http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.25, No.3, August 2020, pp.1-13
https://doi.org/10.7838/jsebs.2020.25.3.001

다중 스태킹을 가진 새로운 앙상블 학습 기법

A New Ensemble Machine Learning Technique with Multiple Stacking

이수은(Su-eun Lee)*, 김한준(Han-joon Kim)**

초 록

기계학습(machine learning)이란 주어진 데이터에 대한 일반화 과정으로부터 특정 문제를 해결할 수 있는 모델(model) 생성 기술을 의미한다. 우수한 성능의 모델을 생성하기 위해서는 양질의 학습데이터와 일반화 과정을 위한 학습 알고리즘이 준비되어야 한다. 성능 개선을 위한 한 가지 방법으로서 앙상블(Ensemble) 기법은 단일 모델(single model)을 생성하기보다 다중 모델을 생성하며, 이는 배깅(Bagging), 부스팅(Boosting), 스태킹(Stacking) 학습 기법을 포함한다. 본 논문은 기존 스태킹 기법을 개선한 다중 스태킹 앙상블(Multiple Stacking Ensemble) 학습 기법을 제안한다. 다중 스태킹 앙상블 기법의 학습 구조는 딥러닝 구조와 유사하고 각 레이어가 스태킹 모델의 조합으로 구성되며 계층의 수를 증가시켜 각 계층의 오분류율을 최소화하여 성능을 개선한다. 4가지 유형의 데이터셋을 이용한 실험을 통해 제안 기법이 기존 기법에 비해 분류 성능이 우수함을 보인다.

ABSTRACT

Machine learning refers to a model generation technique that can solve specific problems from the generalization process for given data. In order to generate a high performance model, high quality training data and learning algorithms for generalization process should be prepared. As one way of improving the performance of model to be learned, the Ensemble technique generates multiple models rather than a single model, which includes bagging, boosting, and stacking learning techniques. This paper proposes a new Ensemble technique with multiple stacking that outperforms the conventional stacking technique. The learning structure of multiple stacking ensemble technique is similar to the structure of deep learning, in which each layer is composed of a combination of stacking models, and the number of layers get increased so as to minimize the misclassification rate of each layer. Through experiments using four types of datasets, we have showed that the proposed method outperforms the exiting ones.

키워드 : 자동분류, 기계학습, 앙상블, 스태킹
Classification, Ensemble, Machine Learning, Stacking

본 연구는 국토교통부 도시건축연구 사업의 연구비지원(20AUDP-B100356-06)에 의해 수행되었으며, 또한 본 연구는 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었습니다(IITP-2020-2018-0-01417).

* First Author, Master, School of Electrical and Computer Engineering, University of Seoul (tndms0224@naver.com)
** Corresponding author, School of Professor Electrical and Computer Engineering, University of Seoul(khj@uos.ac.kr)

Received: 2020-05-28, Review completed: 2020-07-08, Accepted: 2020-07-28

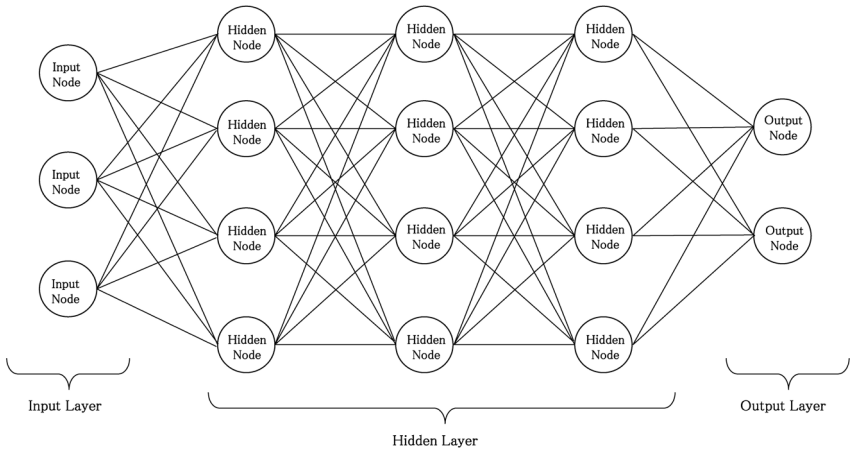
1. 서 론

기계학습 기법 중의 하나인 앙상블(Ensemble) 기법[2, 10, 22]은 다수의 약한 학습기(weak learner)를 결합하여 하나의 강한 학습기(strong learner)를 생성한다. 앙상블 기법에는 투표(voting) 방식을 이용한 배깅(Bagging) 기법[1], 가중 투표(weighted voting) 방식을 이용한 부스팅(Boosting) 기법[14], 그리고 단일 모델(single model)로부터 얻어낸 예측 값을 학습 데이터로 삼는 스태킹(Stacking) 기법[15]이 있다.

본 논문은 앙상블 학습 기법 중에서 가장 성능이 우수한 스태킹 기법을 개선하기 위해 딥러닝(deep learning) 아키텍처와 유사한 다중 스태킹 앙상블(Multiple Stacking Ensemble) 기법을 제안한다.

