딥러닝 기반 앙상블을 이용한 유방암 분류

최도연¹, 정광모¹, 임동훈²

¹부산대학교 통계학과, ²경상대학교 정보통계학과

Breast Cancer Classification using Deep Learning-based Ensemble

Do-Yeon Choi¹, Kwang-Mo Jeong¹, Dong Hoon Lim²

¹Department of Statistics, Busan National University, Busan; ²Department of Information Statistics and RINS, Gyeongsang National University, Jinju, Korea

Objectives: We propose a deep learning-based ensemble for improving breast cancer classification and compare it with existing six models including deep neural network on two UCI data. **Methods:** We propose a deep learning-based stacking ensemble method. We first applied five classifications methods individually, which were k-nearest neighbor, decision trees, support vector machines, discriminant analysis, and logistic regression analysis and then adopt a deep learning to the predictions derived from these methods after using 5-fold cross validation technique. We compared the proposed deep learning-based ensemble method with these methods for two UCI data through classification accuracy, ROC curves and c-statistics. **Results:** Experimental results for two UCI data showed that the proposed deep learning-based ensemble outperformed single k-nearest neighbor, decision trees, support vector machines discriminant analysis, and logistic regression analysis as well as deep neural network in terms of various performance measures. **Conclusions:** We proposed deep learning-based ensemble for improving breast cancer classification. The deep learning-based ensemble outperformed existing single models for all applications in terms of various performance measures.

Key words: Breast cancer, Classification, Deep learning, Ensemble, Performance evaluation

서 론

최근 유방암(breast cancer)의 발병률이 높아지면서 유방암에 의한 사망률을 줄이기 위한 조기 발견 및 정확한 진단은 무엇보다 중요하다. 2017년 12월 보건복지부가 발표한 중앙암등록본부의 통계에 따르면 2015년 우리나라에서 새로 발생한 암환자 수는 214,701명으로 그 중 유방암은 19,219명으로 전체 암의 9.0%로 다섯 번째로 많이 발생한 암이고 여성에게 발생한 암 중에서는 갑상샘암에 이어 두 번째의 발생빈도를 보이고 있으며 1999년 이후 지속적인 증가추세를 보이고 있다[1].

유방암을 진단하는 데 기본적인 방법으로 유방촬영술(mammogra-

phy), 유방초음파(ultrasonography), 세침 검사(fine needle aspiration, FNA) 등이 있다[2,3]. 유방촬영술은 가장 대표적이고 기본적인 유방검사법이고 유방초음파 검사는 고해상도 초음파 기기를 이용하여 유방 질환을 진단하는 검사로 주로 유방촬영술에서 치밀 유방을 보이는 여성에게 사용되는 검사이다. 세침 검사는 가는 주사바늘을 종양이 있는 부위에 찔러 세포를 채취하여 암인지 여부를 검사하는 방법으로 수술하지 않고 검사하기 때문에 간편하나 충분한 양의 조직을 얻기 힘들기 때문에 정밀한 진단을 내리는 데 한계를 갖고 있다.

최근 빠른 컴퓨터 기술 발달로 인하여 머신러닝(machine learning)이 암 진단에 활용되고 있다. 여기서 머신러닝이란 컴퓨터를 인간처럼

Corresponding author: Dong Hoon Lim

501 Jinjudae-ro, Jinju 52828, Korea
Tel: +82-55-772-1465, E-mail: dhl@gnu.ac.kr

Received: February 25, 2018 Revised: May 14, 2018 Accepted: May 24, 2018

*This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2017R1E1A1A03071057).

No potential conflict of interest relevant to this article was reported.

How to cite this article:

Choi DY, Jeong KM, Lim DH. Breast cancer classification using deep learning-based ensemble. J Health Info Stat 2018;43(2):140-147. Doi: <https://doi.org/10.21032/jhis.2018.43.2.140>

© It is identical to the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permit unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2018 Journal of Health Informatics and Statistics

