J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

기계 학습 기반 분석을 위한 다변량 정형 데이터 처리 및 시각화 방법: Titanic 데이터셋 적용 사례 연구☆

Multi-Variate Tabular Data Processing and Visualization Scheme for Machine Learning based Analysis: A Case Study using Titanic Dataset

성 주 형[1] 권 기 원[1] 박 경 원[1] 송 병 철[1*]
Juhyoung Sung Kiwon Kwon Kyoungwon Park Byoungchul Song

요 약

정보 통신 기술의 기하급수적인 발전에 따라 확보 가능한 데이터의 종류와 크기가 증가하고 있다. 이러한 대량의 데이터를 활용하기 위해, 통계 등 확보한 데이터를 분석하는 것이 중요하지만 다양화되고 복잡도가 증가한 데이터를 일반적인 방법으로 처리하는 것에는 명확한 한계가 있다. 한편, 연산 처리 능력 고도화 및 자동화 시스템에 대한 수요 증가에 따라 다양한 분야에 기계 학습을 적용하여 그동안 해결하지 못하였던 문제들을 풀고자 하는 시도가 증가하고 있다. 기계 학습 모델의 성능을 확보하기 위해서 모델의 입력에 사용되는 데이터를 가공하는 것과 해결하고자 하는 목적 함수에 따라 모델을 설계하는 것이 중요하다. 많은 연구를 통해 데이터의 종류 및 특성에 따라 데이터를 처리하는 방법이 제시되었으며, 그 방법에 따라 기계 학습의 성능에는 큰 차이가 나타난다. 그럼에도 불구하고, 데이터의 종류와 특성이 다양해짐에 따라 데이터 분석을 위하여 어떠한 데이터 처리 방법을 적용해야 하는지에 대한 어려움이 존재한다. 특히, 기계 학습을 이용하여 비선형적 문제를 해결하기 위해서는 다변량 데이터를 처리하는 것이 필수적이다. 본 논문에서는 다양한 형태의 변수를 포함하는 Kaggle의 Titanic 데이터셋을 이용하여 기계 학습 기반으로 데이터 분석을 수행하기 위한 다변량 정형 (tabular) 데이터 처리 방법에 대해 제시한다. 데이터 특성에 따른 통계 분석을 적용한 입력 변수 필터링, 데이터 정규화 등의 처리 방법을 제안하고, 데이터 시각화를 통해 데이터 구조를 분석한다. 마지막으로, 기계 학습 모델을 설계하고, 제안하는 다변량 데이터 처리를 적용하여 모델을 훈련시킨다. 그 이후, 훈련된 모델을 사용하여 탑승객의 생존 여부 예측 성능을 분석한다. 본 논문에서 제시하는 다변량 데이터 처리와 시각화를 적용하여 다양한 환경에서 기계 학습 기반 분석에 확장할 수 있을 것으로 기대한다.

☞ 주제어 : 기계 학습, 데이터 시각화, 데이터 처리, 정형 데이터, 통계 분석, Kaggle, Titanic dataset

ABSTRACT

As internet and communication technology (ICT) is improved exponentially, types and amount of available data also increase. Even though data analysis including statistics is significant to utilize this large amount of data, there are inevitable limits to process various and complex data in general way. Meanwhile, there are many attempts to apply machine learning (ML) in various fields to solve the problems according to the enhancement in computational performance and increase in demands for autonomous systems. Especially, data processing for the model input and designing the model to solve the objective function are critical to achieve the model performance. Data processing methods according to the type and property have been presented through many studies and the performance of ML highly varies depending on the methods. Nevertheless, there are difficulties in deciding which data processing method for data analysis since the types and characteristics of data have become more diverse. Specifically, multi-variate data processing is essential for solving non-linear problem based on ML. In this paper, we present a multi-variate tabular data processing scheme for ML-aided data analysis by using Titanic dataset from Kaggle including various kinds of data. We present the methods like input variable filtering applying statistical analysis and normalization according to the data property. In addition, we analyze the data structure using visualization. Lastly, we design an ML model and train the model by applying the proposed multi-variate data process. After that, we analyze the passenger's survival prediction performance of the trained model. We expect that the proposed multi-variate data processing and visualization can be extended to various environments for ML based analysis.

☞ keyword : data processing, data visualization, Kaggle, machine learning, statistical analysis, tabular data, Titanic dataset

[1] Smart Network Research Center, Korea Electronics Technology Institute, Seoul, 03924, Korea.

* Corresponding author (songbc@keti.re.kr)
[Received 5 July 2024, Reviewed 10 July 2024, Accepted 3 August 2024]
☆ 이 논문은 2024년도 정부(해양수산부)의 재원으로 해양수산 과학기술진흥원의 지원을 받아 수행된 연구임 (RS-2022-KS221571)

1. 서 론

하드웨어(hardware, HW)의 연산 처리 능력 향상으로 컴퓨팅 기기의 인공지능 성능은 꾸준히 발달하고 있다. 기계 학습은 인공지능 기술의 하위 분야로 데이터를 기반으로 규칙을 학습하여 추론하는 소프트웨어(software, SW) 기술을 의미한다. 비선형적인 연산에 특화된 기계 학습의 주요 특징을 이용하여 그동안 일반적인 최적화 방법으로 해결하지 못하였던 NP 난해(Non-deterministic Polynomial-time hard, NP-hard)한 문제들을 해결하고 있는 추세이다 [1, 2]. 이와 같은 흐름에 따라 다양한 분야에서 기계 학습 기법을 접목시켜 원하는 목적 함수(objective function)를 최적화시키는 연구가 활발하게 진행되고 있으며 성과가 나타나고 있다.