다중 스태킹 앙상블 기법은 <Figure 1>과 유사한 형태를 띠며, 입력층(input layer)은 단일 모델, 은닉층(hidden layer)은 각 모델의 조합으로 이루어진 스태킹 모델로 구성되고 출력층(output layer)은 최종 분류기에 해당한다. 은닉층과 출력층으로 이루어진 인공신경망(artificial neural network)[5]과 단일 모델, 메타 모델(meta model)로 분류하는 스태킹 기법이 2개의 레이어로 구성되어 있다는 점을 착안하여 은닉층의 개수가 2개 이상인 딥러닝 아키텍처와 유사하게 구성하였다. 노드간의 가중치를 학습하는 딥러닝 아키텍처와는 다르게 제안 기법의 학습 아키텍처는 단일 모델의 예측 값에 대한 가중치를 학습한다는 것이다. 관련 연구로는 스태킹 기법을 이용하여 멀티 계층 앙상블 학습 기법을 제안한 논문[4, 11, 20]이 있다. El-Khatib et al.[11]은 스태킹을 확장하여 기본 계층, 앙상블 계층, 일반화 계층의 3개의 계층으로 구성된 알고리즘을 제안하였다. 이에 대해 본 논문은 El-Khatib et al.[11]의 제안 기법보다 여러 개의 계층을 추가하여 딥러닝 구조와 유사한 형태로 생성했다는 점이 차별화 된다.

본 논문의 구성은 다음과 같다. 제2장은 본 논문에 대한 연구배경을 살펴보고 제3장은 본 연구에서 제안하는 다중 스태킹 앙상블 기법을 설명한다. 제4장은 다양한 데이터를 사용하여



〈Figure 1〉 Deep Learning Architecture

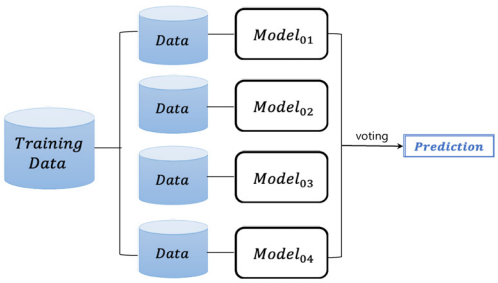
실험한 결과를 보이며 마지막으로 제5장은 결론 및 향후 연구를 서술한다.

2. 배경 지식 및 관련 연구

이 절은 본 연구에서 사용하는 다중 스태킹 기법의 기반이 되는 앙상블 기법에 대해 설명한다.

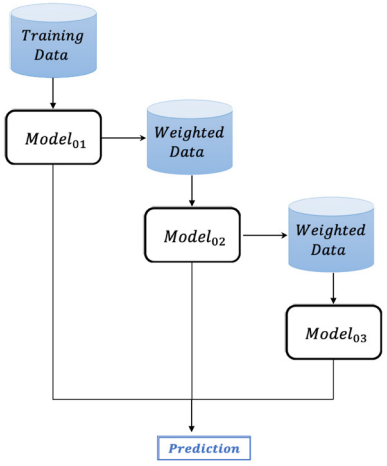
앙상블 기법은 주어진 학습 데이터로부터 다수 개의 모델을 생성하고 이들을 조합하여 개선된 모델을 생성한다. 특히 개별 모델이 50% 이상의 정확도를 가질 때 성능이 우수하다. 앙상블 기법은 배깅, 부스팅, 스태킹으로 구분된다.

배깅(Bagging)은 Bootstrap Aggregating의 약자로서, 주어진 학습 데이터로부터 임의로 복원 추출하여 동일한 크기로 다수 개의 소규모 학습 데이터를 구성한다(<Figure 2> 참조). 이에 대한 학습 과정으로부터 얻어진 다수 개의 약한 학습기를 가지고 최종 분류(또는 예측)를 위해 투표 방식을 수행한다. 이러한 배깅 과정은 개별 모델들의 분산을 크게 하여 모델의 성능을 높인다.



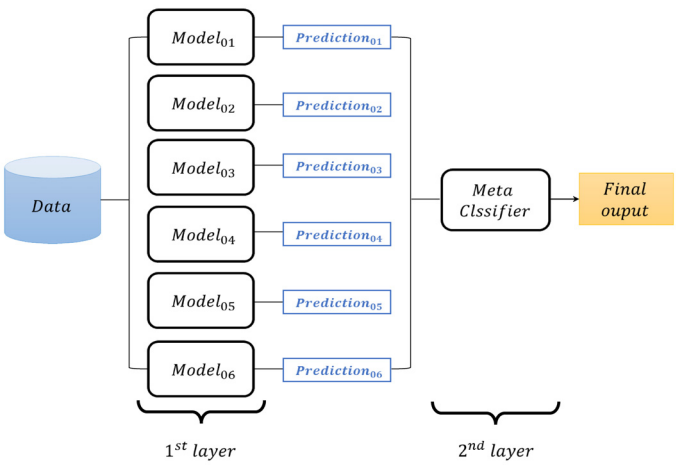
〈Figure 2〉 Bagging Architecture

부스팅 기법이란 배깅 기법과는 달리 생성된 다수 개의 모델들이 동등하지 않게 취급되며, 각 모델이 부여된 가중치가 최종 분류를 위한 투표에 반영된다(<Figure 3> 참조). 초기 단계에서 주어진 학습 데이터를 모두 학습하여 모델을 생성한 후, 생성된 모델로부터 데이터를 분류하였을 때 잘 분류된 데이터는 가중치를 적게 주고, 잘못 분류된 데이터는 가중치를 높게 줌으로써 가중치가 조절된다. 이렇게 가중치가 부여된 데이터에 대하여 임의 추출 과정을 통해 연속적으로 새로운 모델을 생성하게 된다. 결과적으로 신뢰도가 상이한 여러 모델들로 가중 투표 방식을 통해 분류 작업을 수행한다. 이러한 부스팅 과정은 편향(bias)을 조정함으로써 모델의 성능을 높인다.