학습시켜, 스스로 규칙을 생성하도록 하는 기술을 말하는데 대표적인 방법으로는 k-근접 이웃(k-nearest neighbor), 의사결정나무(decision tree), 신경망(neural network), 서포트 벡터 머신(support vector machine), 판별분석(discriminant analysis), 로지스틱회귀분석(logistic regression analysis) 등이 있다[4-8]. k-근접 이웃 방법은 가장 유사한 상위 k개의 데이터로부터 다수결의 원칙(majority voting)에 의해 새로운 데이터의 분류를 결정하는 방법으로 높은 정확도를 갖는 반면에 계산비용이 높다는 단점을 갖고 있다. 의사결정나무는 분류함수를 의사결정 규칙으로 이루어진 나무모양으로 그려서 분류하는 방법으로 이해하기 쉬운 장점을 갖고 있으나 가지분할을 위한 변수 선택에 따라 결과가 달라지는 어려움이 있다. 그리고 신경망은 사람의 두뇌가 의사 결정하는 형태를 모방하여 분류하는 방법으로 과대적합(over-fitting)과 국소최적화(local optimization) 등의 한계점을 갖고 있다. 그리고 서포트 벡터 머신은 학습과정에서 마진(margin)을 최대화하는 초평면(hyper-plane)을 추정한 후 새로운 데이터를 분류하는 방법으로 신경망에서 비해 과대적합이 덜하고 예측의 정확도가 높으나 결과에 대한 설명력이 떨어진다는 단점이 있다. 판별분석은 판별함수를 이용하여 개체들이 소속 그룹을 결정하는 방법으로 독립변수들이 다변량 정규분포(multivariate normal distribution)를 따른다는 가정하에서 효과적이고, 로지스틱 회귀분석은 종속변수가 범주형인 경우 사용하는 회귀분석의 확장방법으로 엄격한 가정을 요구하는 판별분석에 대한 대안으로 많이 사용한다. 위에서 살펴본 바와 같이 모든 경우에 우수한 성능을 갖고 있는 만능의 머신러닝 기술은 존재하지 않으며 제각기 장단점을 갖고 있다.

최근 들어 많이 이용되고 있는 딥러닝(deep learning)은 머신러닝 기술의 한 분야로, 인공신경망(artificial neural network)에 기반을 둔 심층신경망(deep neural network)으로 다양한 분야에서 뛰어난 성능을 보이고 있다[9].

본 연구에서는 유방암 분류성능 향상을 위해 딥러닝기반 앙상블(deep learning-based ensemble) 모형을 구축하고 기존의 단일 모형들과의 성능을 비교하고자 한다. 앙상블 방법으로 배깅(bagging), 부스팅(boosting), 스택킹(stack) 등이 있다. 배깅은 주어진 학습자료에서 부트스트랩 표본(bootstrapped sample)을 생성하여 단일모형 결과들의 통합 전략(aggregation strategy)에 의해 분류하는 방법이고, 부스팅은 잘 분류되지 않는 약한 분류모형들을 결합하여 강한 예측모형을 만드는 방법이다. 스택킹은 배깅, 부스팅보다는 덜 알려졌지만 형태가 다른 여러 모형(heterogeneous models)을 결합함으로써 보다 안정적인 결과값을 도출할 수 있는 모형이다[10,11].

본 연구에서 제안한 딥러닝기반 앙상블 모형은 두 단계에 걸쳐 이루어진다. 첫 번째 단계에서는 스택킹 앙상블 방법을 수행한 후 두 번째

단계에서는 딥러닝을 적용하여 최종 분류한다.

모형의 성능 평가는 University of California at Irvine (UCI)의 머신러닝 데이터베이스(UCI Machine Learning Repository) [12]에서 제공하는 유방암 자료를 가지고 평가하고자 한다.

연구 자료

본 연구에서 분류모형의 성능평가에 사용한 자료는 UCI의 유방암에 대한 Wisconsin Original Breast Cancer (WOBC) 자료와 Wisconsin Diagnostic Breast Cancer (WDBC) 자료이다.

WOBC 자료

WOBC 자료는 1992년 UCI Machine Learning Repository [12]에 의해 제공되었고 많은 연구자에 의해 패턴 인식과 기계학습에 사용되고 있다. WOBC 자료는 699명에 대해 조사되었고 Table 1과 같이 클래스를 나타내는 변수와 FNA의 세포 특성을 나타내는 9개의 변수로 구성되어 있다.

여기서 세포 특성값은 1부터 10까지 스케일된 값을 가지고 1에 가까울수록 양성을 나타내고 10에 가까울수록 악성을 나타낸다.

WDBC 자료

WDBC 자료는 1995년 UCI Machine Learning Repository [12]에 의해 제공되었고 WOBC 자료와 함께 패턴 인식과 기계학습에 널리 사용되고 있는 자료이다. WDBC 자료는 569명에 대해 조사하였고 Table 2와 같이 양성과 악성을 나타내는 클래스 변수, 즉 Diagnosis 변수 1개와 세포 특성을 나타내는 10개의 변수, 즉 radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension에 대해 각각 평균과 표준편차 그리고 이상값(혹은 최대값)을 조사한 총 30개의 변수로 구성되어 있다.

Table 1. WOBC data attribute information

Variables	Values
Clump thickness	1-10
Uniformity of cell size	1-10
Uniformity of cell shape	1-10
Marginal adhesion	1-10
Single epithelial cell size	1-10
Bare nuclei	1-10
Bland chromatin	1-10
Normal nucleoli	1-10
Mitoses	1-10
Class	2 = Benign, 4 = Malignant

WOBC, Wisconsin original breast cancer.