기계 학습 모델의 추론 성능 확보를 위해서는 모델 구조를 설계하는 것뿐만 아니라 학습 모델이 입력된 데이터의 특성(feature)을 추출하는 것이 매우 중요하다 [3]. 이를 위하여, 기계 학습 모델에 입력하는 데이터의 전처리와 학습 모델의 추론 결과를 후처리하는 데이터의 가공이 필수적이다. 다양한 연구를 통해 기계 학습을 적용하기 위해 데이터를 처리하는 방법이 제시되었다 [4-6]. 모델 입력 데이터의 처리 방법은 크게 2가지로 구분할 수 있다. 첫 번째로 수치 연산 기반으로 수행되는 기계 학습을 위하여 숫자 이외의 이미지, 문자열 등 다양한 포맷의 데이터를 수치적인 값으로 변환시키는 수치 변환 처리가 있다. 두 번째로는 수치적인 값으로 변환된 데이터를 정규화시키는 처리가 있다. 일반적인 경우, 기계 학습 모델의 입력값으로 사용되는 데이터의 종류는 매우 많고, 값의 형태나 범위 또한 그 종류에 따라 다양하다. 따라서, 입력 데이터의 특성에 따라 데이터를 처리하는 방법은 무수히 많으며, 이는 전적으로 기계 학습 모델의 설계자에 의해 결정된다. 목적 함수에 따라 조합된 모델의 입력 데이터는 하나의 벡터 또는 다차원의 행렬 형태로 표현된다. 처리 방식에 따라 가공된 입력 데이터는 동일한 목적 함수를 해결하기 위한 기계 학습 모델이더라도 추론 성능은 크게 차이가 발생한다 [7]. 그럼에도 불구하고, 사용할 수 있는 데이터가 급증함에 따라 다변량 데이터를 사용하여 기계 학습 기반으로 데이터 분석을 수행하기 위하여 공통적인 데이터 처리를 적용하는 것은 현실적인 어려움이 있다.

본 논문에서는 기계 학습을 적용하여 효율적인 데이터 분석을 수행하기 위해, 다변량 데이터의 처리 방법에 대해 다룬다. 특히, 데이터의 많은 부분을 차지하고 있는 정형(tabular) 데이터에 집중하여 기계 학습을 적용하기 위한 가공 방법에 대해 제시한다. 제안된 데이터 처리 방법을 기반으로 Kaggle의 Titanic 데이터셋 [8]에 적용하여 기계 학습 기반으로 탑승객의 생존 여부에 대한 분석을 수행한다. Kaggle의 Titanic 데이터셋은 탑승객의 정보에 대해 여러 가지 포맷의 다변량 변수로 구성되어 있고, 결측 데이터가 포함되어 있어 다양한 환경에서 접할 수 있는 데이터셋과 유사하여 본 논문에서 제안하는 기법을 적용하기에 적합하다. 기존에는 기계 학습 모델의 예측 성능을 높이기 위한 모델 구조에 대한 연구가 수행되었으나 성능 최적화를 위한 모델의 입력 데이터 처리에 대해서는 다루지 않았다 [9]. 본 논문에서는 탑승객 정보를 구성하는 다변량 데이터 처리 및 시각화를 통해 데이터의 통계적인 특성을 확인하고, 기계 학습 모델의 입력 변수를 필터링한다. 본 논문의 구성은 다음과 같다. 2절에서는 기계 학습과 관련된 배경 이론 및 그 구조에 대해 나타낸다. 3절에서는 데이터 분석을 수행하기 위하여 Titanic 데이터셋을 사용하여 개별 변수에 대한 데이터 처리를 적용하고, 시각화하여 기계 학습 모델의 입력 변수를 선정한다. 4절에서는 개별 데이터 처리를 통합한 다변량 데이터 처리를 통하여 Titanic 데이터셋 기반의 기계 학습 모델을 설계하고 생존 예측 성능을 분석한다. 5절에서는 본 논문의 내용을 요약하고 앞으로의 연구 방향에 대해 나타낸다.

2. 기계 학습 모델링 개요

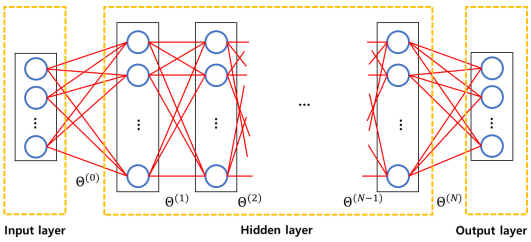
본 절에서는 기계 학습 모델링을 위한 입출력 구조 및 관련 수학식을 소개하고, 목적 함수에 따라 모델을 학습시키는 방법에 대해 나타낸다.

2.1 기계 학습 모델 입출력 구조

기계 학습 모델의 입력값을 X라 하고, 출력값을 Y라 하였을 때, X, Y 사이의 관계식은 다음과 같이 나타낼 수 있다.

$$Y = f(X;\Theta) \tag{1}$$

여기서, $f(\bullet)$는 기계 학습 모델 함수, Θ는 기계 학습 모델 연산에 사용되는 파라미터를 의미한다. X와 Y의 차원은 주어진 데이터의 가공 방법과 기계 학습 모델의 목적 함수에 따라 가변적이다. (1)에 따라 기계 학습 모델의 입력에 대한 출력을 도출하는 과정을 순전파(Forward Propagation, FP)라 한다.



