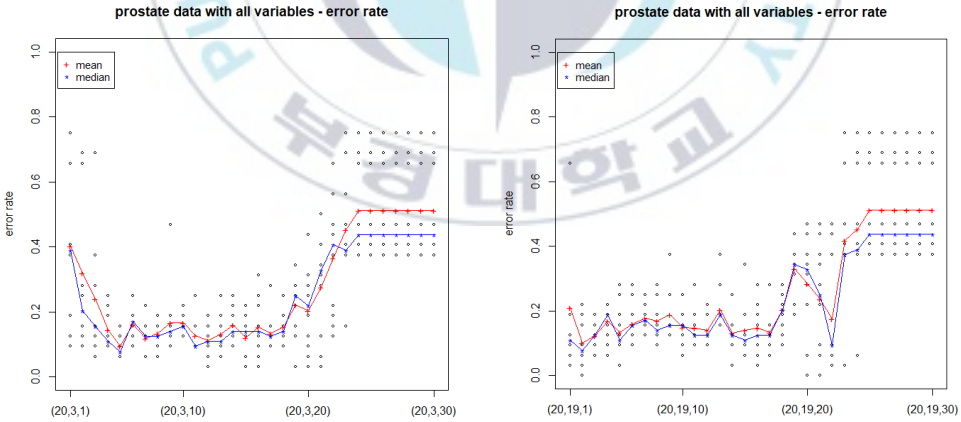

그림 4.1.12. (전립선 암 자료) 은닉층의 수 3, 모든 입력 변수 사용의 경우 (20,3,1)~(20,3,30)과 (20,19,1)~(20,19,30)의 오류율 산점도

본 절에서는 두 고차원 자료에서 모든 입력 변수를 사용했을 때 은닉층의 수가 1, 2, 3인 경우를 분석하였다. 은닉층의 수가 1개일 때 오류율의 평균 및 중앙값이 전반적으로 크고 변동도 컸다. 은닉층의 수가 2개일 때에는 첫 번째 은닉층의 수가 증가할수록 오류율의 평균 및 중앙값이 작아지는 패턴을 보였다. 그리고 은닉층의 수가 3개일 때에는 첫 번째 은닉층의 수가 바뀌면서 오류율의 변동이 크게 발생하였으며, 세 번째 은닉층의 수가 클 때 오류율도 컸다. 하지만 전반적으로 오류율의 값과 변동이 매우 크기 때문에 사용하기에는 적절하지 않은 것으로 보인다.

4.2 입력 변수 선택의 경우 심층 신경망 최적의 은닉층 수와 노드 수 탐색

본 연구에서는 3.3절에서 언급한 SIS를 이용하여 입력 변수를 선택한 후 분석을 하였다. 먼저, 백혈병 자료에 대하여 은닉층이 1개일 때 노드 수를 1~30까지 증가 하면서 오류율과 계산시간을 측정하였다. 그림 4.2.1에서 노드 수가 1~24일 때에는 오류율의 평균 범위는 0.0539~0.0712, 중앙값 범위는 0.062~0.094이다. 하지만 노드 수가 25일 때부터 오류율이 증가

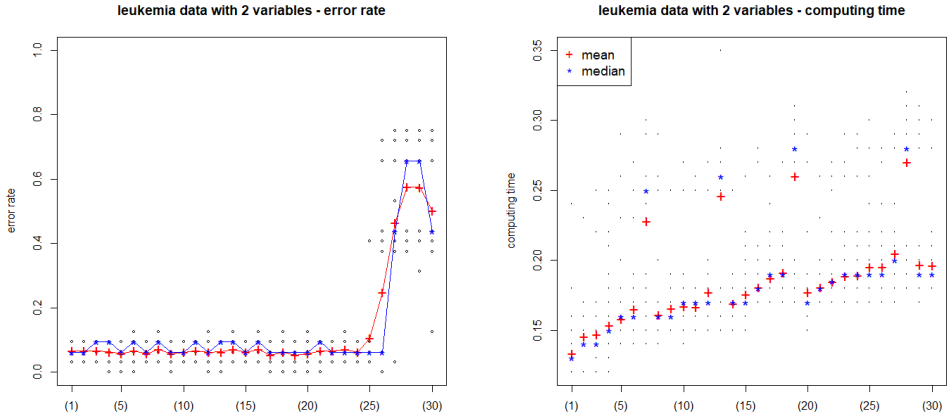

그림 4.2.1. (백혈병 자료) 은닉층의 수 1, 입력 변수 선택의 경우 오류율과 계산시간 산점도

하며 노드 수가 29일 때에는 오류율 평균이 0.5746, 중앙값이 0.6560이다. 또한, 은닉층 1일 때 계산시간 범위는 0.12~0.35초이며, 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보인다.

은닉층의 수가 2개일 때, 첫 번째, 두 번째 은닉층의 노드 수를 각각 1~30까지 설정하여 실험하였다. 그림 4.2.2에서 오류율 산점도를 보면, 첫 번째 은닉층의 수가 변화함에 따라 패턴이 반복되고 있다. 또한, 계산시간 범위는 0.14~0.55초이며, 계산시간 산점도를 보면 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보인다. 오류율의 패턴을 자세히 파악하기 위하여 일부 은닉층의 노드를 분석하였다. 그림 4.2.3과 같이

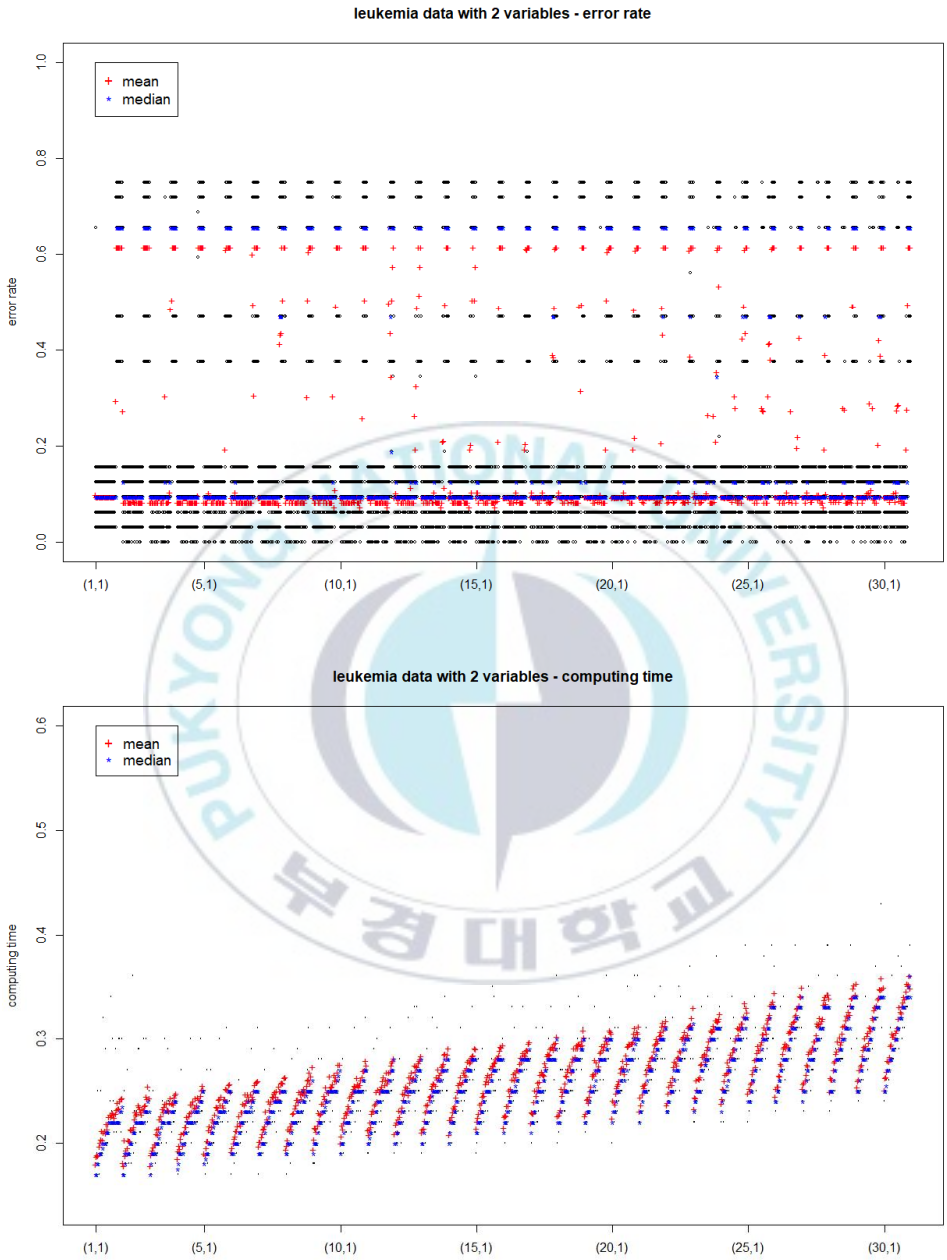

그림 4.2.2. (백혈병 자료) 은닉층의 수 2, 입력 변수 선택의 경우 오류율과 계산시간 산점도

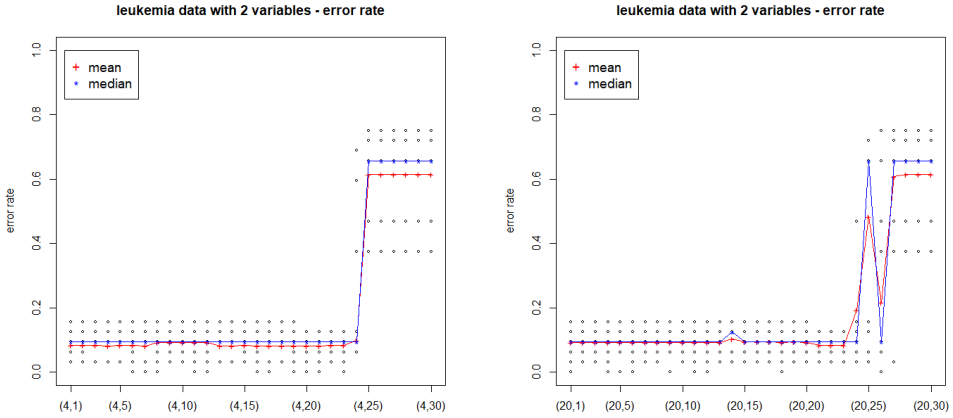

그림 4.2.3. (백혈병 자료) 은닉층의 수 2, 입력 변수 선택의 경우 (4,1)~(4,30)과 (20,1)~(20,30)의 오류율 산점도

(4,1)~(4,30)과 (20,1)~(20,30)를 분석한 결과, 은닉층의 수가 1일 때보다 변동이 조금 더 크지만 비슷한 패턴을 보인다. (4,1)~(4,23)일 때 오류율의 평균 범위는 0.08218~0.09304, 중앙값은 모두 0.094이며, (20,1)~(20,23)일 때 오류율의 평균 범위는 0.0828~0.10358, 중앙값은 0.094~0.125이다.

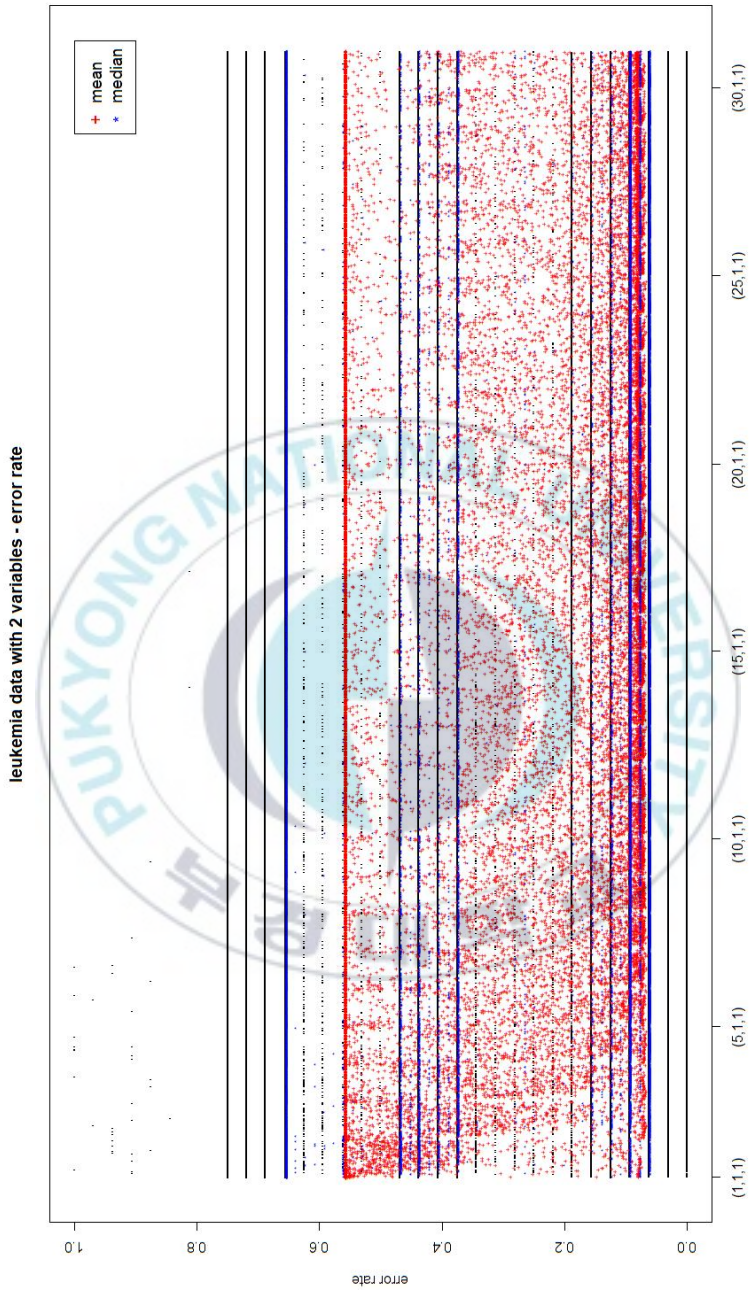

그림 4.2.4. (백혈병 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (1,1,1)~(30,30,30)의 오류율 산점도

