EDISON 플랫폼을 활용한 웹 기반 데이터 전처리 도구 설계 및 구현

한성근, 이정철
한국과학기술정보연구원

{sghan, jclee}@kisti.re.kr

Design and Implementation of a Web-Based Data Preprocessing Tool on EDISON Platform

Sunggeun Han, Jeongcheol Lee
Korea Institute of Science and Technology Information

요 약

데이터 전처리는 데이터 분석이나 머신 러닝 워크플로우에서 중요하고도 시간이 많이 소요되는 단계이다. 본 논문에서는 연구자들의 데이터 전처리 작업을 쉽고 효율적으로 수행하기 위해 EDISON 플랫폼 기반 위에서 개발된 웹 기반 데이터 전처리 도구를 소개한다. 이 도구는 데이터 수집, 탐색, 변환, 피처 엔지니어링과 같은 주요 프로세스를 지원하는 직관적인 사용자 인터페이스를 제공한다. 이를 통해 연구자들은 코딩에 대한 전문 지식이 없이도 누락된 데이터, 이상치, 변환을 효율적으로 처리할 수 있으며, 데이터분석을 위한 데이터셋 준비에 필요한 시간을 줄일 수 있다. 본 논문에서 제안하는 도구는 피처 엔지니어링을 위한 AI 기반 자동화를 통합하고 외부 데이터 소스에 대한 API 지원을 제공하여 연구 협업과 데이터 재사용을 더욱 용이하게 한다.

Ⅰ. 서 론

데이터 전처리는 데이터 분석이나 머신 러닝 프로젝트의 핵심 단계로, 모델 성능에 직접적인 영향을 미친다. 이 과정은 원시 데이터를 알고리즘 또는 모델에 적합한 형태로 변환하는 것으로, 데이터 수집부터 결측치 처리, 이상치 탐지, 인코딩, 스케일링, 피처 선택 및 생성에 이르는 복잡한 작업들을 포함한다. 연구에 따르면, 데이터 전처리는 전체 머신 러닝 프로세스의 50%에서 80%까지 차지할 수 있으며, 이는 데이터의 규모와 복잡성에 따라 증가한다[1]. AutoML(Automated Machine Learning) 도구들은 이러한 복잡한 작업을 자동화하여 사용자의 부담을 줄이고자 등장했다[2]. 그러나 이러한 도구들은 주로 기술적 배경을 가진 사용자를 대상으로 하며, 특정 연구 목적에 맞는 세밀한 조정이 어렵다는 한계가 있다. 특히 연구자들은 다양한 데이터 소스와 복잡한 구조를 다루기 때문에, AutoML 만으로는 모든 문제를 해결하기 어렵다[3]. 데이터 전처리의 초기 단계에서 발생하는 오류, 불일치, 결측치, 이상치 등의 문제는 여전히 사용자의 개입을 필요로 한다. 이러한 문제들을 해결하지 않으면 모델 성능 저하뿐만 아니라 연구 결과에도 심각한 영향을 미칠 수 있다. 따라서 효과적인 전처리 과정은 데이터의 신뢰성 확보와 모델 성능 극대화를 위해 필수적이다.

본 연구에서는 이러한 문제를 해결하기 위해 웹 기반의 데이터 전처리 도구를 설계하고 구현하였다. 이 도구는 직관적인 웹 인터페이스를 통해 데이터 수집, 탐색, 정제 및 변환 과정을 지원한다. 특히 EDISON 플랫폼[4]에서 연구자들이 다양한 데이터를 분석할 수 있도록 사용자 친화적인 UI 를 제공하여, 데이터 분석의 전체 워크플로우를 체계적으로 수행할 수 있게 한다. 제안하는 도구는 기존의 자동화 도구들과 달리, 연구자들이 각 도메인과 데이터 특성에 맞게 전처리 과정을 유연하게 조정할 수 있도록 설계되었다. 결측치 처리, 이상치 탐지 및 처리, 데이터 변환 등의 필수적인 기능을 직관적인 인터페이스에서 쉽게 수행할 수 있어, 연구자들의 시간과 노력을 절감할 수 있다. 이를 통해 연구자들이 데이터 전처리에 소요되는 시간을 줄이고, 더 많은 시간을 실제 데이터 분석과 모델 개발에 할애할 수 있게 한다. 본 논문에서는 제안한 웹 기반 데이터 전처리 도구의 설계와 구현에 대해 설명한다.

