JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 1: 399~406, Mar. 2019

뇌졸중 데이터를 통한 머신러닝, 딥러닝 예측 및 분류 기법 성능비교

김재호[1] · 김장영[2*]

Performance comparison of machine learning, deep learning prediction and classification techniques through stroke data

Jae-Ho Kim[1] · Jang-Young Kim[2*]

[1]Graduate Student, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea
[2*]Associate Professor, Department of Computer Science, The University of Suwon, Hwaseong, 18323 Korea

요 약

세계보건기구(WHO)에 따르면 뇌졸중은 전 세계적으로 두 번째 주요 사망 원인이며 전체 사망의 약 11%를 차지한다. 데이터 세트는 Kaggle에서 가져온 뇌졸중 데이터이며, 각 행마다 환자에 대한 관련 정보를 제공한다. 성별, 연령, 다양한 질병 및 흡연 상태등 다양한 입력 매개변수를 기반으로 환자가 뇌졸중에 걸릴 가능성이 있는지 예측하는 데 사용할 수 있다. 각 데이터 column간의 관계가 있는지 correlation계수와 다중공선성을 확인한다. 분류 기법에 있어서, 여러 가지 머신러닝기법과, DNN, Ensemble 기법을을 사용했다. DNN에서 layer을 5개 쌓아서 적용했고, Ensemble 알고리즘은 RNN과 CNN을 Ensemble하였다. optimizer함수로는 adam을 이용해 기울기를 업데이트했고, Loss Function으로는 Binary Cross Entropy를 사용했다. 평가방법으로는 Accuracy, Recal와 ROC-AUC 을 사용해서 분류기법을 평가한다.

ABSTRACT

According to the World Health Organization (WHO), stroke is the second leading cause of death globally, accounting for approximately 11% of total deaths. The dataset consists of stroke data obtained from Kaggle, with each row providing relevant information about a patient. It can be used to predict whether a patient is likely to have a stroke based on various input parameters such as gender, age, various diseases, smoking status etc. The relationships between data columns are examined using correlation coefficients and multicollinearity. For classification, various machine learning methods, DNN, and Ensemble techniques were used. In the DNN, five layers were stacked and applied. The Ensemble algorithm combines RNN and CNN. The Adam optimizer was used to update gradients, and Binary Cross Entropy was used as the Loss Function. For evaluation methods, Accuracy, Recall, ROC-AUC were used to assess the classification techniques.

키워드 : 뇌졸중, 분류, 머신러닝, 심층신경망, 앙상블 알고리즘

Keywords : Stroke, Classification, Machine learning , Deep Neural Network, Ensemble Algorithm

Received 29 January 2019, **Revised** 29 March 2019, (출판사에서작성) **Accepted** 21 April 2019

* **Corresponding Author** Jang-Young Kim(E-mail:jykim77@suwon.ac.kr, Tel:+82-31-229-8345)
Associate Professor, Department of Computer Science , The University of Suwon, Hwasung, 18323 Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.1.399 pISSN:2234-4772

Ⅰ. 서 론

뇌졸중이란 뇌의 일부분에 혈액을 공급하는 혈관이 막히거나(뇌경색), 혈관이 터짐(뇌출혈)으로써 뇌가 손상되어 나타나는 증상이다. 뇌졸중의 증상으로는 반신마비, 반신 감각 장애, 언어장애, 발음장애, 시력장애, 연하장애, 치매, 어지럼증, 의식장애, 식물인간 상태, 두통 등으로 여러 가지 전조증상이 있으며, 대한민국에서는 주로 고혈압에 의한 뇌출혈이 대부분이다[1]. 대한민국 뿐만 아니라, 세계보건기구(World Health Organization, WHO)에 따르면 뇌졸중은 암 다음으로 전 세계 사망 원인 2위이며, 전체 사망의 약 11%를 차지하며 당뇨병, 심장질환들과 여러 관계가 있으므로 주의해야 하는 질병 중 하나다. [2]

본 논문에서 사용된 데이터셋은 성별, 연령, 다양한 질병 및 흡연 상태와 같은 입력 매개변수를 기반으로 환자가 뇌졸중에 걸릴 가능성을 Classification 알고리즘을통해 예측하는데 사용될 수 있다.