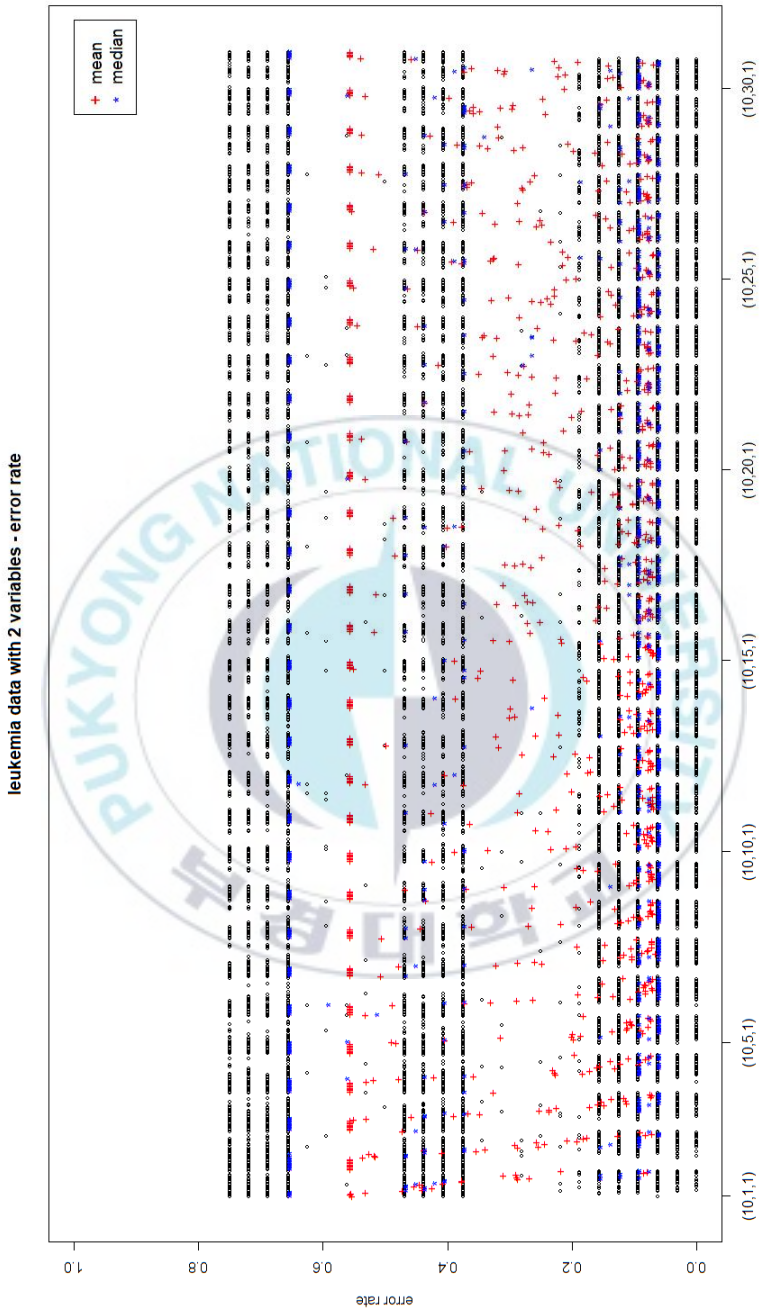

그림 4.2.5. (백혈병 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (10,1,1)~(10,30,30)의 오류율 산점도

은닉층의 수가 3개일 때 첫 번째, 두 번째, 세 번째 은닉층의 노드 수를 각각 1~30까지 설정하여 실험하였다. 그림 4.2.4와 같이 첫 번째 은닉층의 노드 수가 1~5일 때에는 오류율이 상대적으로 크지만 6 이상일 때에는 오류율의 변화가 뚜렷하게 보이지 않았다. 패턴을 자세히 보기 위하여 그림 4.2.5와 같이 (10,1,1)~(10,30,30)의 오류율을 분석한 결과, 두 번째 은닉층의 노드 수에 따른 패턴이 반복되는 것을 확인할 수 있다.

오류율의 패턴을 확인하기 위하여 그림 4.2.6과 같이 일부 은닉층의 노드를 분석한 결과, (10,5,1)~(10,5,12)일 때 오류율의 평균 범위는 0.0752~0.4925, 중앙값은 0.062~0.656이며. 오류율의 평균 및 중앙값이 가장 작을 때에는 노드 수가 각각 19, 9일 때이다. (10,9,1)~(10,9,12)일 때 오류율의 평균 범위는 0.0729~0.4703, 중앙값은 0.062~0.438이며. 오류율의 평균 및 중앙값이 가장 작을 때에는 노드 수가 각각 10, 5일 때이다.

다음으로 전립선 암 자료에 하여 은닉층이 1개일 때 노드 수를 1~30까지 증가시키면서 오류율과 계산시간을 측정하였다. 그림 4.2.7을 보면, 노드 수가 1~23일 때에는 오류율의 평균 범위는 0.05575~0.118, 중앙값 범위는 0.056~0.111이다. 하지만 노드 수가 24일 때부터 변동이 커지면서 노드 수가 27일 때에는 오류율의 평균이 0.5625, 중앙값이 0.5695로 나타났다. 또한, 계산시간 범위는 0.28~0.60초이며, 첫 번째 노드일 때를

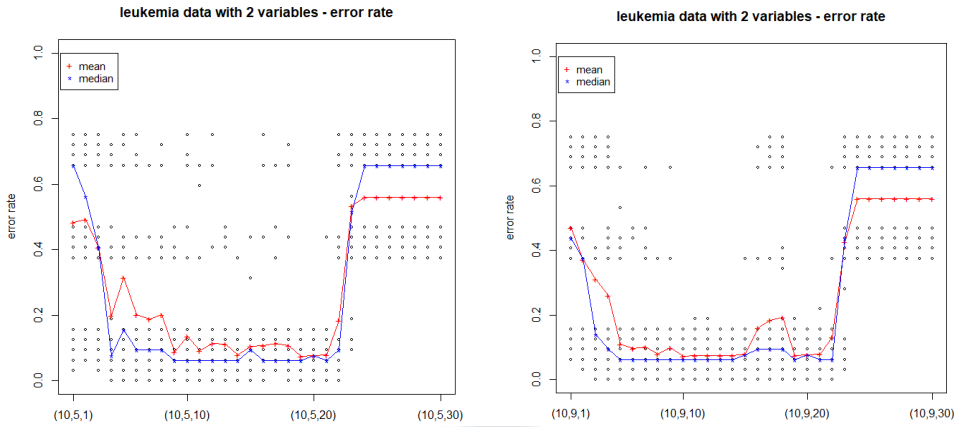

그림 4.2.6. (백혈병 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (10,5,1)~(10,5,30)과 (10,9,1)~(10,9,30)의 오류율 산점도

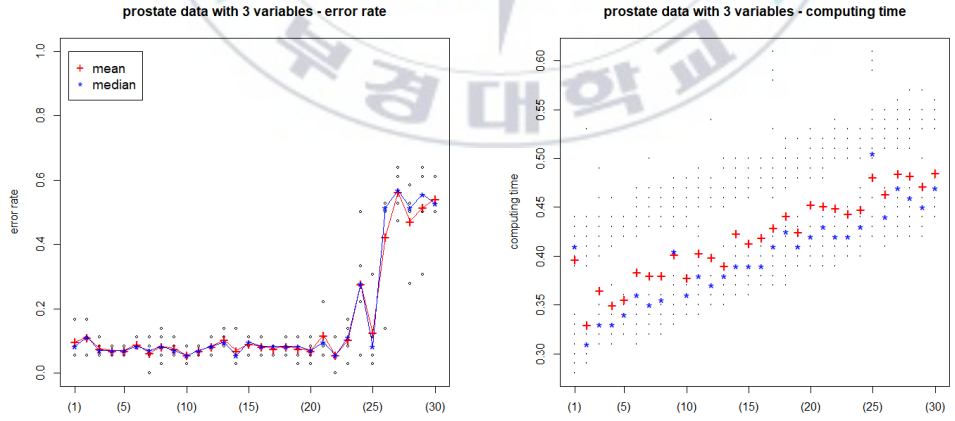

그림 4.2.7. (전립선 암 자료) 은닉층의 수 1, 입력 변수 선택의 경우 오류율과 계산시간 산점도

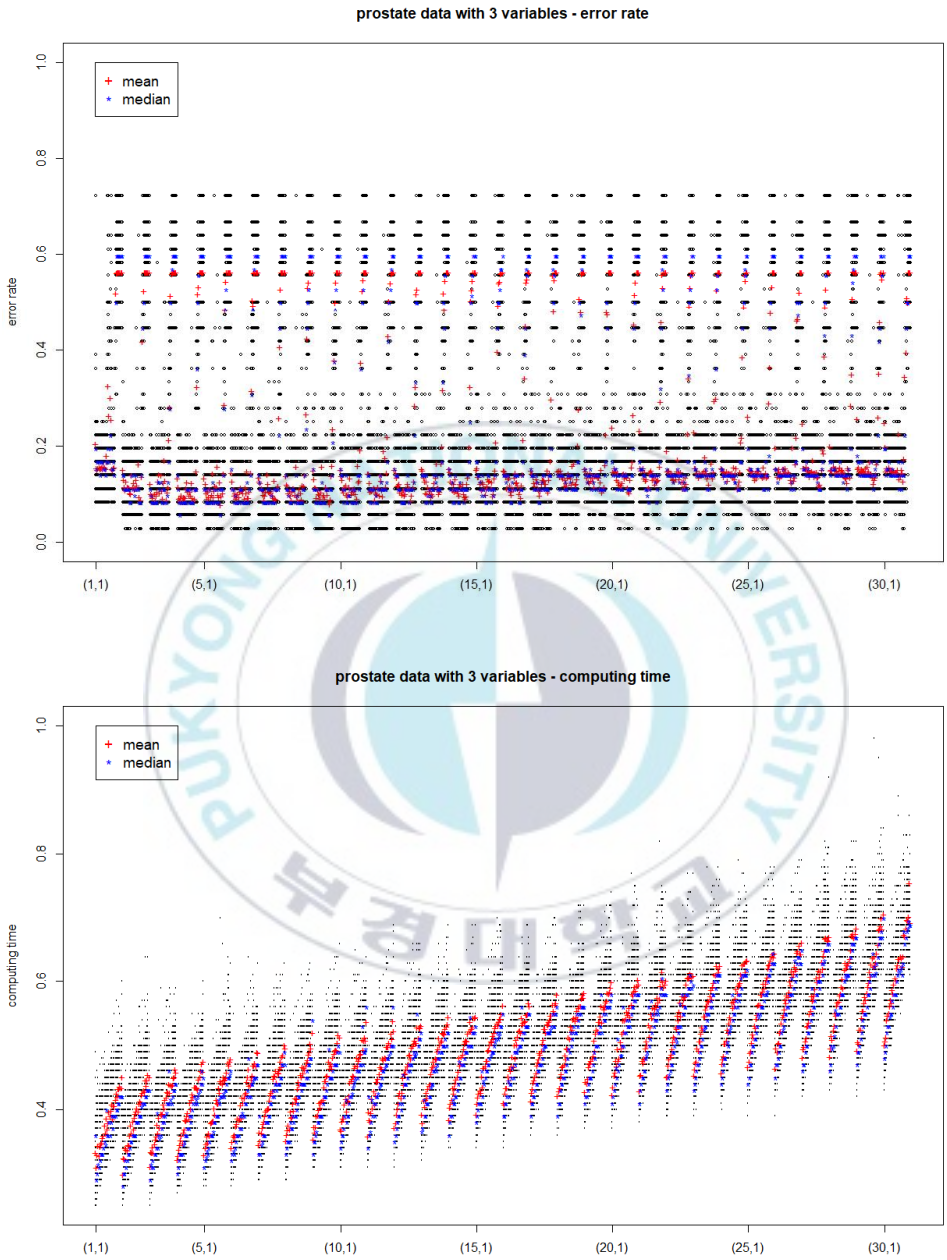

그림 4.2.8. (전립선 암 자료) 은닉층의 수 2, 입력 변수 선택의 경우 오류율과 계산시간 산점도

제외하고는 전반적으로 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보인다.

은닉층의 수가 2개일 때에는 그림 4.2.8과 같이 첫 번째 은닉층의 수가 변화함에 따라 반복되는 패턴을 보인다. 첫 번째 은닉층의 수가 1일 때 오류율이 가장 크며, 첫 번째 은닉층의 수가 증가할수록 오류율이 조금씩 작아지다 점점 커지는 패턴을 보인다. 또한, 계산시간 산점도를 보면 노드 수가 증가함에 따라 계산시간도 증가하는 패턴을 보이며, 계산시간 범위는 0.25~6.89초이다.

