교육용 프로그래밍 언어 기반 Deep al Yourself 실습 플랫폼을 위한 데이터 전처리 블록

이세훈*, 김기태*, 백민주*, 유채원O
OInha공업전문대학 컴퓨터시스템학과,
*인하공업전문대학 컴퓨터시스템학과
e-mail: {seihoon, kkt}@inhatc.ac.kr*, 100qhsk@naver.com*, cecilia0317@itc.ac.krO

Data Preprocessing block for Education Programming Language based Deep al Yourself Hands-on Platform

Se-Hoon Lee*, Ki-Tae Kim*, Min-Ju Baek*, Chae-Won YooO
ODept. of Computer System, Inha Technical College,
*Dept. of Computer System, Inha Technical College

● 요 약 ●

본 논문에서는 머신러닝 학습에 있어 데이터 전처리의 중요성과 기존 데이터 전처리 기능을 가진 교육용 실습 플랫폼 서비스의 단점은 개선할 수 있는 데이터 전처리 학습을 위한 교육용 블록코딩 기반 실습 플랫폼을 제안한다. 머신러닝 모델의 학습데이터는 데이터 전처리에 따라 모델의 정확도에 큰 영향을 미치므로 데이터를 다양하게 활용하기 위해서는 전처리의 필요성을 깨닫고 과정을 정확하게 이해해야 한다. 따라서 데이터를 처리하는 과정을 이해하고 전처리를 직접 실행해 볼 수 있는 교육용 프로그래밍 언어 기반 D.I.Y 실습 플랫폼을 구현한다.

키워드: 데이터 처리(data processing), 교육용 프로그래밍 언어(EPL : Educational Programming Language), 실습 플랫폼(Hands-on Platform)

I. Introduction

데이터 전처리는 머신러닝 학습에 있어 매우 중요한 부분이다. ‘Garbage in, garbage out(GIGO)’ 즉, ‘잘못된 입력은 잘못된 결과를 발생시킨다.’ 이처럼 머신러닝 역시 데이터를 기반으로 모델을 학습시켜 판단하기 때문에 불필요한 혹은 불균일한 데이터로 학습하게 된다면 모델의 정확도는 현저하게 떨어진다[1].

따라서 전처리를 통해 필요한 데이터들만 선별하고 모두 균일한 형태로 맞추어 정확도 높은 모델을 생성하기 위해 데이터 전처리 과정은 꼭 필요하다[2].

이러한 데이터 전처리 기능을 갖는 기존 교육용 블록코딩 기반 서비스들은 데이터 전처리 기능과 개념의 부족과 대량의 데이터를 처리하지 못하는 단점이 존재한다.

따라서 데이터 전처리 학습을 위한 교육용 블록코딩 기반 D.I.Y (이하 Deep AI Yourself)는 데이터 전처리 중요성을 깨닫고 학습의 과정에 대한 높은 이해도를 목표로 한다.

II. Related Works

엔트리는 기존의 텍스트 코딩이 아닌 블록형 언어를 기반으로 제작된 그래픽 기반 교육용 프로그래밍 언어 플랫폼이다. 데이터의 다양한 시각화를 제공하지만 데이터 전처리에 대한 기능이 부족하고 처리된 데이터를 저장할 수 없는 단점으로 처리된 데이터의 활용을 할 수 없다. Dataprep은 직관적인 UI로 복잡한 코드를 작성할 필요 없이 데이터를 정제할 수 있다. 하지만 처리가 셀 단위로 작업 되기 때문에 대량의 데이터 처리에는 어려움이 있다.

III. Function of D.I.Y

D.I.Y는 기존의 서비스들과 달리 다양한 데이터 전처리의 기능이 존재한다. 데이터 전처리 과정을 거치지 않은 데이터를 통해 학습된 머신러닝 모델은 다양한 문제점들이 발생한다. 오버 피팅으로 인한 모델의 정확도 감소, 의도에 어긋나는 모델 생성 등이 있다. 위와 같은 결과는 데이터 전처리 유무에 따라 결정된다. 이를 해결할 데이터 전처리 블록은 다음의 기능을 포함한다. 데이터 전처리를 위해 다양한 파일형식의 데이터를 불러올 수 있으며 테이블 형식으로 데이터를

한눈에 파악할 수 있다.

1. 누락 데이터 처리

누락데이터를 처리하기 위한 블록으로 0, 평균, 중앙값, 최솟값, 최댓값, 행 삭제 6가지의 방법을 갖는다.