Table 2. WDBC data attribute information

Variables	Descriptions
Diagnosis	M=Malignant, B=Benign
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	
Area	
Smoothness	Local variation in radius lengths
Compactness	Perimeter ² / area - 1.0
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	-
Fractal dimension	"coastline approximation" - 1

WDBC, Wisconsin diagnostic breast cancer.

예를 들면, radius 변수에 대해 3개의 변수 즉, 평균을 나타내는 mean radius 변수, 표준편차를 나타내는 radius SD 변수, 그리고 이상 값을 나타내는 worst radius 변수가 존재한다.

연구 방법

기존의 분류 모형

k-근접 이웃

k-근접 이웃 모형은 분류하고자 하는 새로운 데이터에 대해 그 데이터와 가장 유사한 k개의 데이터 집합을 찾아서 그 집합의 데이터들이 속하는 다수결의 집단으로 분류하는 방법이다. 이때 k 값의 선택은 분류 결과에 많은 영향을 준다. k 값이 너무 작으면 근접 이웃 분류 모형은 훈련용 데이터의 잡음 때문에 과대적합될 가능성이 있고 반대로 k 값이 너무 크면 분류하고자 하는 데이터에 가까운 집단으로 분류되지 않을 가능성이 있다.

의사결정나무

의사결정나무는 여러 가지 규칙을 순차적으로 적용하면서 독립 변수 공간을 분할하는 분류 모형으로 단계별로 설명하면 다음과 같다 [13].

단계 1: 여러 가지 독립 변수 중 하나의 독립 변수를 선택하고 그 독립 변수에 대한 기준값을 정한다. 이를 분류 규칙이라고 한다.

단계 2: 전체 학습 데이터 집합(부모 노드)을 해당 독립 변수의 값이 기준값보다 작은 데이터 그룹(자식 노드 1)과 해당 독립 변수의 값이 기준값보다 큰 데이터 그룹(자식 노드 2)으로 나눈다.

단계 3: 각각의 자식 노드에 대해 위의 단계 1과 단계 2를 반복하여 하위의 자식 노드를 만든다. 단, 자식 노드에 한 가지 클래스의 데이터

만 존재한다면 더 이상 자식 노드를 나누지 않고 중지한다.

서포트 벡터 머신

다음의 학습자료 (learning dataset) D 가 주어졌다고 가정하자.

$$D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in R^p, y_i \in \{-1, +1\}\}_{i=1}^n$$

여기서 y_i 는 p 차원 벡터 $\mathbf{x}_i$ 가 속하는 클래스를 나타내는 값으로 +1 혹은 -1이다. 서포트 벡터 머신에서 $\mathbf{x}_i$ 를 두 클래스로 분류하기 위한 최적 분리 초평면 (optimal separating hyperplane)은 각 클래스에 속하는 점들 중에서 서포트 벡터를 지나는 두 개의 평행한 초평면들 사이의 거리, 즉 마진을 최대화 함으로써 결정된다. 우리는 슬랙변수(slack variable) $\xi_i (\geq 0)$ 을 도입하여 다음과 같이 최적화 문제를 형식화할 수 있다.

$$\begin{aligned} \text{minimize} \quad & Q(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \begin{cases} y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1 - \xi_i, & i = 1, \dots, n \\ \xi_i \geq 0, & i = 1, \dots, n \end{cases} \end{aligned} \quad (3.1)$$

여기서 C 는 마진의 최대화와 분류 오류율의 최소화 사이 트레이드-오프(trade-off)를 결정하는 모수이다. 커널 함수(kernel function) $K(\mathbf{x}_i, \mathbf{x}_j)$ 을 도입하여 식(3.1)의 최적화 문제를 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \text{maximize} \quad & Q(\boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \begin{cases} y_i \alpha_i = 0, & i = 1, \dots, n \\ C \geq \alpha_i \geq 0, & i = 1, \dots, n \end{cases} \end{aligned}$$

여기서 α_i 는 라그랑지 배수(Lagrange multiplier)이다. 따라서 서포트 벡터를 사용하여 최적의 분리 평면을 다음과 같이 입력벡터 $\mathbf{x}$ 의 결정함수를 나타낸다.

$$D(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

여기서 S 는 서포트 벡터를 나타내고 b 는 원점으로부터 거리(offset)이다.

판별분석

판별분석은 최초 Fisher [14]에 의하여 체계화된 통계이론으로서 전체 집단을 두 개 이상의 집단으로 분류함에 있어 분류오차를 최소화할 수 있는 판별규칙을 도출하여 새로운 개체를 소속 집단으로 분류하는 방법으로 독립변수들은 다변량 정규분포를 이루고, 또한 각 집

단의 공분산 행렬이 동일하다는 가정하에 성립된다.