(그림 1) DNN 입출력 구조
(Figure 1) DNN Architecture

일반적인 1개의 정형 데이터셋을 입력으로 사용하여 이에 대응하는 1개의 예측 결과를 출력하는 기계 학습 모델을 고려하였을 때, (1)에서 X와 Y는 각각 $M\times N_{in}$ 크기의 행렬, $M\times N_{out}$ 크기의 행렬로 표현된다. 여기서 M은 입력 데이터의 개수, N_{in}은 입력 데이터의 특성의 개수, N_{out}은 출력 데이터의 특성의 개수를 의미한다. 그림 1은 다양한 기계 학습 모델 중 비선형성 특징을 학습하는 것에 용이한 심층 신경망(Deep Neural Network, DNN) 구조를 나타낸 것이다. 이와 같은 DNN 구조를 사용하는 경우, (1)을 풀어서 나타내면 다음과 같다.

$$Y=f^{(N)}(f^{(N-1)}(\cdots f^{(0)}(X;\Theta^{(0)})\cdots;\Theta^{(N-1)});\Theta^{(N)}) \quad (2)$$

DNN의 입력층 (input layer)과 출력층 (output layer)의 차원은 각각 N_{in}과 N_{out}이 되며, N은 DNN의 은닉층 (hidden layer)의 개수를 의미한다. 입력이 X_k 일 때, k번째 layer에서 연산은 다음과 같다.

$$f^{(k)}(X_k;\Theta^{(k)})=A^{(k)}(X_kW^{(k)}+b^{(k)}) \quad (3)$$

(3)에서 $\Theta^{(k)}$를 구성하는 $W^{(k)}$와 $b^{(k)}$는 각각 가중치 (weight)와 바이어스(bias)를 의미하며, $A^{(k)}$는 활성 함수(activation function)를 의미한다.

2.2 기계 학습 모델 학습 방법

2.1에서 다룬 기계 학습 모델을 학습 하기 위해서는 학습 목적에 따른 목적 함수(objective function)의 설계가 선행되어야 한다. 목표로 하는 값 (Y_{target})과 모델을 통한 예측값의 차이를 손실 함수(loss function)라 하며, 손실 함수와 목적 함수를 각각 $L(\Theta)$, $G(\Theta)$라 할 때, 손실 함수와 목적 함수는 아래 수식과 같이 표현된다.

$$L(\Theta)=f(X;\Theta)-Y_{target} \quad (4\text{-a})$$

$$G(\Theta)=\text{minimize } E[\mid L(\Theta)\mid^{n}] \quad (4\text{-b})$$

(4-b)에서 $n>0$은 손실 함수를 최소화시키는 과정에서 차이에 대한 기울기를 얼마나 반영할 것인지에 대해 설계자에 의해 결정된다. n의 값이 커질수록 손실 함수 값이 큰 경우에 대해 수렴하는 속도가 빨라지지만 손실 함수 값이 작은 경우 국소 최적(local optima)에 빠지는 경우가 발생할 수 있다. 반대로 n의 값이 작아질수록, 손실 함수 값이 작은 경우 local optima에 빠질 가능성이 낮아지지만 손실 함수 값이 큰 경우에 대해 수렴 속도가 느려질 수 있다. 따라서, 경우에 따라, 손실 함수 값에 따라 n의 값을 가변시켜 학습시킬 수도 있다. 학습이 진행되면서 Θ는 목적 함수를 최적화시키도록 업데이트되며, 이를 위해 출력층 통해 입력층으로 역산하는 역전파(Back Propagation, BP)를 이용한 경사 하강법(gradient descent) 기반으로 학습이 이루어진다. 더욱 세부적이고 다양한 학습 방법에 대해서는 다른 연구를 참고한다. [10-12]

3. Kaggle Titanic 데이터 셋을 이용한 개별 데이터 처리 및 시각화

본 절에서는 기계 학습 기반 분석을 수행하기 위한 결측 데이터 처리, 상관관계 분석을 통한 데이터 필터링 등 데이터 처리 방법을 제안하고, 시각화한 결과를 나타낸다. 이를 위해, Kaggle의 Titanic 데이터셋 [8]을 사용한다.

3.1 Titanic 데이터셋 구조 및 기계 학습 모델 목적 함수

Titanic 데이터셋은 1912년 Titanic 호의 침몰 사고가 발생했을 당시 실제 탑승자 중 891명에 대한 정보와 생존 여부가 CSV (Comma-Separated Values) 형식의 정형 데이터로 구성되어 있다 [8]. CSV를 구성하는 열(column) 데이터는 표 1과 같은 다변량 변수로 구성된다. 어떠한 승객에 대해서는 결측치가 있는 경우도 존재한다. Titanic 호의 사고에서 역사적으로 알려진 사실로는 여성 탑승객이 남성 탑승객보다 생존율이 높았다는 것이고, 1등석의 탑승객의 생존율이 타 등급 탑승객 대비 높았다는 것이

(표 1) Kaggle의 Titanic 데이터셋 변수 구조
(Table 1) Variable structure of Kaggle Titanic Dataset

변수명	정의	데이터 포맷
PassengerId	승객 ID	Int
Survived	생존 여부	Int
Pclass	탑승 클래스	String
Name	승객 이름	String
Sex	성별	String
Age	나이	Float
SibSp	탑승객 중 배우자 +형제 숫자	Int
Parch	탑승객 중 부모 +자녀 숫자	Int
Ticket	탑승권 번호	String
Fare	탑승권 가격	Float
Cabin	객실 번호	String
Embarked	출항지	String