〈Figure 3〉 Boosting Architecture

마지막으로 스태킹 기법은 <Figure 4>에서 보는 바와 같이, 두 개 이상의 학습 알고리즘을 이용하여 생성된 모델로부터 얻어진 예측 값 자체를 학습 데이터로 삼아 메타 모델(또는 메타 분류기)을 생성하는 기법이다. 스태킹 기법은 각 개별 모델이 독립적이라고 가정하기 때문에 이상치(outlier)에 대응력이 높아 단일 모



〈Figure 4〉 Stacking Architecture

델의 오분류율보다 작은 값을 갖게 되어 성능이 우수하다.

이에 본 논문은 개별 모델의 조합을 사용하여 여러 개의 새로운 계층을 추가한 다중 스태킹 앙상블 학습을 제안한다. 각 모델이 독립적이라는 스태킹의 가정으로부터 임의의 조합을 통해 딥러닝과 유사한 구조를 생성한다. 해당 아키텍처는 각 계층의 오분류율을 최소화함으로 기존 스태킹 기법보다 우수한 성능을 나타낼 수 있다.

3. 다중 스태킹 앙상블 기법

제2장에서 설명한 배경지식을 활용하여 본 논문의 주제인 다중 앙상블 스태킹 기법에 대해 설명한다.

3.1 다중 스태킹 기법의 개요

성능 개선 방법 중에 하나인 스태킹 기법은 입력 데이터로 학습시킨 단일 모델의 예측 결과를 학습 데이터로 하여 메타 모델을 통해 분류한다. 스태킹 기법의 구조는 <Figure 4>와 같이 단일 모델 부분으로 구성된 $1^{st} layer$와 메타 모델 부분인 $2^{nd} layer$로 총 2개의 레이어를 갖는다.

이에 본 논문은 DNN(Deep Neural Network)의 형태를 가지며 스태킹 기법으로 학습한 모델들로 각 레이어를 구축하는 다중 스태킹 앙상블 기법을 제안한다. 해당 기법은 인공 신경망과 스태킹 기법이 2개의 레이어로 구성되어 있다는 점을 이용하였다. 인공 신경망 아키텍처의 은닉층을 2개 이상 가지고 있는 DNN[7]의 형태와 유사하게 레이어 수를 증가시킴으로 분류 성능 개선을 목표로 한다. 2개의 레이어 형태인 스태킹 기법을 n개의 레이어로 증가시키고 스태킹 기법으로 학습한 모델들로 각 레이어를 구성한다. 이에 따라 n개의 레이어를 생성할 때 학습한 모델의 예측 값들을 입력 데이터로 하여 다음 레이어를 학습하는 과정을 반복한다. 앙상블 학습에서 모델의 수에 따라 성능이 개선되기 때문

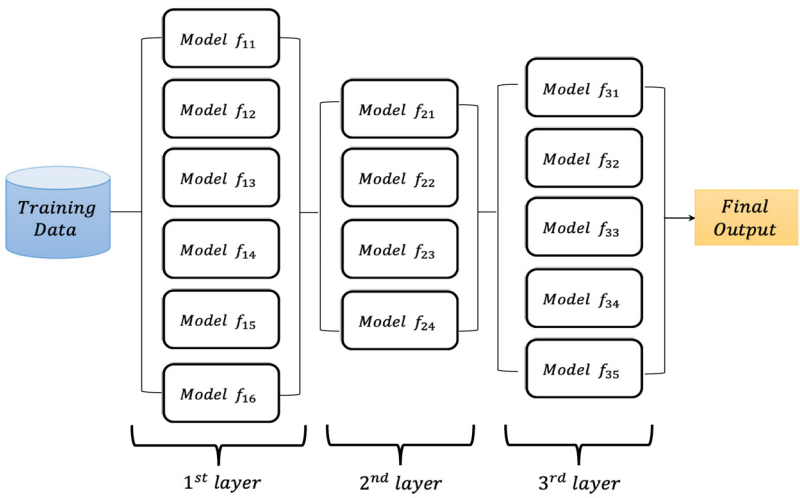
에 다양한 알고리즘을 사용하였으며 과적합(overfitting)의 가능성을 배제하기 위하여 나타낼 수 있는 최대 개수로 모델을 선택한다.