로지스틱 회귀분석

로지스틱 회귀분석이란 분석하고자 하는 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우 개별 관측치들이 어느 집단으로 분류될 수 있는가를 분석하는 방법으로 n 개의 독립변수를 $X_1, \dots, X_n$ 이라 할 때 로지스틱 회귀모형은 다음과 같다.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

여기서 $\beta_0, \dots, \beta_n$ 는 회귀계수이다. p 는 개별 관측치가 어떤 집단에 속할 확률이고 $1-p$ 는 그 집단에 속하지 않을 확률이다. 따라서 이 값이 높을수록 그 집단에 속할 확률이 높다는 것을 의미한다.

딥러닝 기반 양상불 모형

딥러닝

Figure 1은 입력층(input layer)과 출력층(output layer) 사이에 여러 개의 은닉층(hidden layer)을 갖고 있는 심층 신경망의 예를 보여주고 있다. 심층신경망은 많은 은닉층의 개수를 늘림으로써 여러 비선형 변환 기법의 조합을 통해 높은 수준의 추상화(abstractions)가 가능한 신경망 구조이다.

Figure 1에서 층의 수가 n_ℓ 개이고 층 ℓ 를 L_ℓ 이라 표현하자. 그러면 L_1 은 입력층, L_{n_ℓ} 은 출력층, 그리고 $L_2, \dots, L_{n_\ell-1}$ 은 은닉층을 나타낸다. 그리고 신경망에서 학습할 모수는 다음과 같다.

$$\begin{aligned} \mathbf{W} &= \{W^1, W^2, \dots, W^{n_\ell}\} \\ \mathbf{b} &= \{b^1, b^2, \dots, b^{n_\ell}\} \end{aligned}$$

여기서

$$\begin{aligned} W^\ell &= \{W_{ij}^\ell\}, j = 1, \dots, s_\ell, i = 1, \dots, s_{\ell+1}, \ell = 1, \dots, n_\ell \\ b^\ell &= \{b_{ij}^\ell\}, j = 1, \dots, s_\ell, i = 1, \dots, s_{\ell+1}, \ell = 1, \dots, n_\ell \end{aligned}$$

Figure 1. An example of deep neural network structure.

이때 W_{ij}^ℓ 은 층 ℓ 에서 뉴런 j 와 층 $\ell+1$ 에서 뉴런 i 를 연결하는 가중치(weight)이고 b_i^ℓ 는 층 ℓ 에서 뉴런 i 의 바이어스(bias)이다.

심층신경망을 통해 학습할 m 개의 훈련 데이터는 다음과 같다.

$$\{(x^1, y^1), (x^2, y^2), \dots, (x^m, y^m)\}$$

우리는 확률적 경사 하강법(stochastic gradient descent)을 통하여 심층신경망을 학습하고자 한다. 이를 위해 비용함수(cost function)를 다음과 같이 정의한다.

$$\begin{aligned} J(\mathbf{W}, \mathbf{b}) &= \frac{1}{m} \sum_{i=1}^m J(W, b; x^i, y^i) + \frac{\lambda}{2} \sum_{\ell=1}^{n_\ell} \sum_{j=1}^{s_{\ell-1}} \sum_{i=1}^{s_\ell} (W_{ij}^\ell)^2 \\ &= \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{2} \|h_{W, b}(x^i) - y^i\|^2 \right) + \frac{\lambda}{2} \sum_{\ell=1}^{n_\ell} \sum_{j=1}^{s_{\ell-1}} \sum_{i=1}^{s_\ell} (W_{ij}^\ell)^2 \end{aligned} \quad (3.2)$$

여기서 첫 번째 항은 평균 제곱 오차 항(mean square error term)이고 두 번째 항은 정규화 항(regularization term)을 나타내고 λ 은 두 항의 상대적인 중요도를 조절하는 가중치 감소 파라미터(weight decay parameter)이다. 그리고 비선형 활성화 함수 $h_{\mathbf{W}, \mathbf{b}}(x)$ 는 다음과 같이 정의한다.

$$h_{\mathbf{W}, \mathbf{b}}(x) = f(\mathbf{W}^T x + \mathbf{b}) \quad (3.3)$$

식(3.2)의 비용함수를 최소화하는 파라미터는 다음과 같이 반복적인 방법에 의해 구한다.

$$W_{ij}^\ell = W_{ij}^\ell - \alpha \frac{\partial}{\partial W_{ij}^\ell} J(\mathbf{W}, \mathbf{b}) \quad (3.4)$$

$$b_i^\ell = b_i^\ell - \alpha \frac{\partial}{\partial b_i} J(\mathbf{W}, \mathbf{b}) \quad (3.5)$$

여기서 α 는 학습률을 나타낸다.