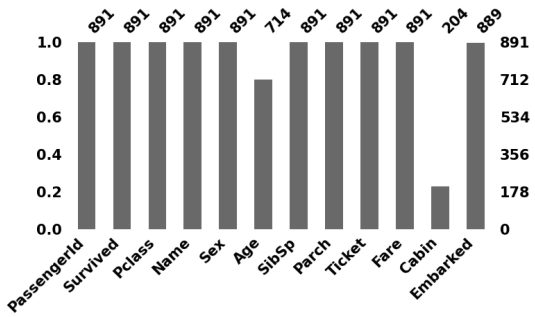
다. 그러나 위와 같이 단순한 성별과 좌석 등급 등 단일 변량 변수에 따라 해당 탑승객의 생존 여부를 예측하는 것은 정확도가 떨어진다. 기존에는 다변량 변수를 이용하기 위해, 여러 가지 단일 변량 값에 따른 세부 조건을 조합하여 예측을 시도하였으나 다변량 변수가 조합되는 경우 비선형적인 특성을 반영하기 어렵기 때문에 유의미한 효과를 획득하는 것이 제한된다. 따라서, 본 논문에서는 다변량 데이터를 활용하여 기계 학습 모델을 설계하여 생존 여부의 예측 정확도를 향상시키고자 한다. 이를 위한 기계 학습 모델의 목적 함수를 나타내면 다음과 같다.

$$\text{minimize} \ \frac{1}{M}\sum_{j=1}^{M} \mid f(x^{(j)};\Theta) - y_{target}^{(j)} \mid \quad (5)$$

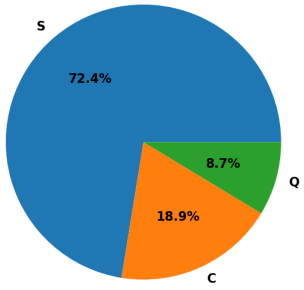
(5)에서 M은 학습 데이터의 개수를 의미하고, $x^{(j)}$와 $y^{(j)_{target}}$는 각각 $1\times N_{in}$ 크기의 입력 데이터, 0 또는 1의 탑승객의 생존 여부를 나타내기 위한 이진 변수를 나타낸다.

3.2 Titanic 데이터셋 결측치 분석 및 처리

Titanic 데이터셋에는 결측 데이터가 포함되어 있고, 기계 학습 모델을 통해 처리하기 위해 이를 처리하는 것이 필요하다. 891개의 Titanic 데이터셋의 각 변수에 대해 결측 데이터를 시각화하면 그림 2와 같다. 그림 2에서 확



(그림 2) Titanic 데이터셋 구성 변수에 따른 결측치 시각화
(Figure 2) Visualization of missing variables in Titanic dataset



(그림 3) Titanic 데이터셋에서 승객 별 출항지 비율
(Figure 3) Passenger ratio of embarkation port in Titanic dataset

인할 수 있듯이 탑승객의 나이(Age), 객실 번호(Cabin), 출항지(Embarked)에서 결측치가 존재하는 것을 알 수 있고, 객실 번호에서는 결측치 발생 비율이 77%를 초과하므로 해당 변수는 결측치를 보상하는 것보다 제외하는 편이 합리적이다. 한편, 탑승객의 나이에서 결측치가 발생한 경우, 결측치 발생 비율이 상대적으로 낮기 때문에 성별에 따른 탑승객 나이의 중앙값으로 결측치를 보상한다. 그림 3은 Titanic 데이터셋에서 3가지의 출항지에 대한 탑승객의 비율을 나타낸다. 2개의 결측치가 발생한 출항지 정보를 보상하기 위해, 전체 데이터셋에서의 출항지가 'S'로 72% 이상 편향되어 있으므로 결측치를 'S'로 처리한다. 한편, 탑승권 번호를 의미하는 'Ticket' 변수에 대해서는 문자열에 해당하는 값이 탑승권의 등급과 연관되어 있고, 동일한 값을 갖는 경우 동반자를 나타내는 변수인 'SibSp', 'Parch'와 유사한 효과를 나타내기 때문에 다른 변수와 중복되는 효과가 있어 기계 학습 모델의 학습 과정에서 이를 배제하기 위해 'Ticket' 변수를 제외한다.

3.3 변수 유형에 따른 상관관계 분석 및 기계 학습 모델 입력 변수 선별

표 1에 속한 변수에는 명백하게 승객 ID(PassengerId)와 같이 생존 여부에 영향을 주지 않는 데이터가 존재하고, 성별(Sex)과 같이 생존 여부에 영향을 주는 데이터도 존재한다. 따라서, 본 절에서는 생존 여부를 나타내는 Survived 변수를 종속 변수로 하여 Titanic 데이터셋을 구성하는 표 1에 속한 변수가 Survived 변수에 얼마나 유의미한 영향을 미치는지 변수 유형에 따라 다른 상관관계 분석 방법을 적용하여 확인한다. 그 이후, Survived에 영향을 주지 않는 변수는 기계 학습 모델의 입력 변수에서 제외하여 모델이 과적합(overfitting)되지 않도록 한다. 이를 위해, Titanic 데이터셋을 구성하는 모든 변수에 대해 생존 여부에 영향을 주는 정도를 확인하기 위해 통계적 접근을 이용한다. Survived 변수는 탑승객이 생존하였을 경우 1, 생존하지 않았을 경우 0으로 나타내는 범주형(categorical) 변수이다. 이와 유사하게 데이터 포맷이 Int 또는 String으로 되어 있는 변수는 마찬가지로 범주형의 변수로 볼 수 있다. 한편, 승객의 나이(Age)와 탑승권 가격(Fare)과 같이 Float 타입의 변수의 경우 연속(continuous) 변수로 분류된다. 변수의 유형에 따라 Survived 변수에 영향을 미치는 정도를 분석하기 위해 아래와 같은 지표를 사용한다.