3.2 다중 스태킹 기법 생성 방법

다중 스태킹 앙상블 기법은 스태킹 구조의 계층을 확장시킨 것으로 생성 방법은 다음과 같다. m개의 알고리즘 중에서 임의로 k개의 학습 알고리즘을 선택하여 학습하면 하나의 레이어는 $\binom{m}{k}$개의 모델 조합으로 구성되고 이와 같은 레이어가 m개 생성될 수 있다. 이때, 레이어 생성 방법은 모델 $\binom{m}{k}$개의 조합 중에서 임의로 하나의 조합을 선택하여 구성하며 계층의 수 또한 사용자의 임의로 생성할 수 있다.

예를 들어, <Figure 5>와 같이 레이어의 수가 3개이고 각 레이어의 모델 수는 6개, 4개, 5개로 생성한다고 가정하자. 먼저 6개의 알고리즘을 학습하여 $1^{st} layer$를 생성한다. 이어서 $1^{st} layer$에 사용한 6개의 알고리즘 중 임의로 선택한 4개의 알고리즘을 학습하여 $2^{nd} layer$를 생성한다. 마지막으로 6개의 알고리즘 중 임의로 선택한 5개의 알고리즘을 학습하여 $3^{rd} layer$을 생성한다. 즉, m개의 학습 알고리즘 중에서 임의로 선택한 k개의 알고리즘을 이용하여 $\binom{m}{k}$개의 모델을 가지는 레이어를 생성하며, 이러한 과정을 사용자가 지정한 레이어 개수만큼 반복한다.

이때 앙상블 기법은 기존에 약한 분류기의 결합으로 강한 분류기를 생성하기 때문에 사용한 분류기에 따라서 성능이 달라진다. 예를 들어 <Figure 5>와 같은 과정을 시행하였을 때와 같이 사용자가 임의로 설정한 레이어의 수와 사용한 알고리즘의 종류 및 개수에 따라서 다중 스태킹 앙상블 기법의 성능이 달라진다. 따라서 본 논문은 특정 분류 알고리즘과 레이어 수를 제시하여 우수한 성능을 보일 수 있는 다중 스태킹 기법의 생성 방법을 제안한다(<Table 1> 참조).



〈Figure 5〉 The Process of Creating Multiple Stacking Architecture

〈Table 1〉 Pseudo Code of the Proposed Multiple Stacking Ensemble Algorithm

Algorithm1 Multiple Stacking Ensemble algorithm

Input: Training data D
Learning algorithms $C = \{c_1, \cdots, c_m\}$
Number of layers n
Output: A Multiple Stacking Ensemble classifier M
Notation f_{ji}: the ith classifier (model) in the jth layer
Step 1: Learn m classifiers in the first layer
1: Generate the classifiers $f_{11}, \cdots, f_{1m}$ learned by C with D
2: for $i \leftarrow 1$ to m do
3: D' = classification results by classifiers except for f_{1i} with D
4: Learn classifiers f_{2i} with D'
5: end for
Step 2: Learn m classifiers in the second and later layers
6: for j ← 2 to n do
7: for $i \leftarrow 1$ to m do
8: D' = classification results by classifiers except for f_{ji} with D
9: Learn classifiers $f_{(j+1)i}$ with D'
10: **end for**
11: **end for**
12: **return** M with n layers

m개의 알고리즘에서 m-1개를 선택한 후 학습하면 m개의 메타 모델로 구성된 하나의 레이어가 생성된다. 여기서, j번째 레이어에서 생성된 i번째 메타 모델을 Stacking model f_{ji}라고 표현한다. m개의 f_{ji} 중에서 다시 m-1개를 선택하여 학습하면 다음 레이어에 새로운 m개의 f_{ji}이 생성된다. 이와 같은 과정을 반복하면 하나의 레이어를 생성하는 모델 조합의 개수는 $\binom{m}{m-1}$개이고 결과적으로 m개의 f_{ji}으로 구성된 레이어가 m개 생성된다. 따라서 하나의 레이어를 구성하는 f_{ji}은 알고리즘의 수와 같으며 다중 스태킹 기법의 구조의 레이어 개수를 알고리즘 수와 동일하게 지정한다.