심층신경망 학습은 식(3.3)식에 의해 활성화 함수값을 계산하고 피드 포워드 방식에 의해 최종 출력값과 실제값의 오차를 계산한 후 오차만큼 역전파를 통해 식(3.4)과 식(3.5)에 의해 가중치와 바이어스를 업데이트한다. 여기서 $\frac{\partial}{\partial W_{ij}^\ell} J(\mathbf{W}, \mathbf{b})$ 과 $\frac{\partial}{\partial b_i} J(\mathbf{W}, \mathbf{b})$ 은 역전파 알고리즘(back propagation algorithm)을 이용하여 계산한다.

스태킹 양상불 모형

제안된 딥러닝기반 양상불 모형은 Figure 2에서 보는 것처럼 크게 두 단계에 걸쳐 이루어진다. 첫 번째 단계에서는 입력 데이터에 스태킹 양상불 방법을 수행한 후 두 번째 단계에서는 딥러닝을 적용하여 최종 분류한다.

5-조각 교차 타당성하에서 Figure 2의 딥러닝기반 양상불 모형을 단계별로 설명하면 다음과 같다.

[First Stage]

Figure 2. A flowchart of deep learning-based ensemble model.

단계 1. 주어진 자료 D 를 조각 크기가 같도록 5-조각 $D_1, D_2, \dots, D_5$ 으로 분할한다. 여기서 $D_k = \{x_k, y_k\}, k = 1, \dots, 5$ 이다.

단계 2. 분할된 5-조각 중에서 4조각의 합집합 즉, $\bigcup_{k=2}^5 D_k$ 을 훈련용 자료로 이용하고 나머지 1조각, $D_1 = \{x_1, y_1\}$ 을 테스트용 자료로 이용한다.

단계 3. 훈련용 자료 $\bigcup_{k=2}^5 D_k$ 에 기본 분류모형 i , 여기서 $i = 1, \dots, 5$ 를 적용하여 모형 h_i 를 생성하고 테스트용 자료 $D_1 = \{x_1, y_1\}$ 에 생성된 모형을 적용하여 예측확률 $h_i(x_1)$ 을 계산한다. 그리고 기본 분류모형에 대한 결과값들을 모은다(assemble). 즉,

$$H_1 = [h_1(x_1), h_2(x_1), h_3(x_1), h_4(x_1), h_5(x_1)]$$

단계 4. 단계 3에서 얻은 모형의 예측확률 H_1 과 주어진 라벨(목표변수) y_1 을 합병(merge)하여 새로운 자료 $D'_1 = \{H_1, y_1\}$ 을 생성한다.

단계 5. 5-조각 교차 타당성에 의해 훈련용 자료와 테스트용 자료에 대해 역할을 교환하면서 위의 단계를 5번 반복 수행하여 새로운 데이터 셋 $D'_1, D'_2, \dots, D'_5$, 여기서 $D'_k = \{H_k, y_k\}, k = 1, \dots, 5$ 가 얻어진다.

[Second Stage]

단계 6. 단계 5에서 얻어진 새로운 데이터 셋 $D'_1, D'_2, \dots, D'_5$ 에서 $D'_2, \dots, D'_5$ 의 합집합 즉, $\bigcup_{k=2}^5 D'_k$ 을 훈련용 자료로 이용하고 나머지 데이터 셋, $D'_1 = \{H_1, y_1\}$ 을 테스트용 자료로 이용한다.

단계 7. 훈련용 자료 $\bigcup_{k=2}^5 D'_k$ 에 심층신경망을 적용하여 모형을 생성하고 테스트용 자료 $D'_1 = \{H_1, y_1\}$ 에 생성된 모형을 적용하여 최종 분류한다.

단계 8. 5-조각 교차 타당성에 의해 위의 단계를 5번 반복 수행하여 최종 분류한다.

본 연구에서 사용된 기본 분류모형으로는 k-근접 이웃, 의사결정나무, 서포트 벡터 머신, 판별분석, 로지스틱 회귀분석을 고려한다.

성능평가 척도

본 연구에서 제안한 딥러닝기반 앙상블 모형의 성능 평가를 위해 성능 평가 척도인 정확도(accuracy), Receiver Operating Characteristics

(ROC) 곡선과 c-통계량(c-statistics) 등을 가지고 비교하고자 한다. 제안한 딥러닝기반 앙상블 모형의 출력값은 이진 분류 값이며, 성능측도는 5-조각 교차타당성에 의해 생성된 5개의 데이터 셋에 대해 평균을 가지고 얻어진다.