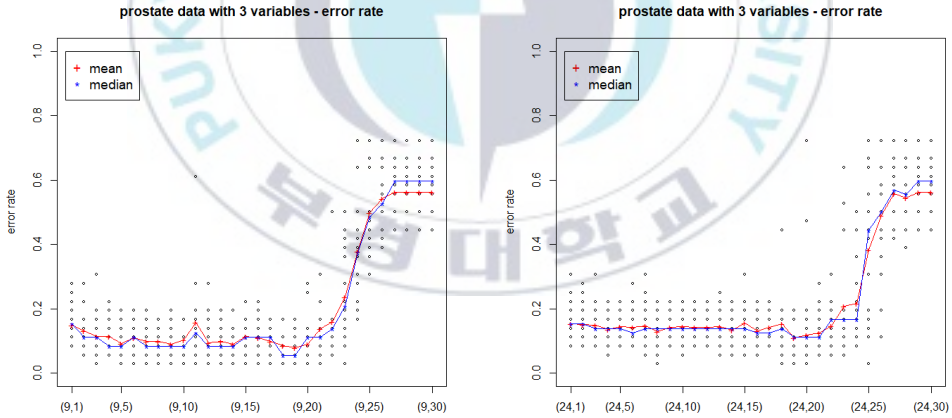

그림 4.2.9. (전립선 암 자료) 은닉층의 수 2, 입력 변수 선택의 경우 (9,1)~(9,30)과 (24,1)~(24,30)의 오류율 산점도

그림 4.2.9와 같이 일부 은닉층의 노드인 (9,1)~(9,30)과 (24,1)~(24,30)를 분석한 결과, 은닉층의 수가 1일 때의 패턴과 비슷하지만 변동이 조금 더 큰 것을 확인할 수 있다. (9,1)~(9,22)일 때 오류율의 평균 범위는 0.0791~0.1592, 중앙값은 0.056~0.153이며, (24,1)~(24,22)일 때 오류율의 평균 범위는 0.1097~0.15582, 중앙값은 0.111~0.167이다.

은닉층의 수가 3개일 때 그림 4.2.10과 같이 첫 번째, 두 번째, 세 번째 은닉층의 수를 각각 1~30까지 설정하여 실험을 하였으며, 첫 번째 은닉층의 노드 수에 따른 오류율의 변화가 반복되고 있다. 그림 4.2.11의 (8,1,1)~(8,30,30) 오류율 산점도를 보면, 두 번째 은닉층의 노드 수가 커질수록 오류율의 평균 및 중앙값이 상대적으로 작아지는 패턴을 보인다. 조금 더 자세히 패턴을 파악하기 위하여 그림 4.2.12와 같이 분석한 결과, 전반적으로 두 경우 모두 비슷한 패턴을 보인다. (8,3,1)~(8,3,21)일 때 오류율의 평균 범위는 0.0976~0.5525, 중앙값은 0.083~0.528이며, 세 번째 은닉층의 노드 수가 20일 때 오류율의 평균 및 중앙값이 가장 작았다. (8,19,1)~(8,19,20)일 때 오류율의 평균 범위는 0.1499~0.5525, 중앙값은 0.139~0.528이며, 오류율의 평균 및 중앙값은 세 번째 은닉층의 노드 수가 각각 21, 20일 때 가장 작다.

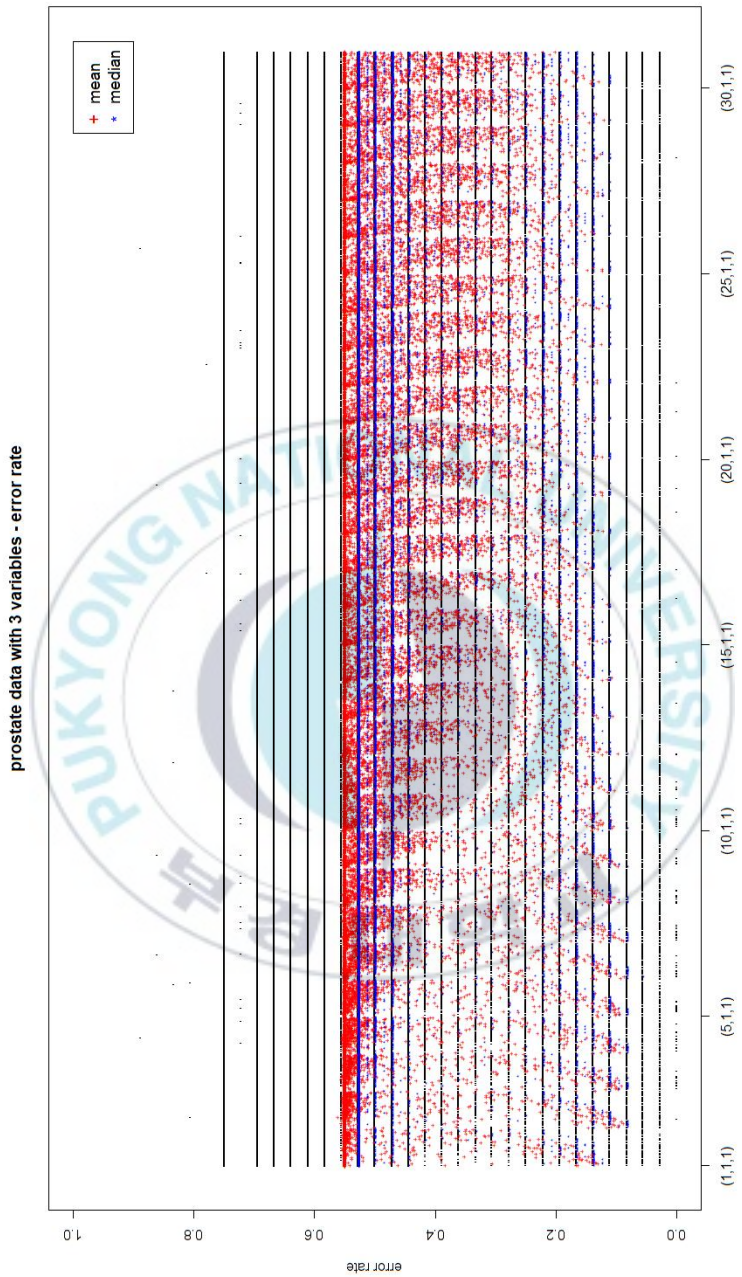

그림 4.2.10. (전립선 암 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (1,1,1)~(30,30,30)의 오류율 산점도

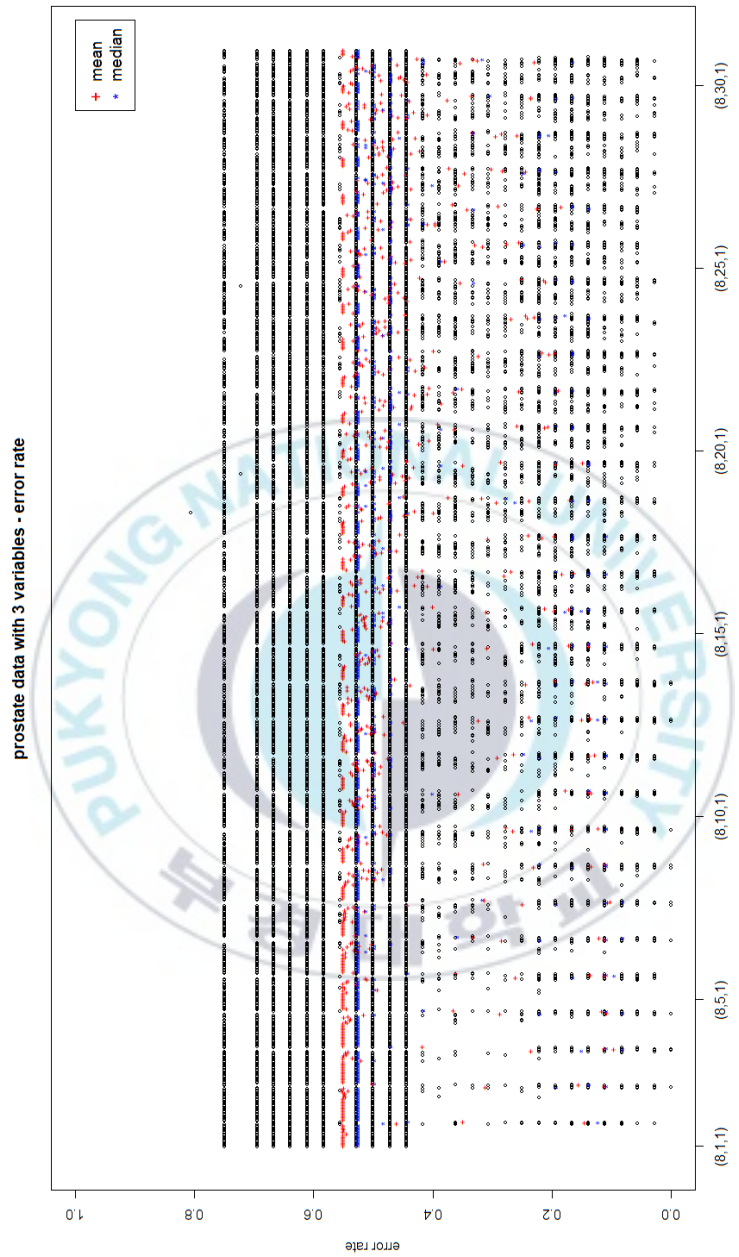

그림 4.2.11. (전립선 암 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (8,1,1)~(8,30,30)의 오류율 산점도

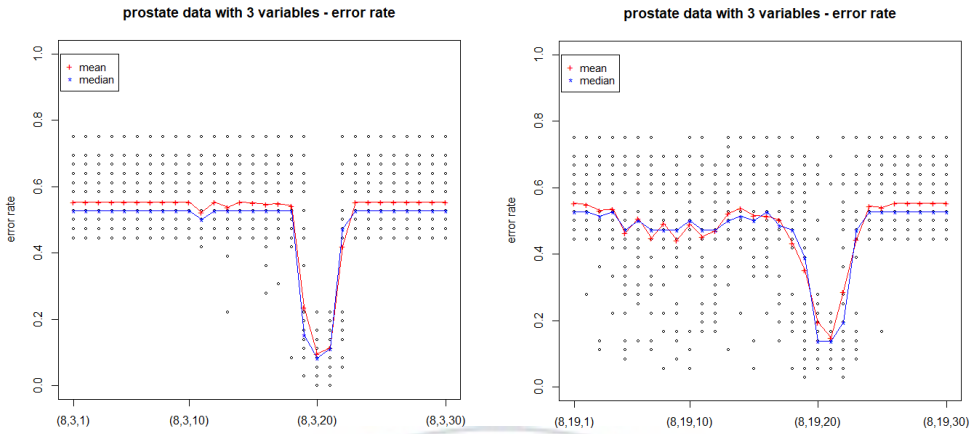

그림 4.2.12. (전립선 암 자료) 은닉층의 수 3, 입력 변수 선택의 경우 (8,3,1)~(8,3,30)과 (8,19,1)~(8,19,30)의 오류율 산점도

본 절에서는 두 고차원 자료에서 입력 변수 선택의 경우 은닉층 1, 2, 3일 때를 분석해보았다. 분석 결과, 은닉층의 수와 노드 수가 커질수록 계산시간뿐만 아니라 오히려 오류율도 증가했다. 따라서 두 고차원 자료 분석을 통하여 은닉층의 수 1, 노드 수 3~20을 사용하는 것이 적절하다고 판단된다.

4.3 심층 신경망 최적의 은닉층 수와 노드 수 탐색 비교

4.1절과 4.2절에서는 각각 모든 입력 변수를 사용한 경우와 입력 변수 선택을 사용한 경우를 살펴보았다. 본 절에서는 고차원 자료에서 두 경우의 결과를 서로 비교하고자 한다.