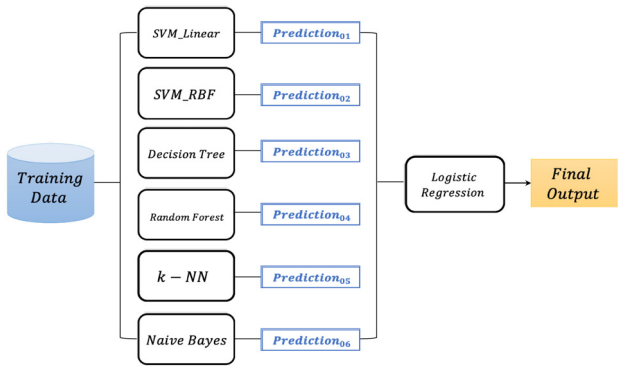
4. 실 험

4.1 실험 환경

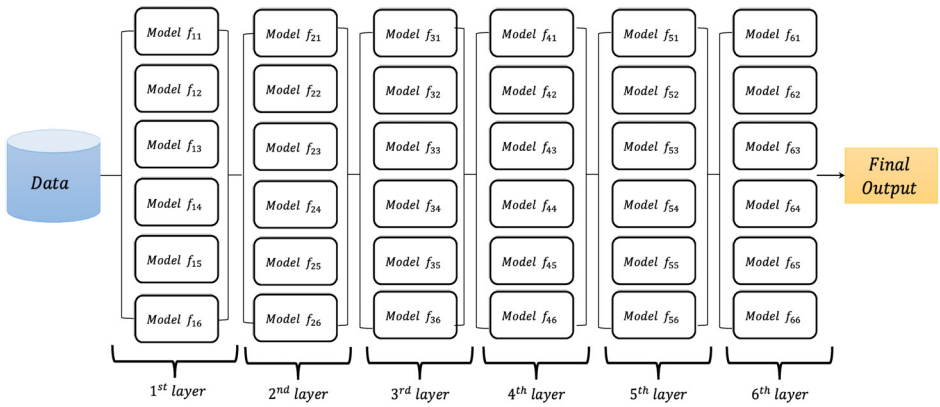
본 논문에서 제안하는 다중 스태킹 앙상블 기법의 효율성을 보이기 위하여 사용한 분류 알고리즘은 SVM(Support Vector Machine)[6], 의사결정나무(Decision Tree)[13], 랜덤포레스트(Random Forest)[3], k-최근접 이웃(k-Nearest Neighbor, k-NN)[21], 나이브 베이즈(Naive Bayes)[12]이다. SVM 알고리즘의 경우 선형분리 커널(Linear kernel)과 RBF 커널(RBF kernel)을 이용하여 실험에 사용한 분류 알고리즘은 총 6개이다. 해당 알고리즘으로 생성된 모델로부터 각 레이어 별 성능평가를 위해 로지스틱 회귀(Logistic regression) 알고리즘을 사용하였다(<Figure 6> 참조). 실험을 진행한 데이터는 1) 타이타닉 생존예측[8], 2) 성인들의 임금예측[16], 3)독일인의 신용 대출 여부 예측[18], 4) 자궁질환 암 예측[17]으로 총 4가지이며, Kaggle[9]와 UCI Repository[19]에서 참고하였다. 본 실험은 4가지 데이터 모두 동일한 조건으로 진행하였으며 모든 데이터는 이진분류(binary classification)를 목적으로 한다. 또한 주어진 데이터에서 학습 데이터(training data)와 테스트 데이터(test data)를 8:2의 비율로 나누어 실험하였으며 k-겹 교차검증법(k-fold cross validation)기법을 사용한다.

<Figure 6>과 같이 각 데이터셋에 대하여 앞서 언급한 6개의 알고리즘을 이용하여 학습을 진행하였고, 생성된 모델을 순서대로 한 가지씩 제외하는 방법을 통해 각 레이어 별로 실험하였다.

첫 번째 레이어는 제시된 6가지 알고리즘으로 각 모델을 학습하여 생성한다. 두 번째 레이어 생성 과정은 SVM(Linear kernel)을 제외한 5개의 알고리즘을 학습하고 분류된 데이터를 메타 모델을 통해 분류할 때 각 메타 모델인 Stacking model은 6가지 알고리즘으로 생성한다. 세 번째 레이어는 SVM(RBF kernel)을 제외한 5개 알고리즘을 학습하고 생성되는 데이터를 분류하는 메타 모델은 제시된 6가지 알고리즘으로 학습한다. 결과적으로 1개 알고리즘을 제외한 나머지 5가지 알고리즘으로 모델을 생성하고 Stacking model을 학습할 6개의 알고리즘을 통해 각 레이어를 구성하는 Stacking model이 생성되도록 학습한다. 이와 같은 과정을 반복하여 <Figure 7>과 같이 6개의 단일 모델로 생성되는 다중 스태킹 기법의 최대 레이어 개수는 6개로 지정한다.



〈Figure 6〉 Ensemble Structure of Multiple Stacking Technique



〈Figure 7〉 Multiple Stacking Architecture with the Maximum Number of Layers Composed of 6 Models

4.2 실험 데이터

4.2.1 타이타닉(Titanic) 데이터

첫 번째 실험에 사용한 데이터는 타이타닉 생존 분류(Titanic survived classification)로, 총 11개의 특징(feature)을 가지고 있으며 891개 데이터는 학습 데이터 712개, 테스트 데이터 179개를 포함하고 있다. 생존 여부를 예측하는 데이터 집합으로 Survived일 때 1, Not Survived일 때 0으로 분류한다.

4.2.2 성인(Adult) 데이터

성인의 소득 분류(Adult income classification) 데이터로, 소득 금액이 50K보다 큰 값을 갖게 되면 1, 작은 값은 0으로 분류하는 데이터이다. 총 14개의 특징을 가지고 있으며 48,842개의 데이터 중 결측값을 제외한 32,560개를 사용하였다. 학습 데이터 26,048개, 테스트 데이터 6,512개를 포함한다.

4.2.3 신용(Credit) 데이터

해당 데이터는 독일인의 수입에 따른 신용 여부 데이터로 대출 승인을 위해 Credit Yes는 1, No는 0으로 분류하는 데이터이다. 8개의 특징으로 구성되어 있고 1,000개의 데이터 중 학습데이터는 800개, 테스트 데이터는 200개이다.