정확도는 전체 데이터 중 실제 목표값과 모형의 예측값이 일치하는 정도를 나타내고 ROC 곡선은 분류모형의 (1-특이도(specificity))를 x축으로 하고 민감도(sensitivity)를 y축으로 한 그래프이다. (1-특이도)를 위양성률(false positive rate)이라 하고 민감도를 진양성률(true positive rate)이라고 한다. ROC 곡선은 분류모형의 결과 값인 사후확률(posterior probability)이 변화할 때 위양성률과 진양성률의 변화를 그래프로 나타낸 것으로 ROC 곡선이 좌측 상단으로 더 위에 위치할수록 좋은 모형이다[15]. ROC 곡선 아래의 면적을 c-통계량이라고 하는데 어떤 모형의 ROC 곡선 아래의 면적이 다른 모형의 면적보다 크면 평균적으로 더 우수한 모형이라 할 수 있다[16].

연구 결과

제안한 딥러닝기반 앙상블 모형의 성능을 평가하기 위해 기존의 단일모형으로 k-근접 이웃, 의사결정나무, 서포트 벡터 머신, 판별분석, 로지스틱 회귀분석과 심층신경망을 고려하였고, UCI 데이터베이스에 있는 WOBC 자료, WDBC 자료를 사용하여 평가하였다. 여기서 WDBC 자료는 표준화, 즉 평균을 빼주고 표준편차로 나누는 작업을 수행 후 사용하였다.

본 논문에서 딥러닝은 H2O 플랫폼을 사용하여 수행하였다. H2O 플랫폼의 성능은 은닉층의 크기와 노드 수, 반복 수 그리고 활성화 함수 등 여러 가지 모수에 의해 영향을 받는다. 여기에서는 H2O 플랫폼에 주어진 기본환경, 즉 은닉층의 수는 2개이고 각 층의 노드는 각각 200개, 200개이고 반복 수(epochs)는 10이고 활성화 함수(activation function)는 Rectifier Linear Unit (ReLU)을 사용하고 있다.

본 연구는 통계 프로그램 R (ver. 3.4.1)을 사용하고 딥러닝은 “h2o” 패키지[17]를 사용하였고, 그리고 k-근접 이웃, 서포트 벡터 머신, 의사결정나무, 판별분석, 로지스틱 회귀분석은 “caret” 패키지[18]의 각각 “knn”, “svmRadial”, “rpart”, “lda”, “glm” 패키지에서 기본옵션을 사용하였다. 여기서 “knn”은 유클리디안 거리(Euclidean distance) 하에서 k=5를 사용하고, “svmRadial”은 래디얼 커널(radial kernel)을 사용하고 WOBC 자료에서 $C=0.25$, $\sigma=0.755$, WDBC 자료에서 $C=0.25$, $\sigma=0.0464$ 그리고 “rpart”는 CART 알고리즘을 사용하여 불순도(impurity)의 척도로서 지니지수(Gini index)를 이용하여 이진분리(binary split)하였다. 참고로, 딥러닝에 많이 사용하는 R 패키지로는 h2o와 deepnet 패키지가 있다. deepnet 패키지는 제한된 볼츠만 머신(restrict-

ed Boltzmann machine), 딥 신뢰 망(deep belief network), 스택 오토코더(stacked autoencoders) 등과 같은 다양한 딥 아키텍처(deep architecture)를 제공하나 GPU 지원이 미흡하고 또한 튜닝 파라미터를 제공하고 있지 않아 방대한 데이터 처리용으로 사용하는 것을 권장하지 않는다. 반면에 h2o 패키지는 H2O 플랫폼 상에서 분산 처리가 가능하고, 높은 수준의 사용자 경험을 제공하며 또한 빠른 학습 속도와 많은 튜닝 파라미터를 제공한다.

Figure 3은 분류모형을 WOBC 자료에 적용하여 얻은 ROC 곡선들을 나타내고 있다. 여기서 kNN은 k-근접 이웃, DT는 의사결정나무, SVM은 서포트 벡터 머신, GLM은 로지스틱 회귀분석, LDA는 판별분석, DNN은 심층신경망 그리고 Ensemble은 제안한 딥러닝기반 앙상블모형을 나타낸다.

Figure 3으로부터 의사결정나무를 제외한 나머지 모형들의 ROC 곡선들은 왼쪽 위 상단에 위치하고 있어 좋은 성능을 보임을 알 수 있다. 이들 곡선들은 서로 중첩되어 있어서 가시적으로 좋고 나쁨을 구별하는 것은 쉽지 않다. 의사결정나무의 ROC 곡선은 대각 참고선(diagonal

reference line)과 가장 가까이 위치하고 있어 성능이 현저히 떨어짐을 보이고 있다.