- 범주형 변수: χ^2 (chi-square) 검정 [13]
- 연속 변수: 점이진 상관 계수 (point-biserial correlation) [14]

위의 지표를 통해, 2개 변수 사이의 유의 확률(p-value)을 구할 수 있다. p-value는 2개의 변수 A, B에 대해 'A는 B에 영향을 미치지 않는다'는 귀무 가설(null hypothesis)을 기각하기 위한 확률이다. 따라서, p-value 값이 작을수록 두 변수는 종속 관계에 있고, p-value 값이 클 수록 두 변수는 독립 관계에 있다는 것을 의미한다. 본 절에서는 Titanic 데이터셋을 구성하는 변수 A와 Survived 사이의 p-value를 계산하고, 'A는 Survived에 영향을 미치지 않는다'는 귀무 가설을 기각하기 위한 p-value의 기준을 0.05로 설정한다. p-value가 0.05를 초과하는 변수는 생존 여부에 유의미한 영향을 주지 않는 독립 관계에 있다고 판단하여 기계 학습 모델의 입력 변수에서 제외한다.

(표 2) χ^2 검정에 따른 p-value 결과

(Table 2) P-value results according to the χ^2 test

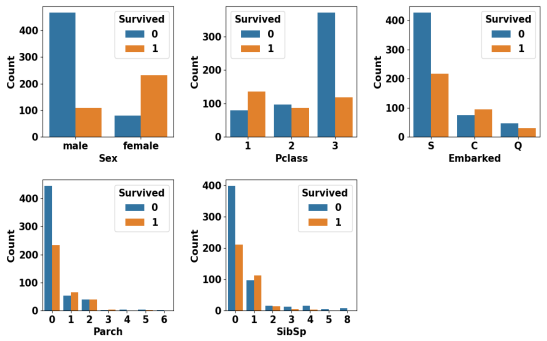
변수명	p-value
PassengerId	0.484248
Pclass	4.549e-23
Name	0.484248
Sex	1.197e-58
SibSp	1.558e-6
Parch	9.703e-5
Embarked	8.294e-7

3.3.1 χ^2 검정을 통한 범주형 변수 상관관계 분석 및 모델 입력 변수 선별

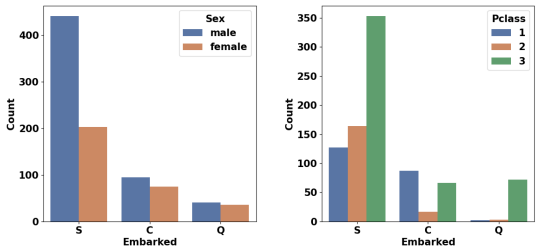
χ^2 검정은 분석을 위한 2개의 변수가 모두 범주형 변수일 때 사용된다 [13]. 이를 통해, Titanic 데이터셋에 속한 범주형 변수가 탑승객의 생존 여부에 유의미한 영향이 있는지 여부를 알 수 있고, 검정을 위한 χ^2 값은 다음과 같이 계산된다.

$$\chi^2 = \sum_{i,j} \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \tag{6}$$

(6)에서 i, j는 각각 생존 여부와 상관도 분석을 위한 범주형 변수의 index, 생존 여부를 나타내는 0 또는 1의 이진 변수를 의미한다. E_{ij}는 두 변수가 독립이라고 가정하였을 경우의 기대 빈도, O_{ij}는 실제 Titanic 데이터셋에서의 관측 빈도를 의미한다. Survived 변수는 2개의 범주만 갖기 때문에, χ^2 검정을 위한 자유도(Degree of Freedom, DoF)는 $|i|-1$이 되며, $|i|$는 상관도 분석 대상 범주형 변수가 가질 수 있는 값의 범위의 총 개수를 의미한다. 3.2절에서 나타낸 방법으로 결측치를 처리한 이후, 표 1의 범주형 변수에 대해 Survived 변수와 χ^2 검정을 수행하여 p-value를 계산한 결과는 표 2와 같다. 성별(Sex)과 좌석 등급(Pclass)이 생존율에 높은 연관성이 있었다는 역사적 사실은 표 2의 매우 작은 p-value를 통해서도 확인 가능하다. 한편, 예상한 것처럼 승객의 ID(PassengerId)와 이름(Name)은 생존 여부와 독립 관계에 있다는 것을 알 수 있고, 이를 통해 기계 학습 모델이 승객의 생존 여부를 예측하는 하기 위해 해당 변수를 사용할 필요가 없다는 결론을 내릴 수 있다. 한편, 동반자가 있는 경우와 출항지 또한 생존 여부와 밀접하게 연관되어 있는 것을 알 수 있다.



(그림 4) 범주형 변수에 따른 생존율 분석
(Figure 4) Analysis of survival ratio according to the categorical variables

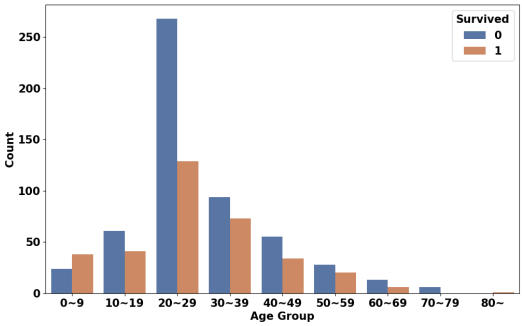


(그림 5) 출항지에 따른 성별 및 탑승권 등급
(Figure 5) Gender and ticket class according to the embarkation ports