4.2.4 자궁질환(Cervical) 데이터

자궁질환 데이터는 여성 질환인 자궁경부암 여부 데이터로 암인 경우 1, 암이 아닌 경우는 0으로 구분하여 분류하는 데이터이다. 총 36개의 특징으로 구성되어 있으나 유의미하지 않은 특징을 제외하여 실제 실험에 사용한 특징은 11개로 858개의 데이터에서 학습데이터는 687개, 테스트 데이터는 171개이다.

<Table 2>는 실험에 사용된 데이터들의 기본적인 명세를 보여준다.

〈Table 2〉 Datasets

Dataset name	Number of features	Number of records	Binary classification
Titanic Data	11	891	Survived : 1 Not survived : 0
Adult Data	11	48842	Income ≤ 50K : 1 Income > 50K : 0
Credit Data	8	1000	Credit Yes : 1 Credit No : 0
Cervical Data	8	858	Cancer : 1 No Cancer : 0

4.3 실험 결과

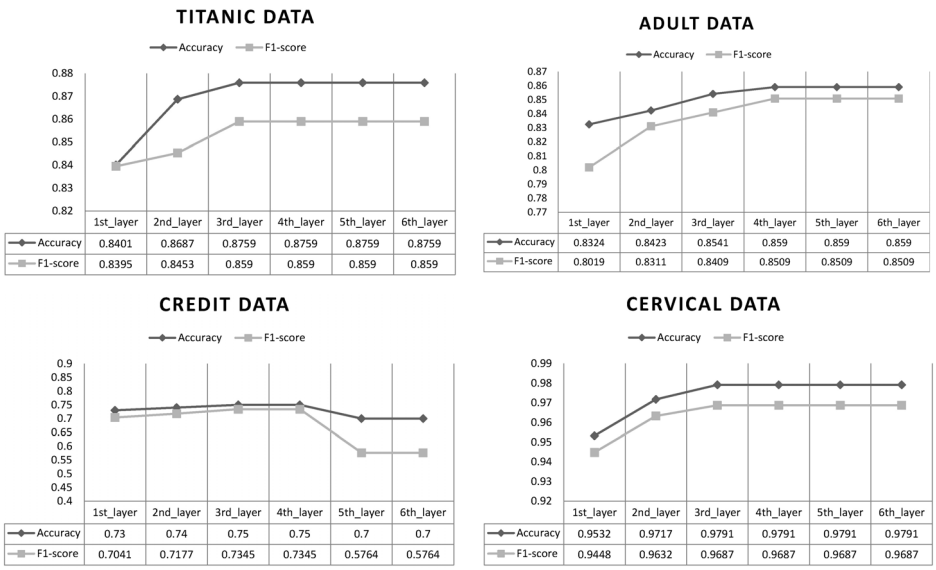
<Table 3>은 제4.2절의 데이터로 실험한 결과를 각 레이어별 성능 개선에 대해 정리한 표로써 정확도(accuracy)와 F1-score를 통해 성능 증가를 나타냈다. 타이타닉, 성인, 자궁암 데이터의 경우 첫번째 레이어($1^{st}layer$)보다 이 후 레이어의 성능이 개선되고 최종 레이어($6^{th}layer$)에 가까워질수록 더 이상 성능 변화가 일어나지 않는다는 것을 알 수 있다. 일반적으로 레이어의 수가 증가하면 각 층의 오분류율이 작아지기 때문에 성능이 개선되고 일정 레이어에 가까워질수록 성능 변화가 일어나지 않는다. 하지만 신용데이터의 경우에는 레이어가 증가함에 따라 성능이 개선되다가 일정 레이어($4^{th}layer$)를 기점으로 성능이 감소하는 것을 확인하였다. 다중 스태킹 앙상블 기법의 형태인 딥러닝 구조에서도 레이어 수가 임계점을 넘어가면 과적합의 결과를 초래하

〈Table 3〉 Classification performance in each layer

Data Performance		Titanic Data	Adult Data	Credit Data	Cervical Data
$1^{st}\,layer$	Accuracy	0.8401	0.8324	0.7300	0.9532
	F1-score	0.8395	0.8019	0.7041	0.9448
$2^{nd}\,layer$	Accuracy	0.8687	0.8423	0.7400	0.9717
	F1-score	0.8453	0.8311	0.7177	0.9632
$3^{rd}\,layer$	Accuracy	**0.8759**	0.8541	**0.7500**	**0.9791**
	F1-score	**0.8590**	0.8409	**0.7345**	**0.9687**
$4^{th}\,layer$	Accuracy	**0.8759**	**0.8590**	**0.7500**	**0.9791**
	F1-score	**0.8590**	**0.8509**	**0.7345**	**0.9687**
$5^{th}\,layer$	Accuracy	**0.8759**	**0.8590**	0.7000	**0.9791**
	F1-score	**0.8590**	**0.8509**	0.5764	**0.9687**
$6^{th}\,layer$	Accuracy	**0.8759**	**0.8590**	0.7000	**0.9791**
	F1-score	**0.8590**	**0.8509**	0.5764	**0.9687**

게 되어 오히려 성능이 좋아지지 않은 경우가 생긴다. 따라서 제안하는 기법의 구조에서도 최적의 레이어 개수를 결정하는 문제는 현재로서 실험적으로 접근하고 있으며 실험 결과를 바탕으로 최대 m개의 레이어가 생성되더라도 더 이상의 우수한 성능 변화가 나타나지 않는 최적의 레이어 개수가 존재한다고 볼 수 있다.