본 데이터는 Kaggle의 Data files © Original Authors에서 가져왔으며 stroke 분류 데이터이다[3]. 5,010개의 데이터와 10개의 column으로 구성되어있으며, 성별, 나이, 고혈압유무, 심장질환의 유무, 결혼유무, 일 유형(children, Govt_job, Never_worked, private, self-employee4개로), 거주유형(도시와 시골), 혈액의 포도당 수준, bmi(body mass index, 체질량지수), 흡연유무, 뇌졸중 유무로 구성되어있다. 데이터 drop에서는 고유식별 id를 제거했고, 범주형데이터로 되어있던 함수는 더미함수,one hot encoding을 활용해 숫자로 표현했다. 성별에 있어서 남자를 1, 여자를0으로, 결혼유무에서 기혼자를 1, 비혼자를 0으로, work_type에서는 private를 0, self-employed를 1, Govt_job을 2, children을 3으로 분류했으며, Residence_type에서는 도시에 사는 사람을1, 시골에 사는 사람을 0으로, smoking_status에서 이전에 흡연했던 사람을 0, 흡연을 전혀 하지않은 사람을 1, 현재 흡연을 하는사람을 2로 표현했다.

모델링에 있어서 scaling은 정규분포화 하는 standard scaler을 사용했으며, data split에 있어서 train data는 0.6 valid data는 0.2, test data는 0.2로 나누었으며, 머신러닝에서는 Logistic Regression, Light GBM, CatBoost, Naive Bayes, Random Forest, Decision Tree 알고리즘을 이용한다. 딥러닝에서는 DNN을 사용했고, layer를 5개 쌓았고, 은닉층의 Activation Function으로 ReLU함수를, 출력층의 Activation Function으로는 sigmoid함수를 사용했다. optimizer함수로는 Adam을 이용해 기울기를 업데이트했고, Loss Function으로는 Binary Cross Entropy를 사용했다. Ensemble model에서는 RNN과 CNN을 결합한 모델을 사용했다. RNN부분에서 데이터를 100개씩, 임베딩 차원은 128로 변환해서 LSTM에 입력을 하였고, hidden state의 크기는 64로, 양방향으로 진행하였다. CNN부분에서도 100개의 단어를 128차원의 임베딩 벡터로 변환을 하고, 합성곱층을 Conv1d로하여, 128크기의 입력을 받고 커널의 수는 32, 커널의 크기는 5로 설정하였다. 다음으로 둘다 Max pooling을 하여, Fully Connected layer에 넘겨주고 2개의 층을 쌓았으며, activation function은 Gelu를, optimizer는 Adam을 사용했으며, drop out을 0.1로 추가하였다.

Ⅱ. 기존연구

2.1 뇌졸중 예측에 대한 통합 기계 학습 접근 방식

뇌졸중은 미국에서 사망의 세 번째 주요 원인이자 심각한 장기 장애의 주요 원인이다. 뇌졸중의 정확한 예측은 조기 개입과 치료에 매우 중요하다. 이 연구에서는 Cox 비례 위험 모델과 CHS(Cardiovascular Health Study)데이터 세트에서 뇌졸중 예측을 위한 기계 학습 접근 방식을 비교한다. 의료 데이터 세트에서 데이터 관계, 특징 선택 및 예측의 일반적인 문제를 고려한다. 제안된 Heuristic 평균을 기반으로 강력한 기능을 선택하는 새로운 자동 기능 선택 알고리즘을 제안한다. SVM(Support Vector Machines)과 결합하여 제안된 기능 선택 알고리즘은 Cox 비례 위험 모델 및 L1 정규화된 Cox 기능 선택 알고리즘과 비교하여 ROC 곡선(AUC) 아래 더 큰 영역을 달성한다. 또한, 우리는 Cox 모델보다 더 나은 일치 지수를 달성하기 위해 마진 기반 분류기의 개념과 중도 회귀 분석을 결합한 마진 기반 중도절단 회귀 알고리즘을 제시한다. 전반적으로, 우리의 접근 방식은 AUC와 일치 지수의 두 메트릭 모두에서 현재의 최첨단 기술을 능가한다. 연구 작업은 전통적인 접근 방식으로 발견되지 않은 잠재적인