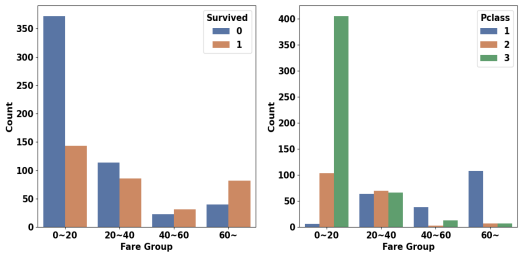
표 2의 결과는 범주형 변수에 대한 생존 여부의 연관 정도와 관련된 내용만 확인 가능하기 때문에 더욱 세부적인 분석을 위하여 그림 4와 같이 생존 여부에 영향을 주는 변수를 구성하는 항목에 따른 생존 여부를 시각화 한다. 그림 4를 통해 역사적 사실인 여성 승객이 남성 승객보다 생존율이 높고, 탑승권 등급이 높은 승객의 생존율이 낮은 등급 승객보다 높다는 것을 알 수 있다. 추가로, Titanic 호에 동반자가 있는 경우(Parch, SibSp), 탑승객의 생존율이 더 높다는 사실을 같이 확인할 수 있다. 한가지 눈에 띄는 점은 생존율에 큰 영향을 주지 않을 것 같은 출항지(Embarked) 변수에 대해서도 출항지에 따라 생존율이 유의미하게 차이가 난 것을 확인할 수 있는데 이를 출항지별 성별에 의한 영향 또는 탑승권 등급에 의한 영향이라 가정하고 그림 5에서 출항지에 따른 성별과 탑승권 등급을 비교한다. 그림 5의 결과로부터 출항지가 'C'인 경우, 생존율이 유의미하게 높게 나타난 이유는 탑승권의 등급이 높은 승객의 출항지가 'C'인 경우가 많기 때문인 것을 확인할 수 있다. 따라서, 본 논문에서는 기계 학습 모델에서 중복되는 파라미터를 배제하여 학습 모델의 불필요한 연산 감소 및 모델 성능을 향상시키기 위해 [3], Embarked 변수는 모델 입력 파라미터에서 제외하고 나머지 변수를 입력 데이터로 사용한다.

(표 3) 연속 변수와 생존 여부 변수(Survived) 사이의 점이진 상관 계수 및 p-value
(Table 3) Point-biserial correlation coefficient and p-value between continuous and 'Survived' variables

변수명	점이진 상관 계수	p-value
Age	-0.0732	0.0286
Fare	0.2573	6.12e-15



(그림 6) 승객 나이 구간에 따른 생존율
(Figure 6) Survival ratio according to the age group of the passengers



(그림 7) 탑승권 가격 구간에 따른 생존율 및 좌석 등급
(Figure 7) Survival ratio and passenger classes according to the fare group

3.3.2 점이진 상관 계수를 통한 연속형 변수 상관관계 분석 및 모델 입력 변수 선별

점이진 상관 계수는 2개의 변수 중 1개는 연속 데이터이고, 1개는 2개의 카테고리를 갖는 이진 변수일 때 사용

한다 [14]. 점이진 상관 계수의 값의 범위는 -1에서 1 사이이며, 음의 값을 갖는 경우에는 음의 상관 관계, 양의 값을 갖는 경우에는 양의 상관 관계가 있다는 것을 의미한다. 절대값이 1에 가까울수록 두 변수는 높은 상관관계를 갖고, 절대값이 0에 가까울수록 두 변수는 독립 관계에 있다. 연속 변수 X에 대하여 점이진 상관 계수 r은 다음 수식과 같이 계산된다.

$$r = \frac{\mu_1 - \mu_0}{\sigma_X}\sqrt{p_0 p_1} \tag{7}$$

(7)에서 σ_X는 X의 표준편차를 의미하고 μ_0, μ_1 은 각각 범주 0과 범주 1에 속하는 X의 평균을 의미한다. p_0, p_1은 각각 관측 결과 기반으로 X가 범주 0과 범주 1에 속해 있는 확률을 의미한다. 추가로, 3.3.1절에서 다룬 χ^2 검정과 유사한 방법으로 p-value를 통하여

서로 다른 변수 사이에 유의미한 상관관계가 있는지 여부를 판단할 수 있다. p-value를 구하기 위한 자유도는 (X의 표본의 개수-2) 이며, 연속 변수인 나이(Age)와 티켓 가격(Fare) 변수에 대해 Survived 변수와의 점이진 상관 계수와 p-value 연산 결과는 표 3과 같다. 세부적으로는 Age와 Survived 사이에는 음의 상관 관계가, Fare와 Survived 사이에는 양의 상관 관계가 있다는 것을 확인할 수 있다. 세부적인 분석을 위해 탑승객의 나이 구간을 10살로 나누어서 생존 여부를 분석한 결과는 그림 6과 같다. 그림 6으로부터 10세 미만의 승객의 생존율이 다른 연령대의 승객 대비 유의미하게 높다는 것을 확인할 수 있다. 그림 7은 탑승권 가격 구간에 따른 생존율과 탑승권의 등급을 나타낸다. 그림 7의 결과로부터 탑승권 가격 구간이 높을수록 생존율이 높은 이유는 탑승권 가격이 높을수록 높은 등급의 탑승권을 소유한 승객의 비율이 높았기 때문이라고 추정할 수 있다. Age와 Fare 변수는 모두 생존 여부와 유의미한 상관관계를 갖기 때문에 본 논문에서는 기계 학습 모델의 입력 데이터로 사용한다.