<Figure 8>은 <Table 3>의 내용을 시각화한 것으로 레이어 수에 따른 성능 변화를 그래프 형태로 나타내었다. 각 데이터의 최대 성능은 타이타닉 87.59%, 성인 85.41%, 신용 75%, 자궁암 97.17%이고 기존 스태킹을 적용한 첫



〈Figure 8〉 Average Classification Performance

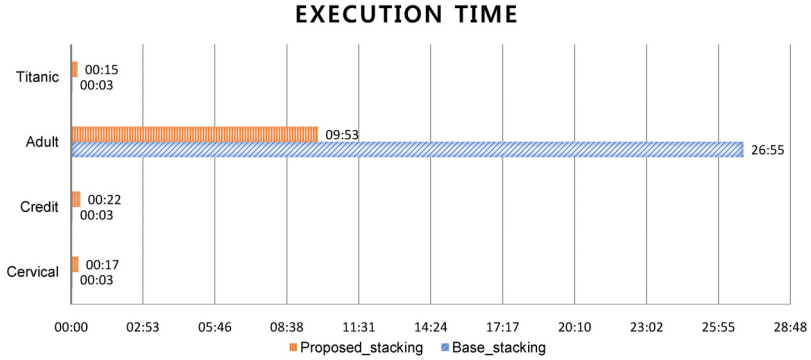
〈Table 4〉 Average Classification Performance

		Base_stacking	Proposed_stacking	Improved ratio
Titanic Data	Accuracy	83.24%	85.72%	2.97%
	F1-score	75.96%	79.98%	5.29%
Adult Data	Accuracy	82.90%	84.15%	1.51%
	F1-score	79.78%	82.41%	3.30%
Credit Data	Accuracy	72.87%	74.24%	1.88%
	F1-score	69.55%	72.51%	4.26%
Cervical Data	Accuracy	93.21%	95.96%	2.95%
	F1-score	92.11%	94.41%	2.50%

번째 레이어의 성능보다 개선된 것을 실험을 통해 확인할 수 있다.

<Table 4>는 각 데이터에 대해 기존 스태킹 기법(Base_stacking)과 제안 기법(Proposed_stacking)의 평균 성능 증가를 비교한 표이다. 기존 스태킹 기법은 다중 스태킹 앙상블 기법의 실험 환경과 동일하게 6가지 알고리즘인 SVM with linear kernel, SVM with RBF, Decision Tree, Random Forest, k-NN, Naive Bayes를 사용하였고 최종적으로 성능을 확인하기 위해 로지스틱 회귀를 통해 분류하였다. 각 실험 데이터에서 분류 성능 결과를 보면 정확도와 F1-score의 증가율이 가장 높은 데이터는 타이타닉 데이터이다. 실험 데이터들의 정확도 증가율은 최소 1.5%에서 최대 2.9%이고, F1-score는 최소 2.5%에서 최대 5.29%로 기존 스태킹 기법보다 다중 스태킹 앙상블 기법의 성능이 향상된 것을 확인할 수 있다.

다중 스태킹 앙상블 기법의 경우 기존 스태킹 기법의 확장이기 때문에 수행 시간이 오래 걸린다는 단점이 있다. 특히 데이터의 수가 가장 많은 성인 데이터의 경우 수행시간이 가장 오래 걸렸다. 이에 대한 내용을 <Figure 9>를 통해 확인할 수 있다. 이때 성인 데이터는 기존 스태킹 기법보다 다중 스태킹 앙상블 기법의 수행시간이 더 짧았는데 이에 대해 데이터 수가 많은 경우 대한 실험을 추후 연구를 통해 비교하는 실험을 진행할 예정이다.



〈Figure 9〉 The Average Execution Time of Stacking Algorithms

다중 스태킹 앙상블 기법을 이용하여 기존의 스태킹 기법보다 평균 약 2~3% 정도 성능이 개선됨을 확인했다. 또한 일정한 개수의 레이어가 생성되면 가장 우수한 성능이 나타나게 되고 이후에는 더 이상 성능이 개선되지 않거나 감소하는 것을 알 수 있다.

다중 스태킹 앙상블 기법을 이용한 실험에서 성능 증가율이 적다고 생각할 수도 있으나 정확도 측면에서 유의한 성능 개선으로 볼 수 있고 데이터의 수가 많은 경우 수행 속도의 개선이 있는 것으로 보아 데이터 수에 대한 비교 실험을 진행해 볼 수 있을 것이다. 또한 딥러닝 아키텍처와 유사하게 스태킹 기법을 응용했다는 것에 대해 의의를 둔다.

5. 결 론

성능 개선을 위한 방법인 앙상블 기법 중에 우수한 성능을 보이는 스태킹 기법을 개선한 다중 스태킹 앙상블 학습 기법에 대해서 소개하였다. 해당 기법은 각 레이어가 스태킹 모델을 구성하면서 딥러닝 구조와 유사한 형태를 띠며 레이어의 개수를 조정할 수 있다. 특히, m개의 단일모델에서 1개를 제외한 m-1개의 모델로 다음 레이어를 생성하는 과정을 반복할 때 성능이 우수하게 나타난다.