위험 요소도 식별한다. 우리의 방법은 결측 데이터가 일반적이고 위험 요소가 잘 이해되지 않는 다른 질병의 임상 예측에 적용될 수 있다[3].

Ⅲ. 분석 알고리즘

3.1 DNN(심층 신경망)

심층신경망은 입력층과 출력층 사이에 n개의 은닉층들로 이루어진 인공신경망(ANN)이다. DNN은 비선형 함수들을 모델링 할 수 있다. 대표적인 신경망 작동 방법에는 Feed-Forward Network(FFD), Backpropagation, Recurrent Neural Network(RNN)이 있으며, 본 논문에서는 FFD, Backpropagation을 다룬다. FFD는 입력층, 은닉층, 출력층으로 한 방향을 쭉 이어지는 것을 의미한다. Backpropagation은 결과의 오차를 줄이기 위해 각 노드에서 다음 노드로 이어지는 가중치를 조절하는 방법이다. 이때, 가중치를 조절하기 위해 뒤로 돌아간다. forward함수를 구현할 때 Gradient Descent를 통해 loss함수를 최소화하는 신경망 파라미터를 찾아야 한다. Backpropagation은 가중치 조절을 해주며, 만약 Backpropagation이 없다면, Loss값을 학습 파라미터로 미분을 해주어야 하는데, Chain Rule에 의해서 필요한 기존 미분 값을 재활용 할 수 있고, 이로 인해 학습시간이 현저하게 줄어든다. Chain Rule은 미분을 다른 변수들의 미분의 곱으로 표현하고, DNN을 합성함수로 표현할 수 있다. Backpropagation을 통해 반복되는 미분 과정을 효율적으로 만들 수 있으며 AutoGrad와 같은 기능을 통해서 feed-forward작업에 대해 자동으로 미분을 수행 할 수 있다[4].

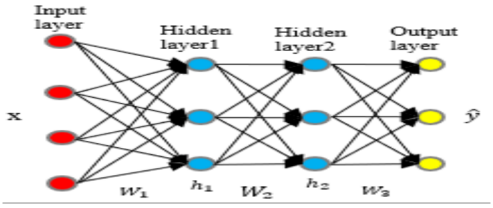

Fig1. DNN

위 그림1은 DNN과정을 보여주며 입력층, h1,과 h2는 은닉층, 출력층을 말하며, W는 가중치를 , x는 입력값, y^은 출력값을 뜻한다.

$$
\begin{aligned}
\hat{y} &= f_3 \circ f_2 \circ f_1 \\
h_1 &= f_1(x) = \delta(x \bullet w_1 + b_1) \\
h_2 &= f_2(x) = \delta(x \bullet w_2 + b_2) \\
h_3 &= f_3(x) = \delta(x \bullet w_3 + b_3) \\
\hat{y} &= f_3(h_2) = h_2 \bullet w_3 + b_3 \cdots
\end{aligned}
\qquad \text{식(1)}
$$

위 식은 심층신경망 모형의 수식이며, Chain Rule에 의해 합성함수로 표현할 수 있다. 위 은닉층(h)들이 수백층이 되면서 식은 복잡해 질수 있다[5].