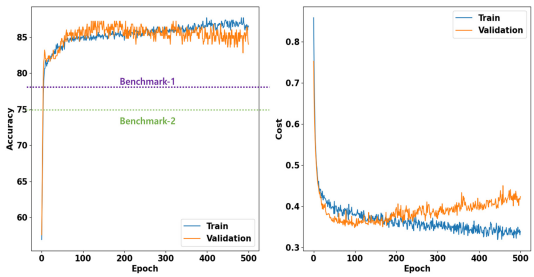
4. 다변량 데이터 처리 기반 Titanic 생존 예측 기계 학습 모델 설계 및 시험 결과

본 절에서는 3절에서 나타낸 가공된 개별의 데이터를 통합한 다변량 데이터 처리를 바탕으로 생존 여부를 예측하는 기계 학습 모델 설계하고, 훈련된 모델의 성능을 분석한다. 3절에서 기계 학습 모델의 입력으로 사용하는 변수로는 Pclass, Sex, Age, Parch, SibSp, Fare로 범주형 데이터와 수치형 데이터가 혼합된 다변량 변수 구조이다. 이 중, Pclass, Sex는 범주형의 변수이므로 one-hot 인코딩을 적용한다. 성별을 의미하는 Sex 변수는 특별한 경우로 2개의 범주만 있으므로 성별에 따라 0 또는 1로 정수형 변환을 처리해도 무방하다. 탑승권 등급을 나타내는 Pclass 변수의 경우, 1등급은 $[0,0,1]^T$, 2등급은 $[0,1,0]^T$, 3등급은 $[1,0,0]^T$ 으로 one-hot 인코딩을 적용한다 [15].

(표 4) 생존 여부 예측을 위한 심층 신경망 hyper-parameter 목록

(Table 4) Hyper-parameter list of DNN for survival prediction

파라미터 명	값
입력층(input layer) 뉴런 개수	8
은닉층(hidden layer) 개수	2
은닉층 뉴런 개수	64×8
훈련 Epoch	500
훈련 배치 크기	50
학습률	0.001
손실 함수	Cross Entropy



(그림 8) 학습 epoch에 따른 생존 예측 모델 정확도 및 손실 함수 평균 값

(Figure 8) Survival prediction accuracy and the average loss function according to the learning epochs

한편, 수치 데이터는 3절에서 가공한 값을 그대로 사용하여도 기계 학습 모델을 훈련 시킬 수 있으나 학습 안정성을 위하여 추가적인 정규화를 수행한다. 수치 데이터의 정규화 방법에는 여러 가지가 있으나 본 논문에서는 변수의 분포를 고려하여 평균값 기반 정규화와 로그 함수 기반 정규화를 적용한다. Age와 Fare 변수는 표준편차 값이 각각 13.02, 49.69로 매우 크기 때문에 3절에서 가공한 데이터를 추가적으로 평균값으로 나누어주는 평균값

기반의 정규화를 수행한다. 한편, Parch와 SibSp 변수는 표준편차 값이 각각 0.80, 1.10으로 작은 값을 갖지만 값의 크기를 작게 하여 역전파를 통한 훈련 과정에서 안정성을 확보하기 위해 (변수 + 1) 값에 로그를 취하여 정규화를 수행한다.

위와 같이 범주형 변수와 수치형 변수를 통합한 다변량 데이터 처리를 적용하면 기계 학습 모델에 입력으로 사용하기 위한 feature의 개수는 8개가 된다. 표 4는 Titanic 데이터셋을 이용한 다변량 데이터 처리를 바탕으로 생존 여부 예측을 위해 설계한 심층 신경망의 주요 hyper-parameter를 나타낸다.

심층 신경망 모델을 훈련할 때, 891개의 Titanic 데이터셋 중 전체의 70% 정도 수준의 데이터인 641개의 데이터를 사용하여 모델을 훈련시키고, 나머지 250개의 데이터를 이용하여 훈련 모델의 성능을 검증한다. 추가로, 훈련된 심층 신경망 모델의 예측 성능 비교를 위하여 다음 2가지의 벤치마크 케이스를 고려한다.

- Case 1: 여성 승객만 모두 생존
- Case 2: 여성 또는 1등급 탑승권 승객만 모두 생존

그림 8은 매 학습 epoch 마다 모델의 생존 여부 예측 정확도와 손실 함수의 평균값, 벤치마크 케이스 별 생존 예측 정확도를 나타낸다. 훈련이 진행됨에 따라 학습 데이터에 대한 정확도가 증가하는 것을 확인할 수 있고, 검증 데이터에 대한 정확도는 증가하다가 특정 시점에서 성능이 포화되는 것을 알 수 있다. 한편, 훈련된 모델의 예측 정확도는 벤치마크 성능 대비 5%p 이상 개선된 것을 확인할 수 있는데 심층 신경망 모델이 주어진 데이터셋에서 이용 가능한 다변량 변수를 모두 활용하여 비선형적인 특성을 잘 학습하였기 때문이라고 결론 낼 수 있다.

5. 결 론

본 논문에서는 기계 학습 기반 분석을 위한 다변량 정형 데이터를 처리하는 방법에 대해 제시하고, 학습 모델에 대한 성능을 분석하였다. 이를 위해, Kaggle에서 제공하는 Titanic 데이터셋을 사용하여 주어진 정형 데이터의 형태를 분석하고, 각각의 데이터 특성에 따른 결측치 처리, 통계를 이용한 모델 입력 변수 필터링 방법을 제시하고, 결과를 시각화하였다. 개별 변수에 대해 처리한 결과를 통합한 다변량 데이터를 사용하여 기계 학습 모델의 입력으로 사용하기 위해, 범주형 변수와 수치형 변수에 대한 추가적인 데이터 처리를 수행하였다. 다변량 데이터처리를 적용한 결과를 입력으로 하여 승객의 생존 여부를 예측하는 기계 학습 모델을 설계하고 모델을 훈련시켜 성능을 분석하였다. 승객의 성별 또는 탑승권 등급 등 제한된 변수만 사용하여 예측하는 벤치마크 성능 대비 다변량 데이터를 사용하여 예측하는 기계 학습 모델의 성능이 더욱 우수한 것을 확인하였다. 향후, 다양한 형태의 정형 데이터에 대해서도 본 논문에서 제안하는 다변량 데이터 처리 방법을 확장하여 기계 학습 모델을 통한 분석에 적용하여 성능을 분석할 예정이다.