Table 3은 WOBC 자료에 적용하여 얻은 분류모형의 정확도와 c-통계량을 보여주고 있다. 먼저, 정확도면에서 보면 제안된 딥러닝기반 앙상블모형이 가장 높은 정확도를 나타내고 있고 다음으로 심층신경망, k-근접 이웃 순으로 나타났고, 의사결정나무가 가장 낮은 정확도를 보였다. c-통계량에서 보면 딥러닝기반 앙상블모형이 가장 높고 다음으로 로지스틱 회귀분석, 판별분석, 심층신경망 순으로 나타났고, 의사결정나무가 가장 낮은 c-통계량을 보였다.

Figure 4는 분류모형을 WDBC 자료에 적용하여 얻은 ROC 곡선들을 나타내고 있다.

Figure 4는 Figure 3과 비슷한 양상을 보임을 알 수 있다.

Table 4는 WDBC 자료에 적용하여 얻은 정확도와 c-통계량을 보여주고 있다.

Table 4의 정확도면에서 보면 WOBC 자료에서처럼 제안된 앙상블모형이 가장 높고 다음으로 심층신경망, k-근접 이웃 순으로 나타났고,

Figure 3. ROC curves of classification models for WOBC data sets. ROC, receiver operating characteristics; WOBC, Wisconsin original breast cancer.

Table 3. Performance comparison of classification models for WOBC data set

Classification models	Accuracy	c-statistics
kNN	0.9639	0.9882
DT	0.8415	0.7951
SVM	0.9564	0.9843
GLM	0.9565	0.9912
LDA	0.9524	0.9906
DNN	0.9710	0.9897
Ensemble	0.9722	0.9914

WOBC, Wisconsin original breast cancer.

Figure 4. ROC curves of classification models for WDBC data sets. ROC, receiver operating characteristics; WDBC, Wisconsin diagnostic breast cancer.

Table 4. Performance comparison of classification models for WDBC data set

Classification models	Accuracy	c-statistics
kNN	0.9653	0.9856
DT	0.8322	0.7916
SVM	0.9565	0.9905
GLM	0.9454	0.9490
LDA	0.9559	0.9911
DNN	0.9808	0.9940
Ensemble	0.9823	0.9935

WDBC, Wisconsin diagnostic breast cancer.

의사결정나무가 가장 낮은 정확도를 보였다. c-통계량 면에서는 심층 신경망이 제안된 앙상블모형보다 약간 높게 나타났고, 다음으로 판별 분석, 서포트 벡터 머신 순으로 나타났고, 의사결정나무가 가장 낮은 c-통계량을 보였다.

고 찰

본 연구에서 유방암의 성능향상을 위해 제안된 딥러닝기반 앙상블 방법은 두 단계에 걸쳐 이루어지고 있다. 첫 번째 단계에서 사용된 스택킹 앙상블은 같은 종류의 모형(homogeneous models)을 앙상블한 배깅, 부스팅과는 다르게 서로 다른 모형들을 결합하여 단일 모형에 비해 성능을 향상시키는 방법이다. 다른 앙상블처럼 첫 번째 단계에서 최종 분류를 할 수 있으나 보다 정확한 유방암 분류를 위해 추가적으로 딥러닝을 적용하여 최종분류하고 있다.

현재 딥러닝을 위한 여러 가지 플랫폼들이 개발되고 있다. 대표적인 플랫폼에는 MXNet, Caffe, TensorFlow, Theano, Torch, H2O 등이 있다. 본 논문에서는 세계 최고의 오픈소스 플랫폼인 H2O 플랫폼을 사용하고 있다. H2O 플랫폼은 일반적인 개발 환경(파이썬, 자바, 스칼라, R), 빅데이터 시스템(하둡, 스파크), 그리고 다양한 데이터 소스(HDFS, S3, SQL, NoSQL)를 통해 머신러닝 알고리즘에 대한 액세스를 제공한다. 그리고 H2O 플랫폼은 데이터 수집, 모형 구축, 그리고 각종 예측을 제공하기 위한 엔드 투 엔드 솔루션으로 개발되어 있어 사용가능하다. 본 논문에서는 H2O 플랫폼의 주어진 기본환경, 즉 자료에 상관없이 주어진 모수, 즉 은닉층의 수, 은닉 노드의 수, 반복 수, 활성화 함수 등의 기본값을 사용하고 있다. 따라서 딥러닝의 성능은 H2O 플랫폼의 최적의 모수를 추정함으로써 개선시킬 수 있다.

결 론

여성의 유방암 발병률은 지속적으로 증가하고 있으며, 현대 의학 기술의 발달로 인해 조기 발견율이 높아지고 환자 삶의 질을 향상시키는 치료방법들이 개발되면서 검진의 비중 역시 늘어나고 있다. 최근 국가 암검진에서 유방암의 진단 정확도가 너무 낮다는 문제가 제기되는 가운데 의사로 하여금 진단 정확도를 높일 수 있는 보조 수단으로 통계적 분류방법이 중요시되고 있다.