3.3 Ensemble Algorithm(RNN + CNN)

신경망에 있어서 RNN(Recurrent Neural Network, 순환신경망)은 보통 순차적 데이터를 입력을 받아 텍스트 처리, 시계열 데이터 처리, 자연어처리에 많이 사용된다. 이전의 출력값을 hidden state로 받아 다음 입력값과 함께 넘겨준다. CNN(Convolutional Neural Network, 합성곱 신경망)의 경우는 보통 이미지 처리에 많이 이용되며, 특징 추출에 유리하다. 한때 자연어 처리에도 영향을 미쳤었다.(Convolutional Neural Networks for Sentence Classification) 본 연구에서는 RNN과 CNN 알고리즘을 Ensemble하여 이진분류를 하는데 이용한다. 이것은 다양한 특징을 추출하며, RNN과 CNN의 서로 다른 모델의 측면에서 학습하므로, 다양한 부분을 고려할 수 있다. 이것은 모델이 특정 데이터 패턴에 과적합되는 것을 방지하고, 더 나은 일반화 성능을 제공 할 수 있다. 또한 다양한 모델의 출력을 결합함으로써 단일 모델의 약점을 보완할 것을 기대할 수 있다. 예를들면, 시퀀스 데이터에서 중요한 패턴을 놓칠때, CNN은 이것을 보완할 수 있으며, 반대도 가능하다. 두 모델이 독립적으로 학습하기 때문에, 하나의 모델이 특정 데이터에서 에러를 범할 경우, 다른 모델이 이를 보정할 수 있다. 이는 모델의 강건성을 높여줍니다. 특징 추출이 병렬로 진행될 수 있어서 데이터셋 처리하는데도 효율적이다[6].

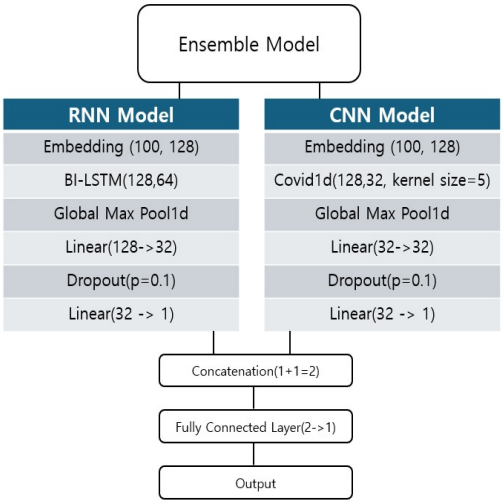

Fig2. Ensemble Model

위 그림2는 RNN과 CNN을 통한 앙상블 알고리즘을 이용한 그림이다.

IV.실험결과 및 분석

데이터 전처리는 Scaling에서는 Standard Scaling을 해주었고, train, valid와 test는 6:2:2 비율로 나누었다. Accuracy는 macro 방식을 사용했고, 또한 precision과 recall을 고려하였다.

4.1 탐험적 데이터분석(exploratory data analysis)

탐험적 데이터 분석(EDA)에 있어서 이진분류인 column이 많기 때문에 히스토그램으로 표현하지 않았고, 상관관계를 중요하게 여겼다.

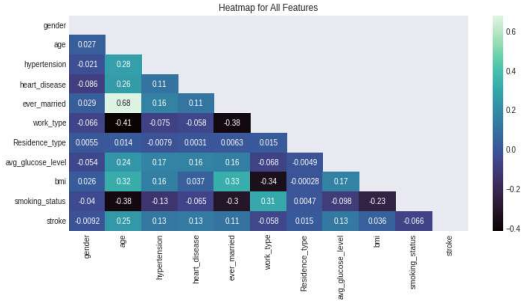

Fig3. Columns Correlation

위 그림 3은 column간에 상관관계를 보여준다.