먼저, 백혈병 자료에 대하여 은닉층이 1개일 때 모든 입력 변수를 사용한 경우 오류율의 범위는 0.031~0.438이었다. 반면, 입력 변수 선택의 경우 노드 수가 1~24일 때 오류율의 평균 범위는 0.05391~0.07124, 중앙값 범위는 0.062~0.094이었다. 그림 4.3.1을 보면 모든 입력 변수를 사

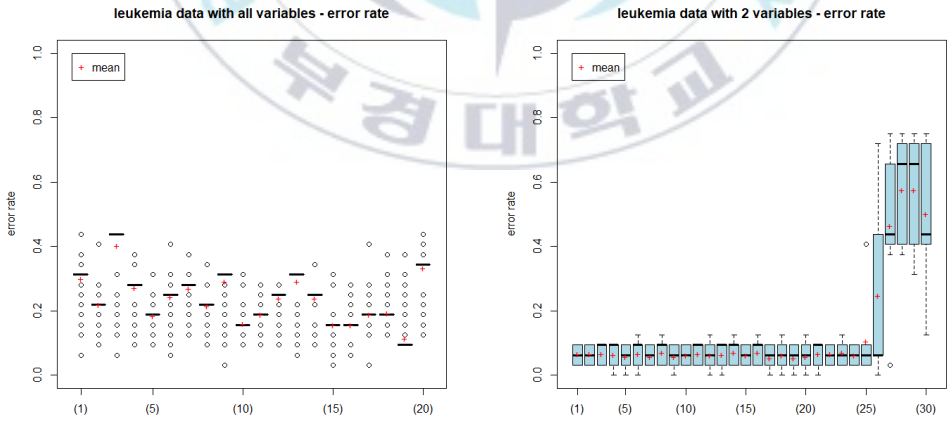

그림 4.3.1. (백혈병 자료) 은닉층의 수 1, 모든 입력 변수 사용과 입력 변수 선택의 경우 오류율 상자그림

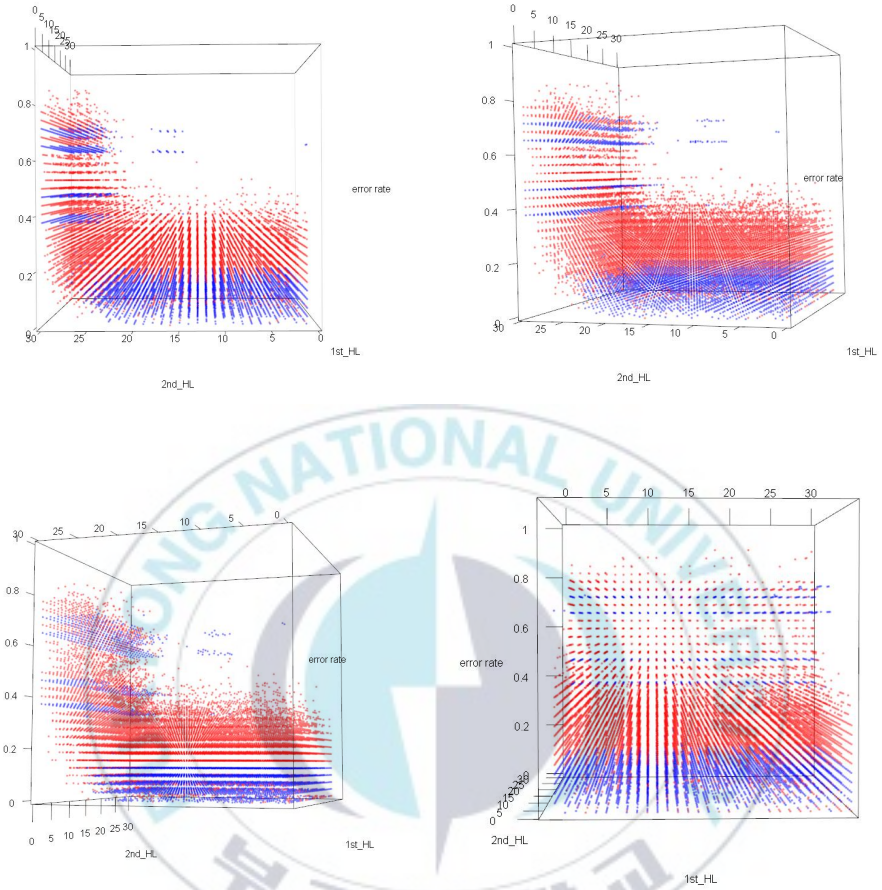

그림 4.3.2. (백혈병 자료) 은닉층의 수 2, 모든 입력 변수 사용(빨강)과 입력 변수 선택(파랑)의 경우 오류율 3차원 그림

용했을 때보다 선택된 입력 변수를 사용했을 때 변동의 폭이 훨씬 더 작다는 것을 확인할 수 있다. 또한, 모든 입력 변수를 사용한 경우 계산시간 범위는 2.71~8.95초였지만 입력 변수 선택의 경우 계산시간 범위는 0.12~0.35초였다. 따라서 오류율, 오류율의 변동 및 계산시간을 고려해보았을 때 입력 변수 선택을 하는 경우 더 좋은 결론을 얻을 수 있다고 판단된다.

그림 4.3.2는 은닉층의 수가 2개일 때 모든 입력 변수 사용과 입력 변수 선택의 경우의 오류율을 합친 3차원 그림이다. 모든 입력 변수 사용을 했을 때는 빨간색, 입력 변수 선택의 경우에는 파란색으로 표현하였다. 그림 4.3.2를 보면, 입력 변수 선택의 경우가 모든 입력 변수를 사용했을 때보다 변동이 훨씬 적으며 오류율이 낮다. 또한, 두 경우 모두 첫 번째 은닉층의 노드 수에 따른 오류율의 변화는 없으며 두 번째 은닉층의 노드 수가 작을 때 오류율이 낮은 것으로 판단된다. 은닉층의 수가 3개일 때에는 4.1.1절과 4.2.1절 결과로부터 입력 변수 선택을 했을 때의 오류율 값 및 변동이 조금 더 좋다는 것을 알 수 있다.

전립선 암 자료에 대하여 은닉층이 1개일 때 모든 입력 변수를 사용한 경우 오류율의 범위는 0.083~0.722이었다. 반면, 입력 변수 선택의 경우 노드 수가 1~23일 때 오류율의 평균 범위는 0.05575~0.118, 중앙값 범위는 0.056~0.111이었다.

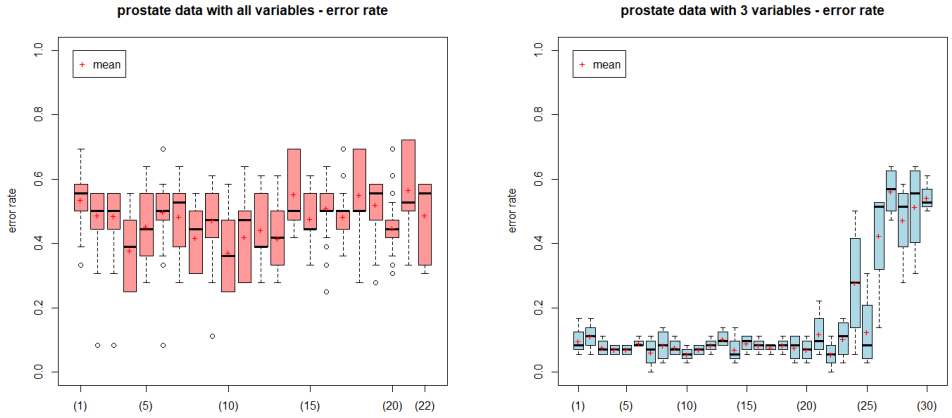

그림 4.3.3. (전립선 암 자료) 은닉층의 수 1, 모든 입력 변수 사용과 입력 변수 선택의 경우 오류율 상자그림

그림 4.3.3을 보면 모든 입력 변수를 사용한 경우에 비해 입력 변수 선택의 경우 변동의 폭이 작다는 것을 확인할 수 있다. 또한, 모든 입력 변수를 사용한 경우 계산시간 범위는 15.87~44.28초였지만 입력 변수 선택의 경우 계산시간 범위는 0.28~0.60초였다. 따라서 오류율, 오류율의 변동 및 계산시간을 모두 고려해보았을 때 입력 변수 선택을 하는 경우 더 좋은 결론을 얻을 수 있다고 판단된다.

은닉층의 수가 2개일 때의 그림 4.3.4를 보면, 입력 변수 선택의 경우가 모든 입력 변수를 사용했을 때보다 오류율이 더 작고 오류율의 변동도 매우 작았다. 또한, 두 경우 모두 첫 번째 은닉층의 노드 수에 따른 오류율의

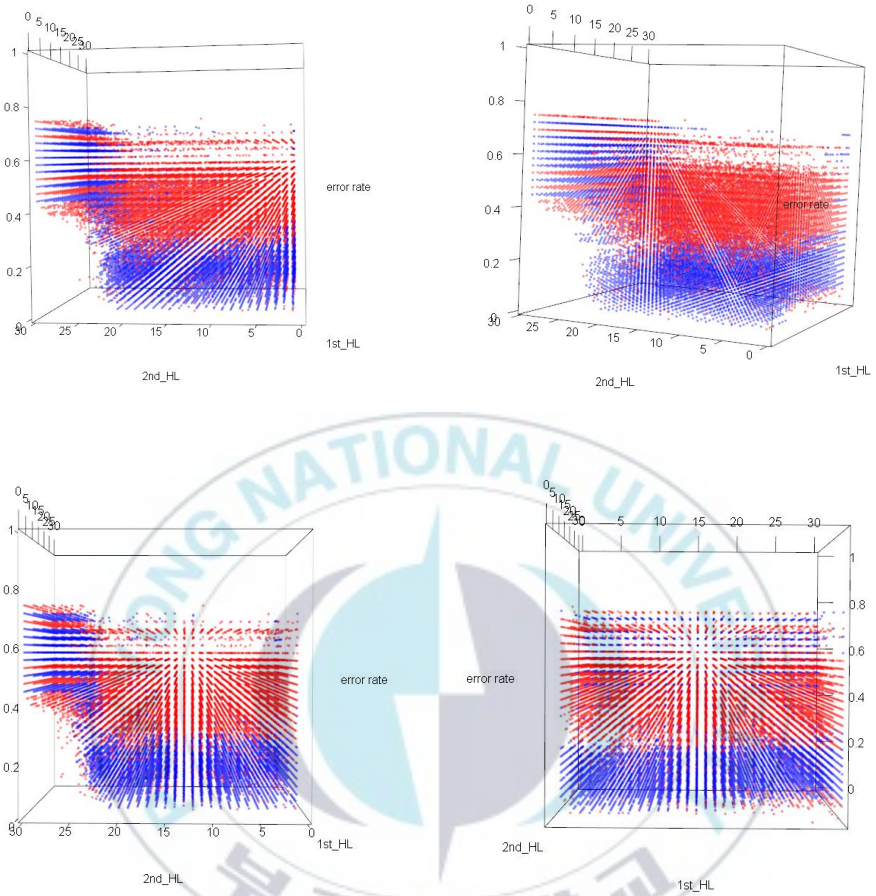

그림 4.3.4. (전립선 암 자료) 은닉층의 수 2, 모든 입력 변수 사용과 입력 변수 선택의 경우 오류율 3차원 그림

변화는 없으며 두 번째 은닉층의 노드 수가 적을 때 오류율이 낮은 것으로 판단된다. 은닉층의 수가 3개일 때에는 백혈병 자료에서와 마찬가지로 4.1.1절과 4.2.1절 결과로부터 큰 차이는 없지만 입력 변수 선택의 경우가 오류율 측면에서 조금 더 좋다는 것을 알 수 있다.

본 장에서는 최적의 은닉층 수 및 노드 수를 탐색하였다. 또한, 모든 입력 변수를 사용한 경우와 입력 변수 선택의 경우를 비교해보았다. 최적의 은닉층 수, 노드 수 및 변수 선택의 사용 여부를 결정할 때 오류율, 오류율의 변동, 계산시간을 고려하였다. 비록 오류율의 평균 및 중앙값이 작더라도 변동이 크다면 실제로 적용하기에는 적합하지 않기 때문이다. 분석 결과, 고차원 자료에서는 은닉층의 수와 노드 수가 커질수록 오류율과 계산시간이 증가하는 것을 확인할 수 있었다. 고차원 자료는 관측값의 수보다 입력 변수의 수가 더 크기 때문에 은닉층의 수와 노드 수가 커질수록 과적합(overfitting) 현상을 초래할 확률이 높으며, 이로 인하여 모형 성능이 떨어진다는 것을 짐작할 수 있다. 또한, 은닉층의 수가 1, 2, 3일 때 모두 입력 변수 선택의 경우가 모든 입력 변수를 사용한 경우보다 오류율이 낮고 변동이 작았으며, 계산시간도 훨씬 적게 걸렸다. 따라서 위의 결과를 바탕으로 고차원 자료에서는 불필요한 입력 변수 없이 간결한 모형일 때 모형 성능이 높아지는 것으로 판단된다.

Ⅴ. 고차원 자료에서 심층 신경망-영향 상자그림을 이용한 영향점 진단

5.1 심층 신경망-영향 상자그림

고차원 자료에서는 영향점을 진단하는 것이 매우 중요하다. 그 이유는 자료의 수 n 보다 변수의 수 p가 훨씬 더 크기 때문에 관측값이 회귀계수 추정에 영향을 줄 수 있으며, 변수 선택을 수행할 때 축소추정량에 영향점의 영향이 더 클 수 있기 때문이다. 관측값의 영향력을 판단하기 위한 방법으로는 수치적 접근법(numerical approach)과 그래픽 접근법(graphical approach)이 있다. 수치적 접근법으로 진단을 한다면 관측값의 영향력에 대한 정확한 값을 알 수 있지만 영향점의 영향 패턴을 이해하기는 어렵다. 반면 그래픽 접근법으로 진단을 한다면 관측값의 영향력에 대한 정확한 값을 그림에서 읽기가 어렵지만 영향점의 영향 패턴을 이해하기가 쉽다. 그래픽 접근법으로 Jang과 Anderson-Cook(2017)은 라쏘추정량 사용 시 영향점의 영향을 평가할 수 있는 라쏘 영향그림을 제안하였다. 안소진 등

(2017)은 고차원 자료에서 영향점을 평가하기 위한 그래픽 방법으로써 라쏘 영향 그림뿐만 아니라 변수선택 순위그림, 삼차원 라쏘 영향그림을 제안하였다.

본 연구에서 영향점을 평가하기 위한 그래픽 접근법으로서 심층 신경망-영향 상자그림을 제안하고자 한다. 심층 신경망 모형 적용 시 전체 입력변수를 사용한 경우와 입력 변수 선택을 한 경우를 구분하여 두 가지 결과를 모두 고려하였다. 분석 과정은 각각 algorithm 5.1과 algorithm5.2에서 설명된다. 이때 입력 변수 선택 방법으로서는 4장에서 살펴본 SIS 방법을 이용하였다.