참고문헌(Reference)

[1] J. Kotary, F. Fioretto, P. V. Hentenryck, and B. Wilder, “End-to-end constrained optimization learning: A survey,” arXiv:2103.16378, 2021. https://doi.org/10.48550/arXiv.2103.16378

[2] S. Tkatek, S. Bahiti, Y. Lmzouari, and J. Abouchabaka, “Artificial intelligence for improving the optimization of NP-hard problems: A review,” Int. J. Adv. Trends Comput. Sci. Appl., vol. 9, no. 5, pp.7411-7420, 2020. https://doi.org/10.30534/ijatcse/2020/73952020

[3] M. A. Wojtas and K. Chen, “Feature importance ranking for deep learning,” in Proc. of Adv. Neural Inf. Process. Syst. (NIPS), pp. 5015-5114, 2020. https://dl.acm.org/doi/10.5555/3495724.3496153

[4] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neurocomputing, vol. 237, pp. 350-361, May 2017. https://doi.org/10.1016/j.neucom.2017.01.026

[5] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” Social Netw. Comput. Sci., vol. 2, pp. 1-21, Mar. 2021. https://doi.org/10.1007/s42979-021-00592-x

[6] S. Athmaja, M. Hanumanthappa, and V. Kavitha, “A survey of machine learning algorithms for big data analytics,” in Proc. of International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-4, 2017. https://doi.org/10.1109/ICIIECS.2017.8276028

[7] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: An overview,” J.

Phys.: Conf. Ser., vol. 1142, pp. 1-15, Dec. 2018. https://doi.org/10.1088/1742-6596/1142/1/012012

[8] Kaggle.com, Titanic: Machine Learning form Disaster. [Online]. Available: https://www.kaggle.com/c/titanic/. [Accessed: 31-May-2024]

[9] Y. Alparslan et al., "Towards searching efficient and accurate neural network architectures in binary classification problems," in Proc. of Int. Jt. Conf. Neural Netw. (IJCNN), pp. 1-8, 2021. https://doi.org/10.1109/IJCNN52387.2021.9533483

[10] S. Ruder, "An overview of gradient descent optimization algorithms," arXiv:1609.04747, 2016. https://doi.org/10.48550/arXiv.1609.04747

[11] D. S. Chen and R. C. Jain, "A robust backpropagation learning algorithm for function approximation," IEEE Trans. Neural Netw., vol. 5, no. 3, pp. 467-479, May 1994. https://doi.org/10.1109/72.286917

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014. https://doi.org/10.48550/arXiv.1412.6980

[13] M. L. McHugh, "The chi-square test of independence," Biochem. Med., vol. 23, no. 2, pp. 143-149, Jun. 2013. https://doi.org/10.11613/BM.2013.018

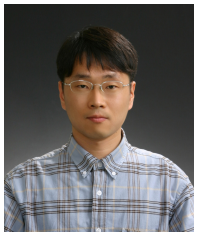
[14] H. Demirtas and D. Hedeker, "Computing the Point-biserial Correlation under Any Underlying Continuous Distribution," Communications in Statistics - Simulation and Computation, vol. 45, no. 8, pp. 2744-2751, 2016. https://doi.org/10.1080/03610918.2014.920883

[15] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," arXiv:1604.06737, 2016. https://doi.org/10.48550/arXiv.1604.06737

◐ 저 자 소 개 ◑

성 주 형 (Juhyoung Sung)
2015년 고려대학교 컴퓨터통신공학부 (공학사)
2019년 한국과학기술원(KAIST) 전기및전자공학과 (공학석사)
2015년~2017년 대한민국 육군 정보통신장교
2019년~2021년 삼성전자 네트워크사업부 연구원
2021년~현재 한국전자기술연구원 스마트네트워크연구센터 선임연구원
관심분야: 무선 통신, 시스템 최적화, 강화학습
E-mail: jh.sung@keti.re.kr

권 기 원 (Kiwon Kwon)
1997년 광운대학교 컴퓨터공학과 (공학사)
1999년 광운대학교 컴퓨터공학과 (공학석사)
2011년 중앙대학교 전자전기공학부 (공학박사)
1999년~현재 한국전자기술연구원 스마트네트워크연구센터 센터장
관심분야: 디지털트윈, 유무선디지털통신시스템, 해양수산 ICT융합
E-mail: kwonkw@keti.re.kr



박 경 원 (Kyoungwon Park)
1999년 중앙대학교 전기공학과 (공학사)
2001년 중앙대학교 전기공학과 (공학석사)
2005년 중앙대학교 전자전기공학부 (공학박사)
2005년~현재 한국전자기술연구원 스마트네트워크연구센터 수석연구원
관심분야: 디지털 통신 및 신호처리, AI/ML/DL
E-mail: kwpark@keti.re.kr

송 병 철 (Byoungchul Song)
1994년 명지대학교 전자공학과 (공학사)
1996년 명지대학교 전자공학과 (공학석사)
1999년~현재 한국전자기술연구원 스마트네트워크연구센터 수석연구원
관심분야: 네트워크 시스템, ICT융합 플랫폼
E-mail: kwonkw@keti.re.kr