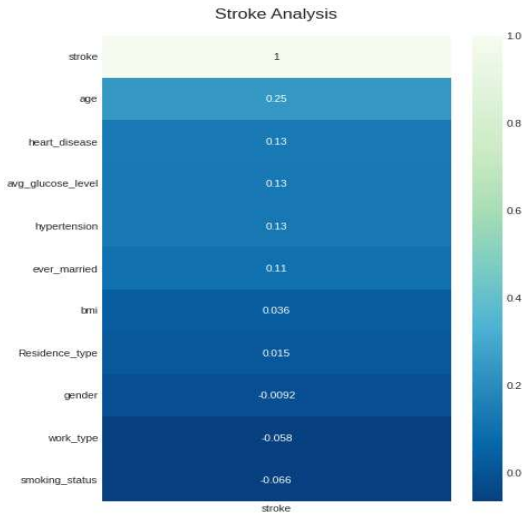

Fig4. Stroke Correlation

위 그림은4는 Stroke와 다른 column간 상관관계 사진이다. 나이와 가장 많은 상관관계를 가지는 것을 확인 할 수 있다. 상관관계가 너무 높다면, 다중공선성을 의심해야한다. 보통은 상관계수 절대값 기준이 0.8 이상인 경우, 의심해야 한다. 본 데이터는 다중공선성으로부터 안전하다고 말할수 있는 근거가 될 수 있다.

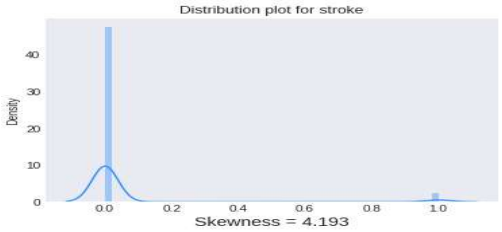

Fig.5 Stroke Data Distribution stroke

위 그림5는 stroke에 대한 분포를 보여주며, 불균형한 데이터임을 보여준다. 본 데이터에서 Accuracy는 대부분 높은 Accuracy를 가지고 있다. 하지만 불균형 데이터인 부분과 한계점에 따라 binary classification의 성능이 바뀔 수 있다는 부분을 고려해서 평가부분에서, Accuracy뿐만 아니라 Recall을 보고 판단한다. 또한 불균형 데이터이기 때문에, Recall을 이용한 방법에서 weighted평균과 macro평균을 동시에 제시한다.

4.2 다중공선성(Multicollinearity)

다중공선성은 독립변수간 상관관계가 있는지 나타내는 지표이다. 본 연구에서 사용한 데이터에서는 각 독립변수는 각 column이 된다. 만약 다중공선성이 높다면, 각 column끼리 비슷하다는 것을 의미하며, 변수간의 차이가 없음을 의미한다. 데이터에서 column을 너무 많이 분할하면 나타나는 현상으로, 문제가 된다면 변수를 삭제한다거나, PCA(주성분 분석) 또는 정규화를 이용해야한다. 상수항(const)은 독립변수 들이 함께 사용될 때 발생하는 자연스러운 결과이며, 독립변수 간의 다중공선성 문제를 반영하지않는다. 본 연구의 데이터에선 각 각 독립변수는 데이터에서 각각의 column을 의미한다. 실제로 중요한값은 독립변수의 VIF값들이다[7].

Table.1 VIF(Variance Inflation Factors)

VIF	Factor features
33.856990	const
1.016757	gender
2.415670	age
1.113385	hypertension
1.112824	heart_disease
1.955416	ever_married
1.339696	work_type
1.001191	Residence_type
1.106672	avg_glucose_level
1.242843	bmi
1.216712	smoking_status
1.087601	stroke

위 표1은 VIF(분산팽창인수)을 나타낸다. 이 값은, 다중공선성을 나타내는 척도다. VIF값이 5보다 크면 다중 공선성에 문제가 있어 회귀 계수에 대한 신뢰할 수 없는 추정치와 해석이 발생할 수 있다. 이 값이 5미만이면 다중공선성 문제가 크지 않다고 판단한다. 위에서 언급했듯이, 상수항이 높은 VIF값은 크게 신경쓰지 않아도 된다[8]. 수식을 쉽게 표현하면 다음과 같다.