Algorithm 5.1. 심층 신경망-영향 상자그림(모든 입력 변수 사용)

1. 입력 변수를 전체 사용하여 신경망 모형을 만든다.
2. $i = 1, 2, ..., n$에 대하여
   (1) (소거법) 전체 관측값을 넣었을 때와 i번째 관측값을 제거했을 때의 오류율(또는 평균제곱오차)을 계산한다.
   (2) (1)번의 작업을 100번 반복하여 계산한다.
3. 결과값을 바탕으로 상자그림을 그린다.
4. 전체 관측값을 넣었을 때와 i번째 관측값을 제거했을 때의 결과를 비교하여 상자그림에 변화가 심하게 나타난 관측값이 있다면 그 관측값을 영향점으로 판단한다.

Algorithm 5.2. 심층 신경망-영향 상자그림(선택된 입력 변수 사용)

1. 선택된 입력 변수를 사용하여 신경망 모형을 만든다.
2. $i = 1, 2, ..., n$에 대하여
   (1) (소거법) 전체 관측값을 넣었을 때와 i번째 관측값을 제거했을 때의 오류율(또는 평균제곱오차)을 계산한다.
   (2) (1)번의 작업을 100번 반복하여 계산한다.
3. 결과값을 바탕으로 상자그림을 그린다.
4. 전체 관측값을 넣었을 때와 i번째 관측값을 제거했을 때의 결과를 비교하여 상자그림에 변화가 심하게 나타난 관측값이 있다면 그 관측값을 영향점으로 판단한다.

5.2 자료 분석 및 결과 비교

본 연구에서는 3장에서 언급하였던 백혈병, 전립선암, 대장암 자료를 이용하여 분석을 하였다. 백혈병 자료와 전립선암 자료는 R에서 기본적으로 제공하고 있는 학습 자료(train set)와 테스트 자료(test set)를 이용하여 각각 38개, 102개의 학습 자료에 대한 영향점을 진단하였다. 대장암 자료에서는 학습 자료(train set)와 테스트 자료(test set)가 구분되어 있지 않아서 학습 자료와 테스트 자료를 7:3으로 세분화시켜 표 5.2.1과 같이 43개의 학습 자료와 12

표 5.2.1. 대장암 자료의 학습 자료와 테스트 자료 관측값 번호

학습 자료 관측값 번호	테스트 자료 관측값 번호
1, 2, 3, 5, 6, 7, 9, 10, 11, 14, 15, 16, 17, 19, 20, 22, 23, 24, 26, 27, 31, 32, 35, 36, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 53, 54, 55, 57, 58, 59, 60	4, 8, 12, 13, 18, 21, 25, 28, 29, 30, 33, 34, 39, 50, 51, 52, 56, 61, 62

개의 테스트 자료로 구분하였다. 이를 토대로 62개의 학습 자료에 대한 영향력을 진단하였다.

첫 번째 예제는 38개의 학습 자료와 34개의 테스트 자료로 이루어진 백혈병 자료이다. 전체 변수들을 고려했을 때에는 그림 5.2.1과 같으며, SIS를 통하여 선택된 V3320, V4847 변수를 고려했을 때에는 그림 5.2.2와 같다. 첫 번째 열인 전체 관측값을 사용했을 때의 결과와 큰 차이가 나는 관측값을 확인해본 결과, 그림 5.2.1에서는 없지만 그림 5.2.2에는 17, 29, 38번이다. 따라서 심층 신경망-영향 상자그림을 통하여 총 3개의 관측값이 영향점으로 판단되었다.

그래픽 접근법의 영향력 측도인 HIM plot과 영향 그림을 이용하여 결과를 비교해보았다. 그림 5.2.3은 HIM plot이며, 가로축은 관측값 번호이고 세로축은 주변 상관에 기초하여 제안된 고차원적 영향력 측도인 D_k 값이다.

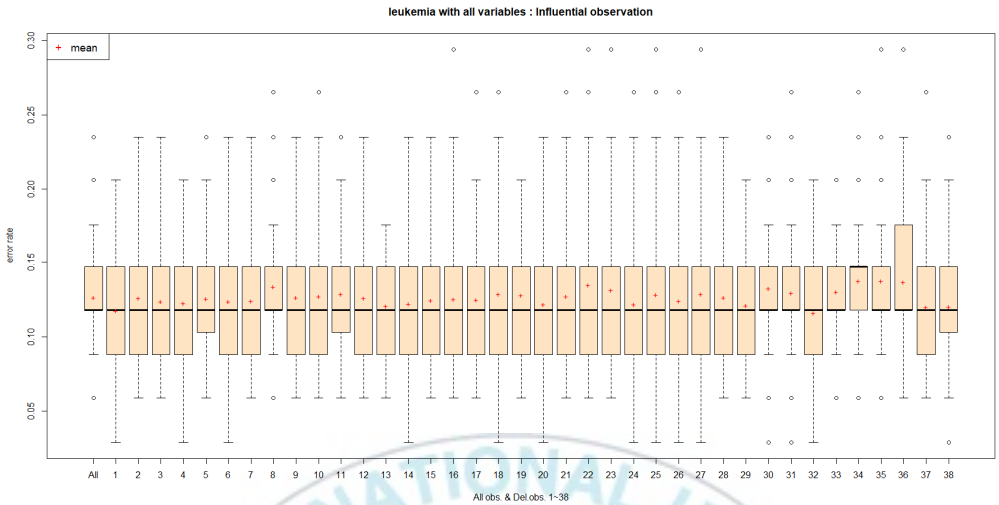

그림 5.2.1. (백혈병 자료) 모든 입력변수 사용했을 때의 심층 신경망-영향 상자그림

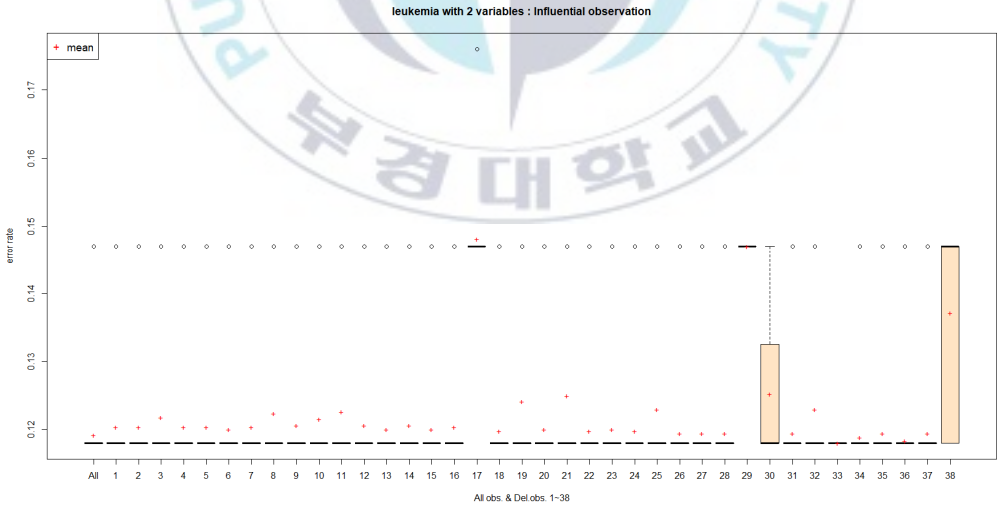

그림 5.2.2. (백혈병 자료) 선택된 입력 변수들을 사용했을 때의 심층 신경망-영향 상자그림

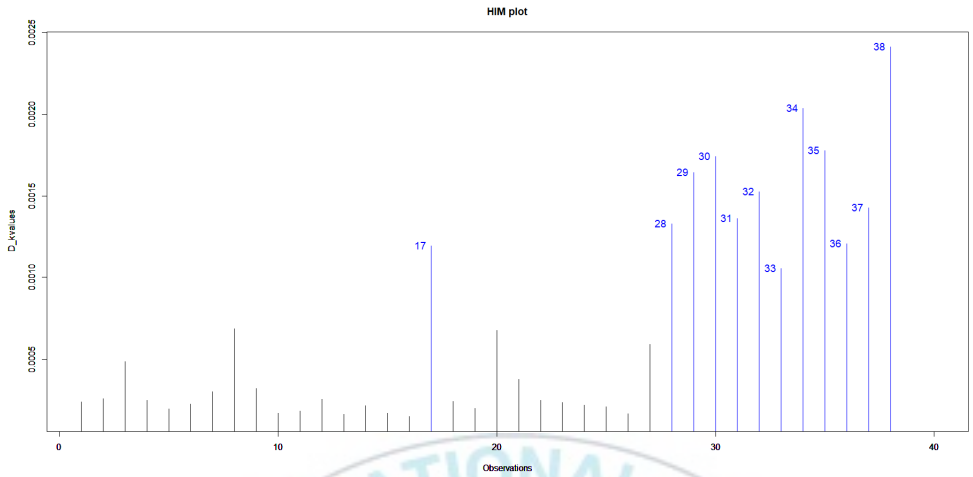

그림 5.2.3. (백혈병 자료) HIM plot

분석 결과 17, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38번 관측값을 제거하였을 때 D_k 값이 상대적으로 크게 나왔다. 따라서 HIM plot에서는 11개의 관측값이 영향점으로 판단되었다. 영향 그림의 분석 결과는 그림 5.2.4와 같다. 모든 관측값에 대하여 SIS를 이용하여 변수 선택을 한 결과, 총 7,129개의 변수 중 V3320, V4847가 유의한 변수로 선택이 되었으며 분석 결과에 대한 영향 그림은 그림 5.2.4의 첫 번째 그림과 같다. 이때 두 변수의 추정된 회귀계수는 그림에서 점선으로 나타내었다. 소거법을 이용하여 변수 선택을 한 결과 2, 3, 17, 21, 37번 관측값에서 V3320, V2020가 유의한 변수로 선택이 되었으며, 추정된 회귀계수는 그림에 실선으로 나타내었다. 따라서 영향 그림에서는 5개의 관측값이 영향점으로 판단

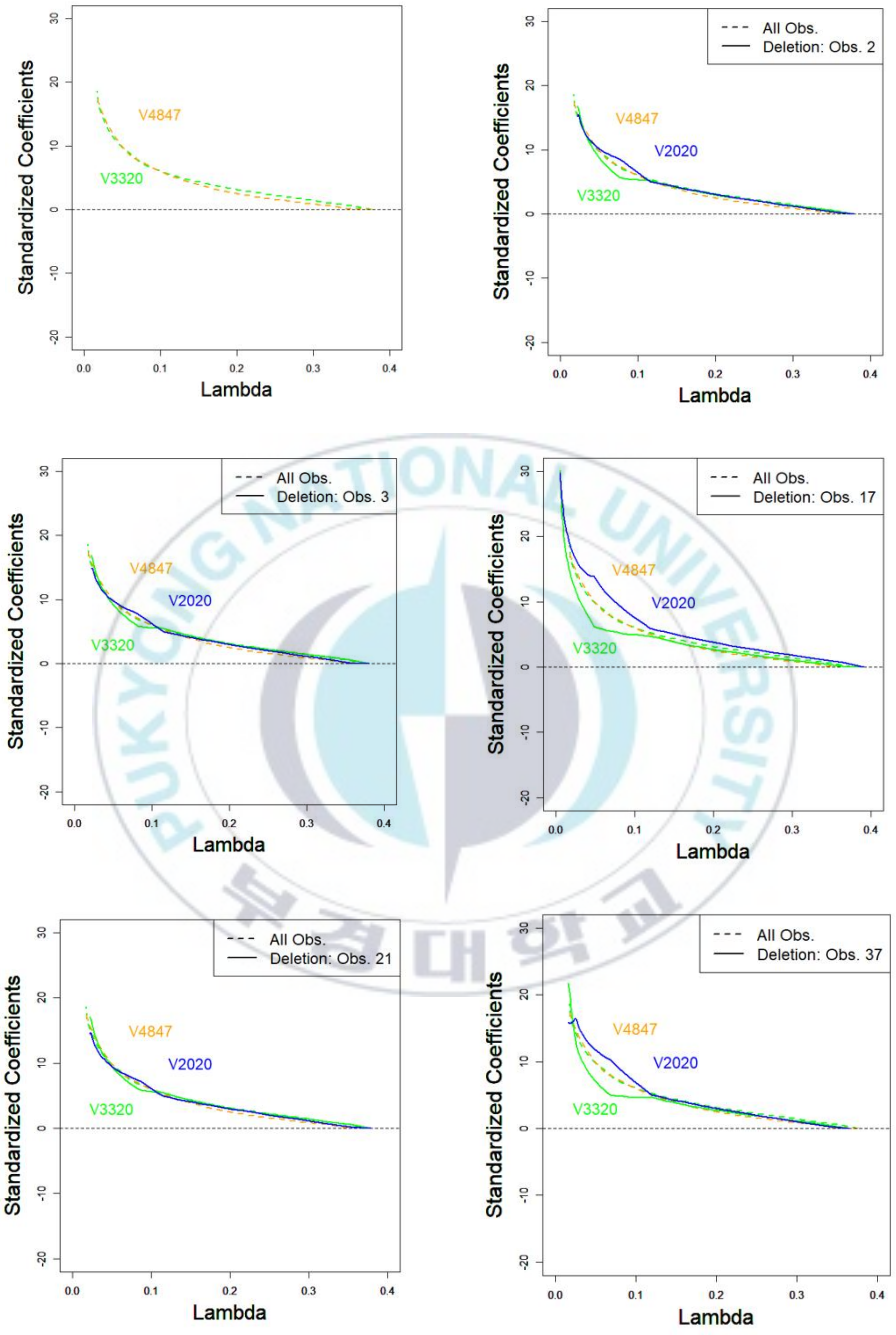

그림 5.2.4. (백혈병 자료) 영향 그림

되었다. 세 가지 영향력 측도를 이용하여 진단한 영향점 결과는 표 5.2.2와 같았다. 세 영향력 측도에서 모두 17번 관측값을 영향점으로 진단하였으며, 29, 38번 관측값은 심층 신경망-영향 상자그림과 HIM plot, 37번 관측값은 영향 그림과 HIM plot에서 각각 공통적으로 영향점이라고 평가하였다.