추후 연구에서 다양한 데이터에서 적용될 수 있는 다중 스태킹 앙상블 기법의 최적 레이어 개수를 찾는 것을 목표로 한다. 또한, 단순한 모델 조합에서 그치는 것이 아니라 실제 DNN 기법과 유사하게 모델의 결과 값에 할당되는 가중치를 자동으로 업데이트할 수 있도록 한다. 더 나아가 업데이트된 가중치로 결과 값의 예측(또는 분류) 성능을 개선할 수 있는 추가적인 알고리즘 개발을 통해 일반화된 기법을 기대할 수 있다.

References

[1] Breiman, L., “Bagging predictors,” Machine Learning, Vol. 24, No. 2, pp. 123-140, 1996.

[2] Brown, G., “Ensemble Learning,” Encyclopedia of Machine Learning, Vol. 312, pp. 15-19, 2010.

[3] Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Joshua, J. G., and Lawler, J. J., “Random forests for classification in ecology,” Ecology, Vol. 88, No. 11, pp. 2783-2792, 2017.

[4] Demir, N., and Dalkiliç, G., “Modi ed stacking ensemble approach to detect network intrusion,” Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 26, No. 1 pp. 418-433, 2018.

[5] El-Khatib, M. J., Abu-Naser, B. S., and Abu-Naser, S. S., “Glass classification using artificial neural network,” International Journal of Academic Pedagogical Research (IJAPR), Vol. 3, No. 2, pp. 25-31, 2019.

[6] Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H., “Comparison of linear, nonlinear, and feature selection methods for EEG signal classification,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 11, No. 2,

pp. 141-144, 2003.

[7] Goyal, M., and Rajapakse, J. C., “Deep neural network ensemble by data augmentation and bagging for skin lesion classification,” Computer Vision and Pattern Recognition, Vol. 1807.05496, 2018.

[8] Kaggle Datase, Titanic Data. https://www.kaggle.com/c/titanic/data.

[9] Kaggle DataSets, https://www.kaggle.com/datasets.

[10] Kim, Y. J., Choi, Y. L., Kim, S. L., Park, K. Y., and Park, J. H., “A study on method for user gender prediction using multi-modal smart device log data,” The Journal of Society for e-Business Studies, Vol. 21, No. 1, pp. 147-163, 2016.

[11] Pari, R., Sandhya, M., and Sankar, S., “A multitier stacked ensemble algorithm for improving classification accuracy,” Computing in Science & Engineering, Vol. 22, No. 4, pp. 74-85, 2020.

[12] Patil, T. R., and Sherekar, S. S., “Performance analysis of Naive Bayes and J48 classification algorithm for data classification,” International journal of computer science and applications, Vol. 6, No. 2, pp. 256-261, 2013.

[13] Ramamurthy, M., and Krishnamurthi, I., “Decision tree based classification type question/answer e-assessment system,” Advances in Natural and Applied Sciences, Vol. 10, No. 1 pp. 22-26, 2016.

[14] Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S., “Boosting the margin: A new explanation for the effectiveness of voting methods,” The annals of statistics, Vol. 26, No. 5 pp. 1651-1686, 1998.

[15] Syarif, I., Zaluska, E., Prugel-Bennett, A., and Wills, G., “Application of bagging, boosting and stacking to intrusion detection,” International Workshop on Machine Learning and Data Mining in Pattern Recognition, Vol. 7376, No. 8, pp. 593-602, 2012.

[16] UCI Repository Adult Data, http://archive.ics.uci.edu/ml/datasets/Adult.

[17] UCI Repository Cervical Data, http://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29.

[18] UCI Repository German Data, http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

[19] UCI Repository, http://archive.ics.uci.edu/ml/datasets.php.

[20] Yang, X., Lo, D., Xia, X., and Sun, J., “TLEL: A two-layer ensemble learning approach for just-in-time defect prediction,” Information and Software Technology, Vol. 87, No. 1 pp. 206-220, 2017.

[21] Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., “Efficient knn classification with different numbers of nearest neighbors,” IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 5, pp. 1774-1785, 2018.

[22] Zhou, Z. H., “Ensemble methods: Foundations and algorithms,” Chapman and Hall/CRC, 2012.

저 자 소 개

이 수 은 (E-mail: tndms0224@gmail.com)
2018년 성신여자대학교 통계학과 (학사)
2019년~현재 서울시립대학교 전자전기컴퓨터공학과 (석사과정)
관심분야 머신러닝, 빅데이터분석, 딥러닝

김 한 준 (E-mail: khj@uos.ac.kr)
1994년 서울대학교 계산통계학과 (공학사)
1996년 서울대학교 전산과학과 (공학석사)
2002년 서울대학교 컴퓨터공학부 (공학박사)
2002년~현재 서울시립대학교 전자전기컴퓨터공학부 정교수
관심분야 머신러닝, 빅데이터분석, 텍스트마이닝, 데이터베이스, 정보검색