$$VIF = \frac{1}{1 - 결정계수(변수설명정도))}$$ 식(2)

4.3 Accuracy와 Recall

Table.2 Accuracy

Model	Accuracy
XGB	0.91
SVC	0.91
Ridge	0.88
LR	0.89
ADA	0.92
KNN	0.89
RF	0.89
DT	0.90
BNB	0.84
GNB	0.85
DNN	0.88
RNN+CNN	0.92

위 표2는 각각 모델별 Accuracy를 보여준다. 전체적으로 높은 정확도를 보여준다. 수식을 쉽게 표현하면 다음과 같다.

$$Acuraccy = \frac{올바르게예측한전체수}{전체예측수}$$

식(3)

Table.3 Recall

Model	Recall-weight	Recall-macro
ADA	0.74	0.50
LR	0.77	0.47
XGB	0.74	0.49
RF	0.84	0.47
KNN	0.75	0.50
SVC	0.73	0.50
BNB	0.74	0.54
Ridge	0.75	0.50
DT	0.70	0.55
GNB	0.79	0.62
DNN	0.87	0.67
RNN+CNN	0.91	0.73

위 표3은 Recall값을 즉, 재현율을 보여준다. 값들은 소수점 두 번째 자리까지 나타낸다. 질병에 관해서는 Recall을 고려한다. weight평가는 클래스의 데이터 비율에 따라 가중치를 주어지는 방법이다. macro평가 는 클래스의 성능 지표를 동일한 비율로 평균을 내는 방법이다. 본 연구에서 사용한 데이터가 5:5 데이터가 아니기 때문에, macro평균을 제시한다. 수식은 다음과 같다.

$$Recall = \frac{올바르게예측한실제양성수}{전체실제양성수}$$

식(4)

Table.4 ROC-AUC

Model	ROC-AUC
ADA	0.51
LR	0.5
XGB	0.51
RF	0.5
KNN	0.5
SVC	0.5
BNB	0.54
Ridge	0.5
DT	0.54
GNB	0.62
DNN	0.81
RNN+CNN	0.83

위 표4는 ROC-AUC값을 보여준다. 데이터가 불균형하기 때문에, 그림.6 Distribution Stroke 그림에서 불균형 데이터 이기 때문에 ROC-AUC평가방법은 의미가 있다[9]. 딥러닝 모델의 결과값이 현저하게 좋음을 보여준다. 보통 0.8이 넘으면, 모델의 점수는 높다고 평가한다.[10] 수식을 쉽게 표현하면 다음과 같다.

$$ROC\ AUC \approx \frac{\text{올바르게 순위 매긴 쌍의 수}}{\text{전체 가능한 쌍의 수}} \quad \text{식(5)}$$

Ⅴ. 결론 및 향후연구

본 연구는 Kaggle 의 stroke data를 가지고와 뇌졸중 데이터를 통해서 머신러닝 분류기법과 딥러닝의 DNN, CNN+RNN의 앙상블 기법을 통해 뇌졸중을 판단한다. 텍스트처리, 자연어처리, 시퀀스 데이터, 시계열 데이터에 자주쓰이는 RNN과 이미지 특징추출에 자주 쓰이고, 가끔은 텍스트 처리에 사용되는 CNN을 앙상블 하여 이용한다. RNN과 CNN을 앙상블 하여 질병 데이터를 가지고 이진분류를 수행하면, 다양한 데이터 특성을 효과적으로 학습하고, 더 나은 일반화 성능과 예측 정확도를 얻을 수 있다. 이는 특히 복잡하고 다양한 형태의 의료 데이터를 처리하는데 매우 유용하다. 그에 대한 평가 방법으로는 Accuracy와 Recall을 사용했고, Accuracy 방면에서는 모두 좋은 결과값을 가졌지만 Recall방면에서는 앙상블 기법이 더 좋은 성능을 가졌다. 또한, ROC-AUC에서는 딥러닝 기법들이 좋은 성능을 가지는 것을 확인할 수 있다.