표 5.2.2. (백혈병 자료) 영향력 측도들을 이용한 영향점 결과

심층 신경망- 영향 상자그림	영향 그림	HIM plot
17, 29, 38	2, 3, 17, 21, 37	17, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38

두 번째 예제는 102개의 학습 자료와 34개의 테스트 자료로 이루어진 전립선 암 자료이다. 전체 변수를 고려했을 때에는 그림 5.2.5와 같으며, SIS를 통하여 선택된 V4231, V6185, V8965 세 개의 변수를 고려했을 때에는 그림 5.2.6과 같다. 그림 5.2.5에서는 전체 관측값을 사용했을 때 그려진 오류율에 대한 상자그림 결과와 크게 차이가 나는 관측값은 없지만 그림 5.2.6에서는 10, 54, 62, 66, 73, 84, 95번 관측값이 전체 관측값을 사용했을 때의 결과와 큰 차이를 보인다. 따라서 심층 신경망-영향 상자그림에서는 총 7개의 관측값이 영향점으로 판단되었다.

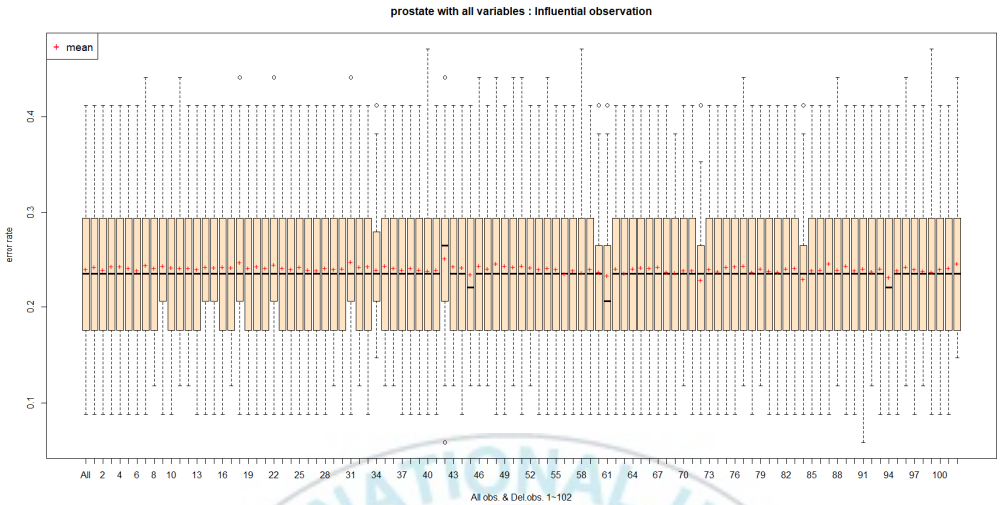

그림 5.2.5. (전립선 암 자료) 모든 입력 변수 사용했을 때의 심층 신경망-영향 상자그림

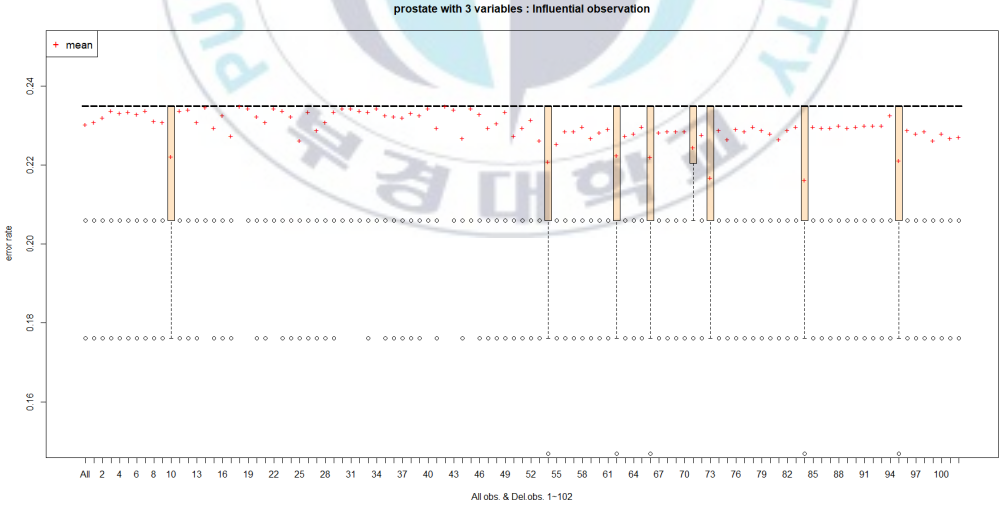

그림 5.2.6. (전립선 암 자료) 선택된 입력 변수들을 사용했을 때의 심층 신경망-영향 상자그림

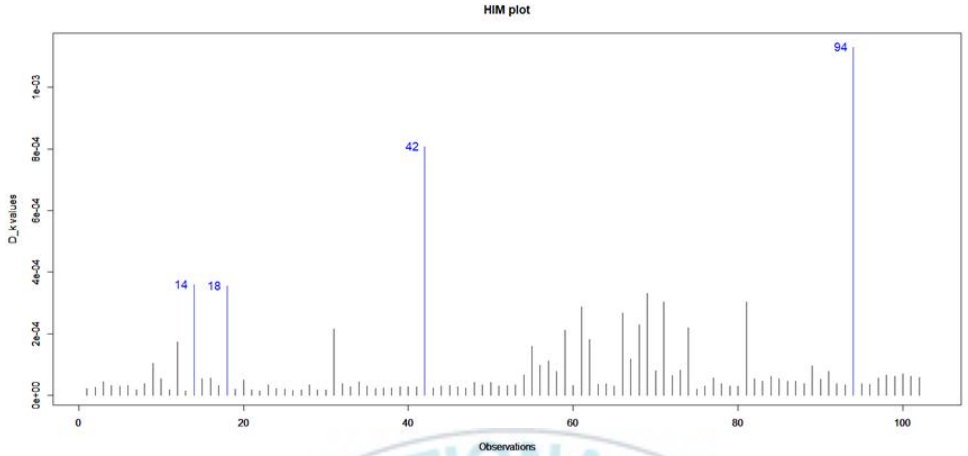

그림 5.2.7. (전립선 암 자료) HIM plot

그림 5.2.7을 보면 HIM plot에서는 14, 18, 42, 94번 관측값의 D_k 값이 크며 그중에서 94번과 42번 값이 매우 큰 값을 나타낸다. 영향 그림의 분석 결과는 그림 5.2.8과 같다. 모든 관측값에 대하여 SIS를 이용하여 변수 선택을 한 결과, 총 12,600개의 변수 중 V4231, V6185, V8965가 유의한 변수로 선택이 되었으며 분석 결과에 대한 영향 그림은 그림 5.2.8의 첫 번째 그림과 같다. 소거법을 이용하여 변수 선택을 한 결과 18, 42, 73, 84번 관측값에서 각각 다른 변수가 유의하다고 선택이 되었다. 세 가지 영향력 측도를 이용하여 진단한 영향점 결과는 표 5.2.3과 같다. 73, 84번 관측값은 심층 신경망-영향 상자그림과 영향 그림, 18, 42번 관측값은 영향 그림과 HIM plot에서 각각 공통적으로 영향점이라고 평가하였다.

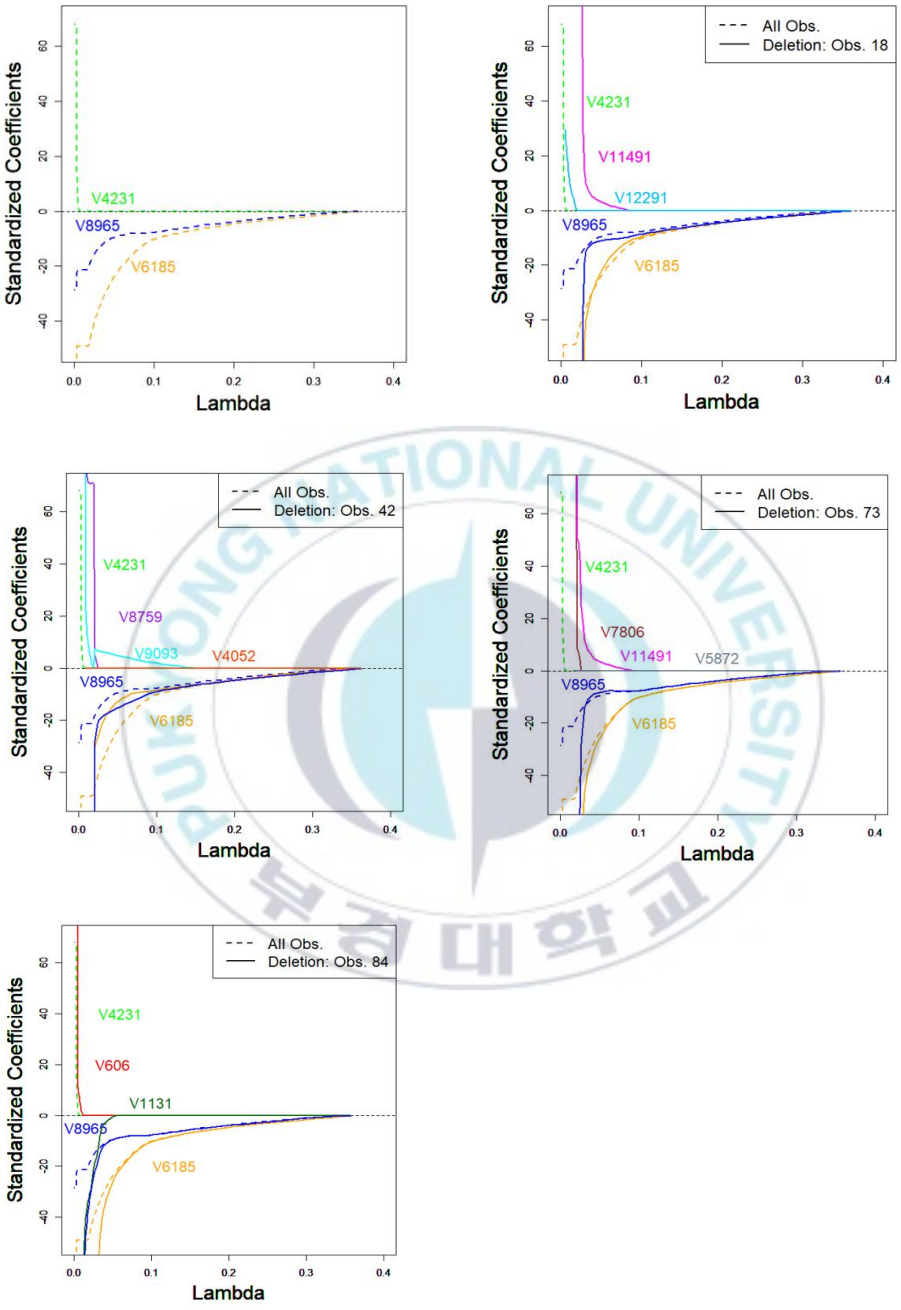

그림 5.2.8. (전립선 암 자료) 영향 그림

표 5.2.3. (전립선 암 자료) 영향력 측도들을 이용한 영향점 결과

심층 신경망- 영향 상자그림	영향 그림	HIM plot
10, 54, 62, 66, 73, 84, 95	18, 42, 73, 84	14, 18, 42, 94

세 번째 예제는 43개의 학습 자료와 19개의 테스트 자료로 이루어진 대장암 자료이다. 전체 변수로 분석했을 때에는 그림 5.2.9와 같으며, SIS를 통하여 선택된 V66, V765 변수로 분석했을 때에는 그림 5.2.10과 같다. 그림 5.2.9에서는 관측값 2, 4, 12, 18, 20, 22, 30, 36, 41을 각각 제거하였을 때 오류율의 변동이 상대적으로 크다. 그림 5.2.10에서는 관측값 2, 5, 8, 10, 11, 15, 18, 21, 30, 32, 33, 38, 39, 42를 제거하였을 때 변동이 크게 나타난다. 따라서 심층 신경망-영향 상자그림에서는 총 20개의 관측값이 영향점으로 판단되었다. 그림 5.2.11은 HIM plot의 결과이며, 17~43번 관측값이 영향점으로 판단되었다. 영향 그림의 분석 결과는 그림 5.2.12 ~ 그림 5.2.15와 같다. 총 2000개의 변수 중 V66, V765가 유의한 변수로 선택이 되었으며 분석 결과에 대한 영향 그림은 그림 5.2.12의 첫 번째 그림과 같다. 소거법을 이용하여 변수 선택을 한 결과 2, 5, 7, 10, 12, 14, 18, 20, 22, 23, 25, 28, 29, 30, 32, 35, 36, 38, 39번 관측값이 영향점으로 판단되었다.