딥러닝은 머신러닝의 한 분야로 영상 인식, 음성 인식 등, 다양한 분야에서 획기적인 결과를 내면서 큰 주목을 받고 있다. 본 연구에서는 유방암 분류성능 향상을 위해 딥러닝기반 앙상블 모형을 구축하고 기존의 단일 모형들과의 성능을 비교하였다.

본 연구에서 제안된 딥러닝기반 앙상블은 두 단계에 걸쳐 이루어졌

다. 첫 번째 단계에서는 형태가 다른 5가지의 기본 모형, 즉 k-근접 이웃, 서포트 벡터 머신, 의사결정나무, 판별분석, 로지스틱 회귀 모형을 가지고 스택킹 앙상블 수행한 후 두 번째 단계에서는 딥러닝을 적용하여 분류성능을 향상시켰다.

본 논문에서는 딥러닝기반 앙상블의 성능을 평가하기 위해 UCI의 유방암 자료에 대해 기존의 분류방법인 k-근접 이웃, 의사결정나무, 서포트 벡터 머신, 판별분석, 로지스틱 회귀분석과 성능평가 측도인 정확도, ROC 곡선 그리고 c-통계량을 가지고 비교분석하였다.

실제 WOBC 자료와 WDBC 자료에서 성능 실험 결과 ROC 곡선에서 의사결정나무를 제외하고는 제안된 딥러닝 앙상블을 포함한 모든 방법이 좋은 성능을 보였고, 정확도 면에서 WOBC 자료와 WDBC 자료 모두에서 제안된 앙상블모형이 가장 높고 다음으로 심층신경망, k-근접 이웃 순으로 성능을 보였다. c-통계량 면에서 WOBC 자료에서는 제안된 앙상블모형이 가장 높고 다음으로 로지스틱 회귀분석, 판별분석 순으로 나타났고, WDBC 자료에서는 심층신경망, 제안된 앙상블모형 순으로 나타났으며 의사결정나무는 가장 성능이 떨어지는 것으로 나타났다.

본 연구에서 딥러닝은 H2O 플랫폼의 주어진 기본환경하에서 수행하였다. 향후 연구에서는 주어진 자료로부터 H2O 플랫폼의 최적의 모수를 추정함으로써 탁월한 성능을 갖는 딥러닝 앙상블모형을 개발하고자 한다.

REFERENCES

1. National Cancer Center. Annual report of cancer statistics in Korea in 2015. Sejong: Ministry of Health and Welfare; 2017 (Korean).
2. Sewak M, Vaidya P, Chan CC, Duan ZH. SVM approach to breast cancer classification. Conference: Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007); 2007, p. 32-37.
3. Fiuzy M, Haddadnia J, Mollania N, Hashemian M, Hassanpour K. Cancer based on fine needle aspiration (FNA) test data and combining intelligent systems. Iran J Cancer Prev 2012;5(4):169-177.
4. Yuan Y, Shaw M. Induction of fuzzy decision trees. Fuzzy Sets System 1995;69(2):125-139.
5. Vapnik VN. The nature of statistical learning theory. New York: John Wiley & Sons; 1996.
6. Zhang GP. Neural networks for classification: A Survey, IEEE Transactions on Systems, Man and Cybernetics—Part C. Applications Reviews 2000;30(4):451-462.

7. Gupta S, Kumar D, Sharma A. Data mining classification techniques applied for breast cancer diagnosis and prognosis. *Indian J Comput Sci Eng* 2011;2(2):188-195.
8. Kitbumrungrat K. Comparison logistic regression and discriminant analysis in classification groups for breast cancer. *Int J Comput Sci Network Secur* 2012;12(5):111-115.
9. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput Methods Programs Biomed* 2018;153:1-9.
10. Lim JS, Oh YS, Lim DH. Bagging support vector machine for improving breast cancer classification. *J Health Info Stat* 2014;39(1):15-24 (Korean).
11. Salunkhe UR, Mali SN. Classifier ensemble design for imbalanced data classification: a hybrid approach. *Procedia Comput Sci* 2016;85:725-732.
12. UCI Machine Learning Repository. University of California, Center for Machine Learning and Intelligent Systems. Available at <http://archive.ics.uci.edu/ml/datasets.html> [accessed on October 13, 2017].
13. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. London, UK: Chapman & Hall/CRC; 1984.
14. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen* 1936;7:111-132.
15. Egan J. *Signal decision theory and ROC analysis*. Cambridge, MA: Academic Press; 1975.
16. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;54(1):17-23.
17. Landry M. *Machine learning with R and H2O*. Mountain View, CA: H2O.ai, Inc.; 2018.
18. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;28(5):1-26.