본 연구는 비교적 작은데이터셋과 작은 모델로 각 column의 유무별로 따져보아 뇌졸중을 판단하는데 도움을 준다. 이것은 환자의 간단한 문진표 작성을 통해 뇌졸중에 경고와 뇌졸중 검사를 장려 할 수 있다. 향후 연구로 문진표를 기반으로, 추가적인 뇌졸중 이미지 데이터를 통해 뇌졸중 자동화 판단을 할 때 근거를 제시하고, 정밀검사까지 장려할 수 있다.

마지막으로, 개인적인 신상정보, 의료데이터를 가지고 실험을 하기 때문에, 항상 윤리적인 사항에 유념을 가지고 데이터를 구성해야한다. 추후에 윤리적, 신상정보를 고려한 더 많은 데이터를 가진다면 더 높은 성능과 알고리즘을 고려할 수 있을 것이다.

REFERENCES

[1] "Stroke", *Korea association of health promotion,* vol.31, issue 1, pp. 30-31, 2007 [Internet].Available:https://www.koreascience.or.kr/article/JAKO200762680456596.page

[2] WHO and FAO announce global initiative to promote consumption of fruit and vegetables, *World Health Organization*, 2003. [Internet].Available: https://www.who.int/news/item/11-11-2003-who-and-fao-announce-global-initiative-to-promote-consumption-of-fruit-and-vegetables.

[3] Kaggle Data set, Stroke Prediction Dataset, 2021. [Internet].Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset .

[4] S. Chen, G. I. Webb, L. Liu, X. Ma, "A novel selective naïve Bayes algorithm", *Knowledge-Based Systems*, vol.192, Mar, 2020. DOI: https://doi.org/10.1016/j.knosys.2019.105361

[5] R. Hataya, J. Zdenek, K. Yoshizoe, H Nakayama, "Faster AutoAugment: Learning Augmentation Strategies using Backpropagation", *Computer Vision-ECCV 2020: 16th European Conference*, Glasgow:USA, pp. 1-16, Aug, 2020. DOI: https://doi.org/10.1007/978-3-030-58595-2_1

[6] A. Mahdaddi, S. Meshoul, M. Belguidoum, "EA-based hyperparameter optimization of hybrid deep learning models for effective drug-target interactions prediction," Expert Systems with Applications, vol.185, Dec, 2021. DOI:doi.org/10.1016/j.eswa.2021.115525.

[7] R. K. Paul, “Multicollinearity: Causes, effects and remedies”, *IASRI, New Delhi,* pp.58-65, 2006.

[8] N. Shrestha, “Detecting multicollinearity in regression analysis”, *American Journal of Applied Mathematics and Statistics*, vol.8, no.2, pp.39-42, 2020. DOI:DOI:10.12691/ajams-8-2-1

[9] B. Jason, “ROC Curves and Precision-Recall Curves,” in *Imbalanced Classification with Python Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*, Independently published., ch.7, pp. 74-84, 2020.

[10] F. S. Nahm, "Receiver operating characteristic curve: overview and practical use for clinicians." Korean journal of anesthesiology vol.75, no.1, pp.25-36, 2022. DOI: https://doi.org/10.4097/kja.21209

김재호(Jae-Ho Kim)

수원대학교 컴퓨터학부 학사
수원대학교 컴퓨터학부 석사과정
※관심분야 : 인공지능

김장영(Jang-Young Kim)

연세대학교 컴퓨터과학 공학사
Pennsylvania State Univ. 공학석사
State University of New York 공학박사
University of South Carolina 교수
수원대학교 컴퓨터학부 교수
※관심분야 : Big data, AI, Cloud computing, Networks