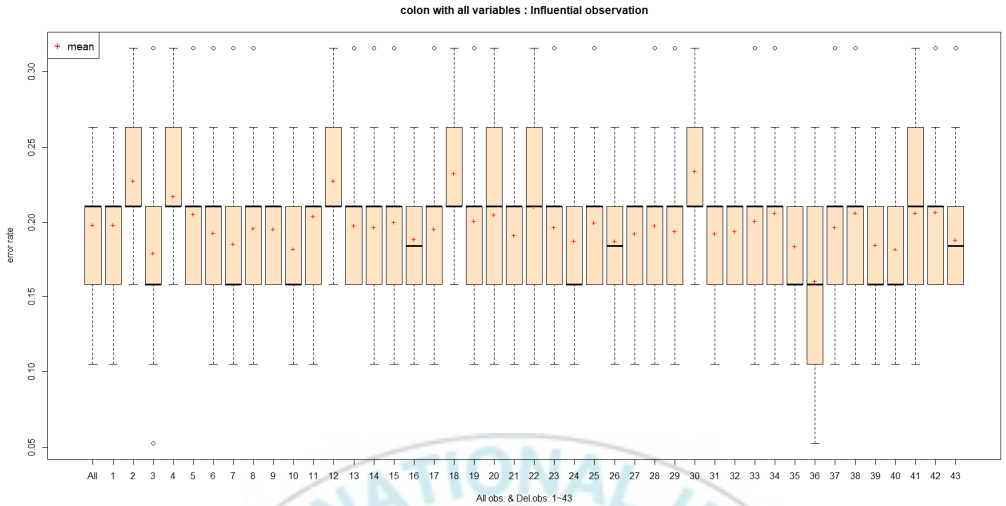

그림 5.2.9. (대장암 자료) 모든 입력 변수 사용했을 때의 심층 신경망-영향 상자그림

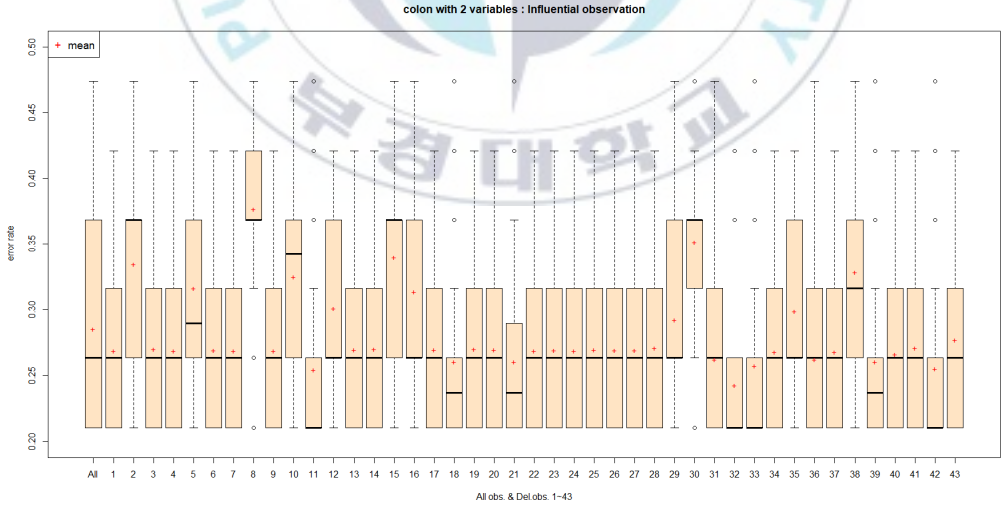

그림 5.2.10. (대장암 자료) 선택된 입력 변수들을 사용했을 때의 심층 신경망-영향 상자그림

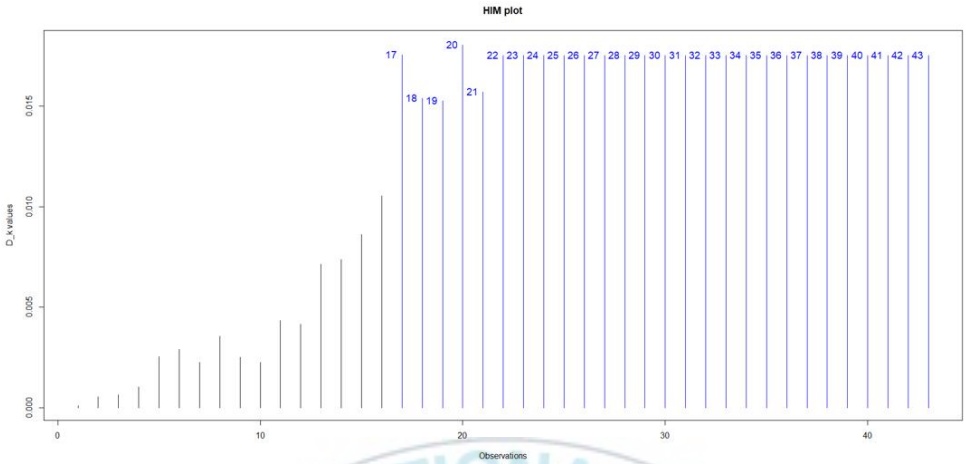

그림 5.2.11. (대장암 자료) HIM plot

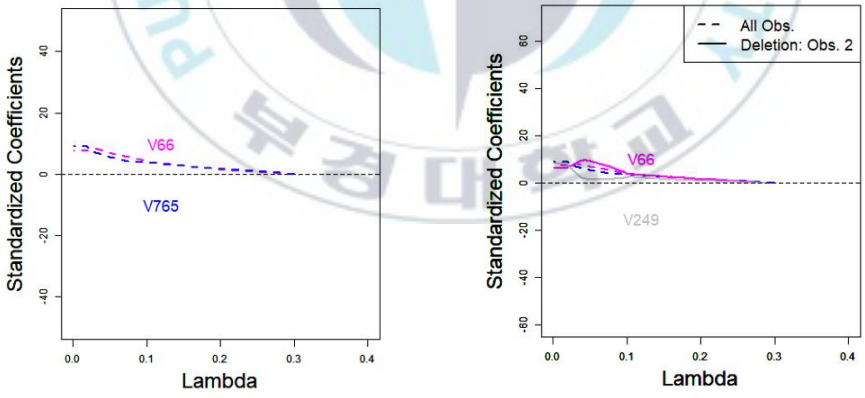

그림 5.2.12. (대장암 자료) 영향 그림 – 1

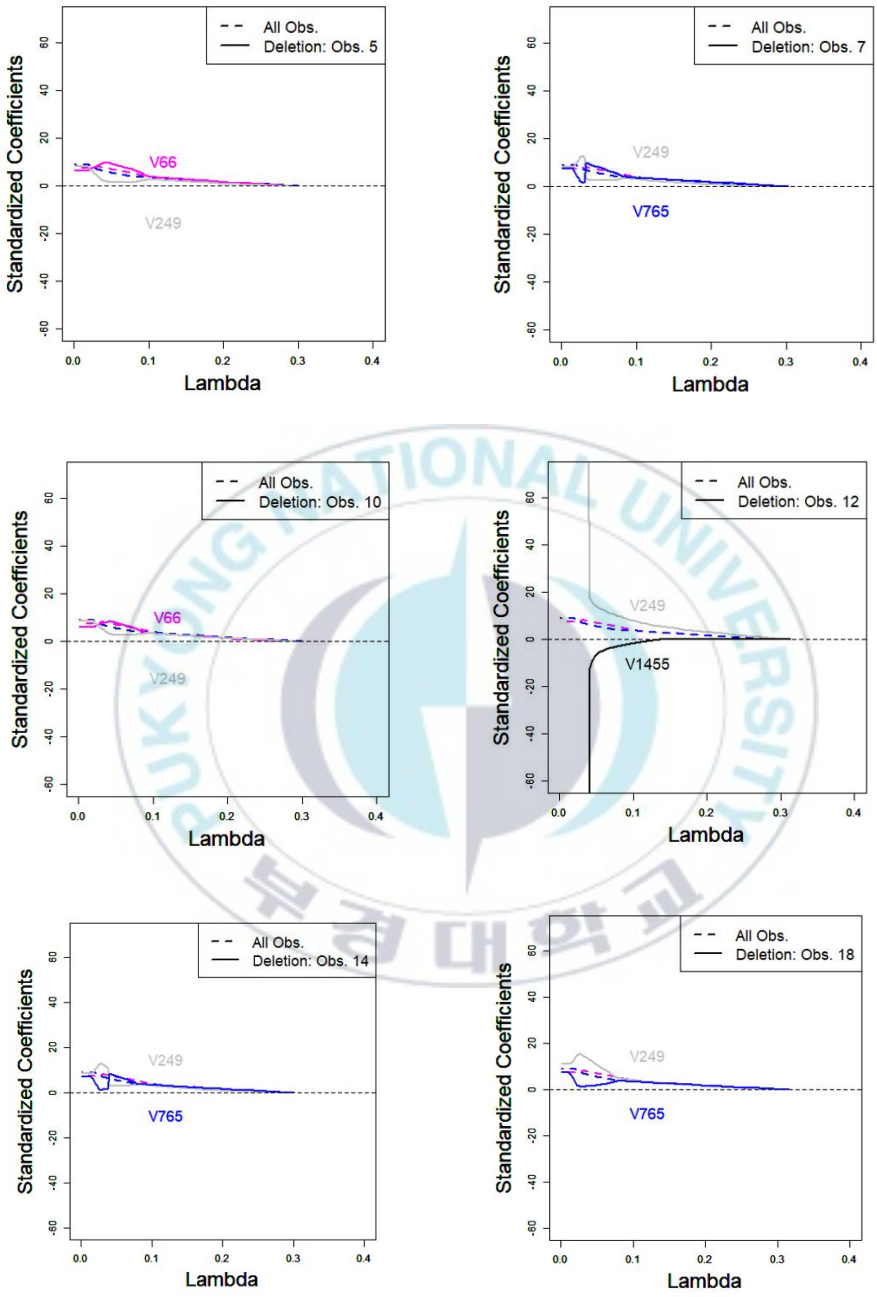

그림 5.2.13. (대장암 자료) 영향 그림 – 2

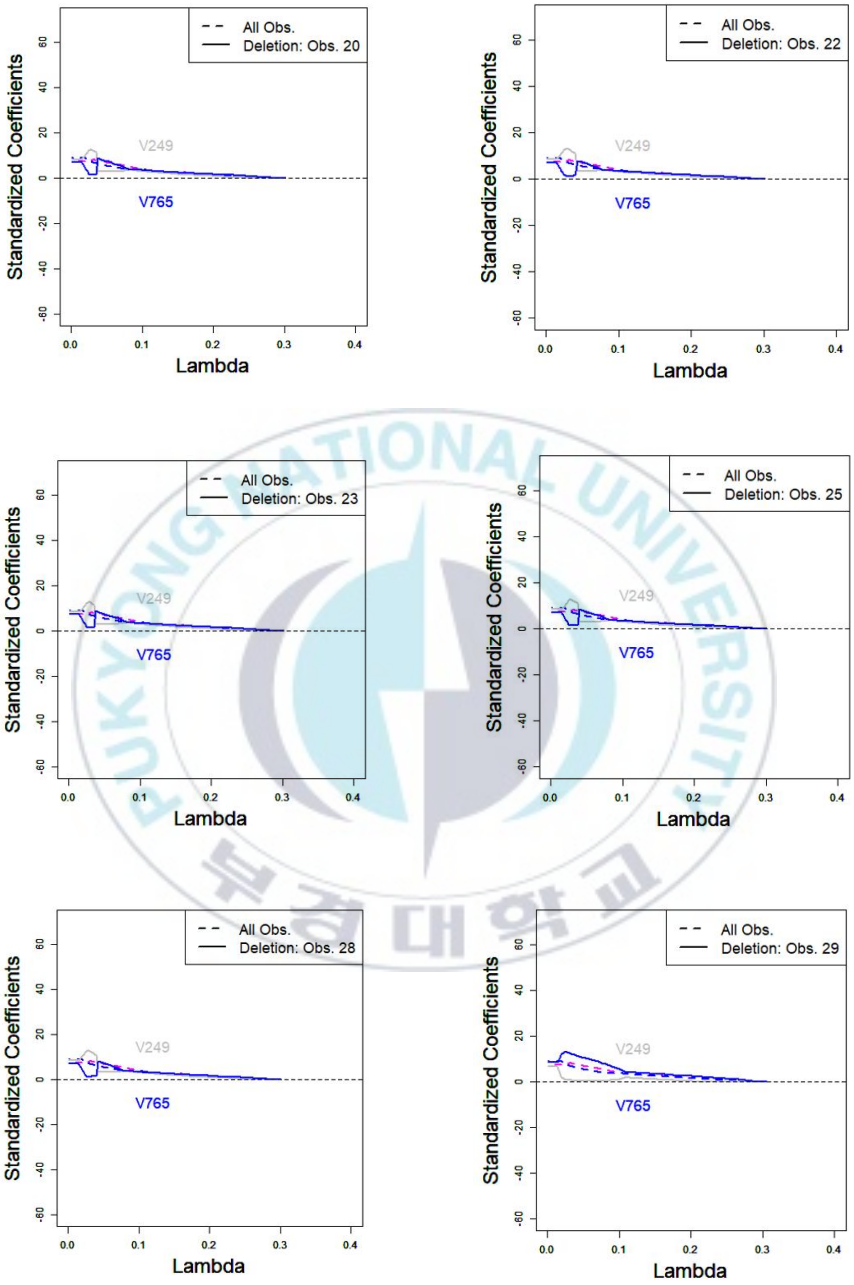

그림 5.2.14. (대장암 자료) 영향 그림 – 3

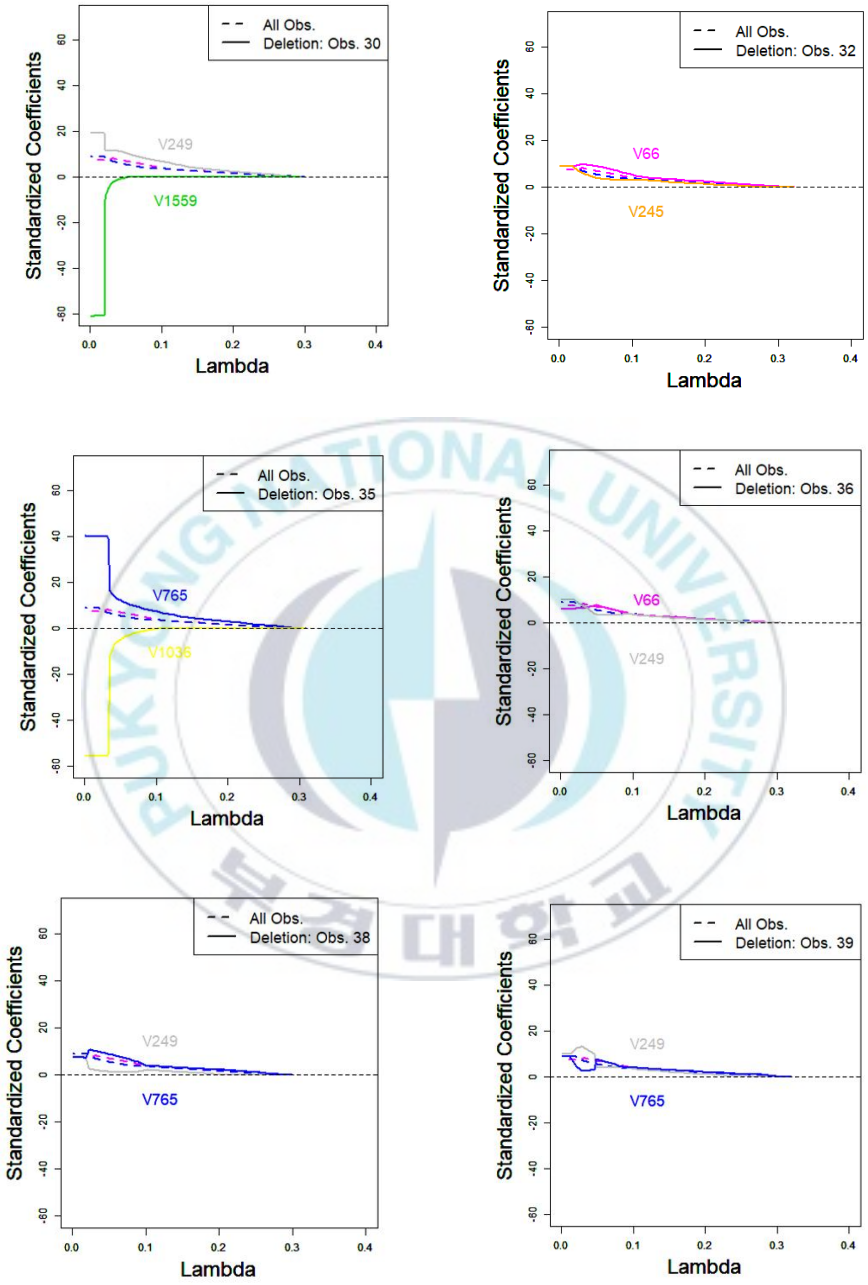

그림 5.2.15. (대장암 자료) 영향 그림 – 4

세 가지 영향력 측도를 이용하여 진단한 영향점 결과는 표 5.2.4와 같다. 세 영향력 측도에서 모두 18, 20, 22, 30, 32, 36, 38, 39번 관측값을 영향점으로 진단하였다. 대장암 자료는 앞의 두 예제와 달리 많은 관측값들이 영향점으로 판단되었다. 이는 고차원 자료의 특성상 자료의 수는 작지만 차원이 커지게 되면서 많은 관측값들이 영향점으로 진단된 것으로 보인다.

표 5.2.4. (대장암 자료) 영향력 측도들을 이용한 영향점 결과

심층 신경망- 영향 상자그림	영향 그림	HIM plot
2, 4, 5, 8, 10, 11, 12, 15, 18, 20, 21, 22, 30, 32, 33, 36, 38, 39, 41, 42	2, 5, 7, 10, 12, 14, 18, 20, 22, 23, 25, 28, 29, 30, 32, 35, 36, 38, 39	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43

본 장에서는 관측값의 영향력을 진단하는 측도로써 그래픽 방법인 심층 신경망-영향 상자그림을 제시하였다. 세 가지 고차원 예제를 통하여 전체 변수를 심층 신경망 모형에 적용하여 분석한 결과와 SIS를 통하여 선택된 변수를 모형에 적용하여 구한 결과를 종합하여 영향점을 진단할 수 있다는 것을 보였다. 그리고 영향그림과 HIM plot 결과들의 비교를 통하여 다른 측도들의 결과와도 함께 비교 할 필요가 있다는 것을 알 수 있었다.

Ⅵ. 결 론

기계학습에서 사용자가 직접 설정해야 하는 초매개변수는 모형의 성능에 큰 영향을 미치기 때문에 초매개변수의 최적화 문제에 대한 중요성이 더욱 커지고 있다. 하지만 이론적으로 어떠한 값을 사용해야 하는지 알려진 바가 없어서 경험적으로 초매개변수를 설정해야 하는 문제를 가지고 있었다. 이를 극복하기 위하여 그동안 초매개변수를 탐색하는 연구가 다양하게 이루어지고 있음을 선행연구들을 통하여 알 수 있었다. 하지만 격자탐색에서는 탐색시간이 오래 걸리는 문제, 임의 탐색에서는 초매개변수에 대한 난수를 이용하는 문제, 베이지안 최적화에서는 초매개변수를 확률변수로 보고 초매개변수의 사전 분포를 통한 통계적 방법을 사용하기 때문에 최적의 값이라고 장담하기 어려운 문제 등 온전히 적용하기에는 여전히 한계가 존재한다. 또한 ‘차원의 저주’로 불리는 고차원 자료는 대부분의 변수들이 소음이므로 위 상관관계와 과적합 문제가 발생할 수 있으며 회귀분석시 모형행렬에 대한 적률행렬의 역행렬을 구할 수가 없어서 전통적인 통계기법들을 적용하기가 어렵다. 이러한 문제를 해결하기 위하여 많은 연구가 이루어지고 있지만 여전히 풀리지 못한 난제 중 하나로 여겨지고 있다.

본 연구에서는 초매개변수 최적화 문제의 한계를 극복할 수 있는 방법으로 실험계획법과 격자탐색을 조합한 하이브리드 탐색을 제안하였다. 본 연구에서 제시한 방법은 실험계획법을 이용하여 초매개변수 값을 미리 선별하기 때문에 탐색시간이 오래 걸리는 문제를 극복할 수 있었다. 또한 선별된 값들에 대하여 격자 탐색을 하기 때문에 많은 조합의 경우를 고려할 수 있는 장점이 있었다. 그리고 세 가지 고차원 예제를 이용하여 심층 신경망에 하이브리드 탐색을 적용해보았으며, 실험계획법과 하이브리드 탐색 방법을 적용한 결과들을 각각 비교해보았다. 고차원 예제에서 심층 신경망의 초매개변수 탐색을 위하여 하이브리드 탐색 방법을 적용하였을 때 백혈병 자료에서는 88.44%, 전립선 암 자료에서는 86.28%의 정확도를 보였다.

본 연구의 4장에서는 탐색적 자료 분석 방법을 통하여 고차원 자료에서 최적의 은닉층 수와 노드 수를 확인하였다. 이때 모든 입력 변수를 사용한 경우와 선택된 입력 변수들을 사용한 경우를 비교분석 해보았다. 오류율, 오류율의 변동, 계산시간을 기준으로 비교분석을 한 결과, 두 가지 고차원 예제에서 모두 은닉층의 수 1, 노드 수 1~20 그리고 선택된 입력 변수를 사용했을 때 높은 성능을 보이는 것으로 판단되었다. 복잡한 모델은 단순한 모델보다 편향(bias)을 쉽게 줄이는 대신 분산(variance)은 커지게 된다. 즉, 훈련오차를 아주 작게 줄일 수는 있지만 학습된 모델이 선택된 훈련자료에만 과도하게 민감하여 과적합(overfitting) 문제를 발생시킬 수가

있다. 따라서 고차원 자료에서는 복잡한 모델보다 단순한 모델을 선호하는 절약의 원리(principle of parsimony)가 적용되어 분산을 줄이는 모델을 더 선호하는 것으로 판단된다.

고차원 자료에서는 변수의 개수가 관측값의 수보다 과도하게 많기 때문에 영향점의 영향이 매우 클 수 있으므로 영향점을 진단하는 것이 매우 중요하다. 이러한 문제를 해결하기 위하여 많은 선행 연구들이 있어왔다. 하지만 방법론에 따라 선택되는 영향점들이 조금씩 다를 수 있기 때문에 결과들을 서로 비교분석하는 것이 필요하다. 본 연구의 5장에서는 영향점을 진단하는 그래픽 방법으로서 심층 신경망-영향 상자그림을 제안하였다. 세 가지 고차원 예제들을 통하여 영향점을 진단할 수 있음을 보였으며, 고차원 자료에서 영향점을 진단하는 그래픽 방법들인 영향 그림과 HIM plot의 결과들을 서로 비교분석을 해보았다.

향후 연구 방향은 다음과 같다. 먼저, 하이브리드 탐색 방법의 성능을 확인하기 위하여 임의탐색, 자동 선택 방법, 베이지안 최적화 등 다양한 최적화 방법들의 결과를 비교분석 해볼 것이다. 두 번째로, 여러 가지 고차원 예제들을 이용하여 최적의 은닉층 수, 노드 수 및 입력 변수 선택 여부를 확인하고, 심층 신경망-영향 그림을 적용해볼 것이다.

참 고 문 헌

[1] 김충락, & 강근석. (2010). 회귀분석. *교우사, 서울*.

[2] 안소진, 이재은, & 장대흥. (2017). 고차원 자료에서 영향점의 영향을 평가하기 위한 그래픽 방법. *한국데이터정보과학회지*, ***28***(6), 1291-1300.