2. 값의 범위 줄이기

값의 범위를 줄이기 위한 블록으로 정규화, 표준화 2가지 방법을 갖는다.

3. 컬럼 필터

사용자가 원하는 열의 삭제, 보존을 위한 블록이다.

4. 이미지 사이즈 재조정

이미지 전처리를 위한 블록으로 이미지의 사이즈를 사용자가 지정한 크기로 모두 변경하여 데이터의 통일성을 갖는다.

5. 이미지 회색조 변환

이미지의 색깔의 단색화시키기 위한 블록이다.

6. 로컬 저장

전처리한 데이터를 사용자의 pc에 저장한다.

Ⅳ. Experiment

데이터는 인공지능 모델의 정확도에 큰 영향을 준다. 따라서 인공지능 모델을 생성하기 전 데이터를 처리하기 위해 D.I.Y의 빅데이터 확장 블록을 사용한 데이터 처리에 대한 실험을 수행한다.

Fig. 1. Block used for data processing

Fig. 1은 데이터 처리를 위해 D.I.Y의 빅데이터 확장블록을 사용한 모습으로 본 실험에서는 보스턴 집에 대한 여러 조건에 따른 집값으로 구성되어 있는 데이터를 처리한다. 첫 번째 블록은 불러온 데이터에 대한 변수명을 설정한 모습이고 두 번째 블록은 불러온 csv파일을 D.I.Y에서 처리할 수 있도록 변수에 담은 것이다. 세 번째, 네 번째 블록은 각각 누락값, 라벨링에 대한 처리를 한 모습이며 마지막은 처리된 데이터의 확인을 위한 시각화를 수행하는 블록의 모습이다.

보스턴집값예측 데이터 테이블

주택매매가격	건축년도	지붕종류	중앙에어컨여부	1층평방피트	2층평방피트	차고유형	주차가능한차량수
1461.00000000	1961.00000000	0.00000000	1.00000000	896.00000000	0.00000000	0.00000000	1.00000000
1462.00000000	1958.00000000	2.00000000	1.00000000	1329.00000000	0.00000000	0.00000000	1.00000000
...							
1560	1910	Gambrel	Y	935	744	Detchd	2

Fig. 2. before data preprocessing

보스턴집값예측 데이터 테이블

건축년도	지붕종류	중앙에어컨여부	1층평방피트	2층평방피트	차고유형	주차가능한차량수
1961.00000000	0.00000000	1.00000000	896.00000000	0.00000000	0.00000000	1.00000000
1958.00000000	2.00000000	1.00000000	1329.00000000	0.00000000	0.00000000	1.00000000
...						
1910.00000000	1.00000000	1.00000000	935.00000000	744.00000000	2.00000000	2.00000000

Fig. 3. after data preprocessing

Fig. 2는 전처리 과정을 거치기 전의 데이터로 누락데이터, 텍스트로 된 데이터가 존재하는 등 전처리가 필요한 데이터로 구성되어 있다.

Fig. 3은 Fig. 1에 해당하는 블록의 실행 결과를 시각화 블록을 사용하여 테이블로 전처리 된 데이터를 한 눈에 확인할 수 있다. 시각화 블록을 통해 데이터 처리 과정 중에 데이터를 확인해 정상적으로 데이터 전처리가 이루어졌음을 알 수 있다.

Ⅴ. Conclusions

본 논문에서는 D.I.Y와 다른 블록코딩과 차별점, 머신러닝 모델 학습에서 데이터 전처리의 중요성과 그 과정을 학생들이 직접 체험하고 이해할 수 있는 EPL기반 D.I.Y 실습 플랫폼을 제안하였다. 이미지와 csv파일 모두 전처리가 가능하며 이를 D.I.Y의 하나의 구성인 머신러닝 학습 부분에서 직접 전처리한 데이터로 모델을 생성할 수 있다. 이처럼 머신러닝과 같이 전처리된 데이터를 활용해야만 하는 분야에 적용해봄으로써 데이터 전처리의 중요성과 학습의 필요성을 통해 제시한 플랫폼의 유의미함을 입증한다.

REFERENCES

[1] Myoungwon Jo, Namgi Kim, Byoung-Dai Lee, "An Image Preprocessing Scheme Based on Machine Learning for ROI Extraction" Journal of KIIT. vol. 14, no. 31, pp. 29-34, Jun. 2016

[2] Jun-Mo Jo, "Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance", Journal of the KIECS. pp. 547-552, vol. 14, no. 3, Jun. 30. 2019