[3] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., &Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, ***96***(12), 6745-6750.

[4] Andrews, D. F., & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society: Series B (Methodological)*, ***40***(1), 85-93.

[5] Bellman, R. (1961). Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ.*

[6] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics John Wiley &Sons. *New York*.

[7] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*,

13(Feb), 281-305.

[8] Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, *35*(6), 2313-2351.

[9] Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18.

[10] Fan, J., Feng, Y., Saldana, D. F., Samworth, R., & Wu, Y. (2017). http://www.stat.columbia.edu/ ~yangfeng/pubs/jss1375.pdf, Package SIS.

[11] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.

[12] Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849-911.

[13] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531-537.

[14] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[15] Huang, G. B. (2003). Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Transactions on Neural Networks*, **14**(2), 274-281.

[16] Jang, D. H., & Anderson-Cook, C. M. (2017). Influence plots for LASSO. *Quality and Reliability Engineering International*, **33**(7), 1317-1326.

[17] Jang, W., Kim, G., & Kim, J. (2016). Current trends in high dimensional massive data analysis. *Korean Journal of Applied Statistics*, **29**(6), 999-1005.

[18] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

[19] Kim, C., Lee, J., Yang, H., & Bae, W. (2015). Case influence diagnostics in the lasso regression. *Journal of the Korean Statistical Society*, **44**(2), 271-279.

[20] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning.

nature, ***521***(7553), 436-444.

[21] Lee, J., & Jang, D. H. (2018a). Influence plots for shrinkage estimators in high-dimensional data. Data Science & Visualisation 2018.

[22] Lee, J., & Jang, D. H. (2018b). Exploratory data analysis on the optimal number of hidden layers and nodes in deep neural network. Computational Statistics 2018.

[23] Lujan-Moreno, G. A., Howard, P. R., Rojas, O. G., & Montgomery, D. C. (2018). Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. *Expert Systems with Applications*, ***109***, 195-205.

[24] Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, ***5***(1), 18.

[25] Shen, H. Y., Wang, Z. X., Gao, C. Y., Qin, J., Yao, F., &Xu, W. (2008). Determining the number of BP neural network hidden layer units. *Journal of tianjin University of Technology*, ***24***(5), 13-15.

[26] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J.,

Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, *1*(2), 203-209.

[27] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951-2959).

[28] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 847-855). ACM.

[29] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

[30] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91-108.

[31] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894-942.

[32] Zhao, J., Leng, C., Li, L., & Wang, H. (2013). High-dimensional influence measure. *The Annals of Statistics*, ***41***(5), 2639-2667.

[33] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, ***67***(2), 301-320.

[34] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, ***101***(476), 1418-1429.