Ⅱ. 데이터 전처리 도구 설계

EDISON 플랫폼에서 지원하는 데이터 전처리 도구는 연구자들이 다양한 데이터를 분석할 수 있도록 사용자 친화적인 웹 기반 UI 를 제공한다. 이를 통해 데이터 수집, 탐색, 정제, 피처 엔지니어링, 저장 등 데이터

분석의 전체 워크플로우를 직관적이고 체계적으로 수행할 수 있다. 본 논문에서 설계한 기능은 다음과 같다.

1. 데이터 수집(Upload): 연구자들이 손쉽게 데이터를 수집하고 불러올 수 있도록 다양한 소스에서의 데이터 업로드를 지원한다. 데이터 소스 위치에 따라 From Local, From SDR, From Remote 로 구성된다.

2. 데이터 탐색 및 가시화(Explore & Visualize): 수집한 데이터를 분석하기 전 탐색하고 시각화하여 데이터의 특성을 쉽게 파악할 수 있도록 한다. 탐색 기능은 데이터 탐색 범위에 따라 All Data, Top-N, Bottom-N, Range 로 구성되며, 가시화 기능은 Heatmap, Missing Bar, Box Plot, Feature Importance 로 구성된다.

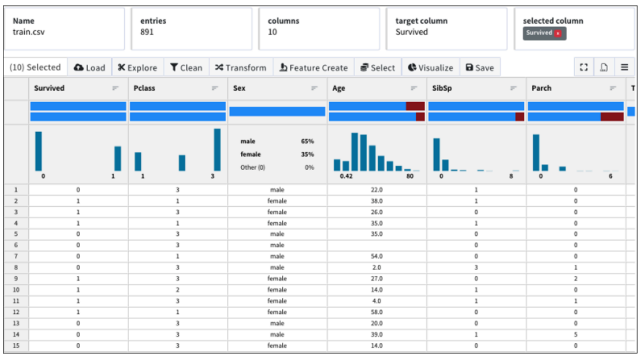
3. 데이터 정제 및 변환(Clean & Transform): 분석에 적합한 데이터를 만들기 위해 필수적인 데이터 정제 및 변환 기능을 제공한다. 데이터 정제 기능은 Missing Data(결측치 처리), Outliers(이상치 처리)로 구성되며, 데이터 변환 기능은 인코딩 알고리즘에 따라 Label Encoding, One-hot Encoding 과 스케일링 알고리즘인 Standard Scaling, MinMax Scaling 으로 구성된다.

4. 특징 생성(Feature Create): 데이터에서 더 의미 있는 특징을 추출하고 가공하는 기능을 제공한다. 특징 생성 알고리즘에 따라 Expert-based, Rule-based, AI-based 기능을 제공하며, 특징 선택을 위해 Sampling 과 Merging, Filter 기능을 제공한다.

5. 데이터 저장(Store): 분석한 데이터와 결과를 EDISON 플랫폼의 리포지터리인 SDR 에 저장하여 다른 연구자들과 공유할 수 있는 Save 기능을 제공한다.

Ⅲ. 데이터 전처리 도구 구현

본 연구에서 구현된 데이터 전처리 도구는 Python 과 Java Liferay 프레임워크를 기반으로 설계되었으며, 웹 기반 사용자 인터페이스(UI)는 JavaScript 를 사용하여 직관적이고 반응형으로 개발되었다. 제안한 데이터 전처리 도구는 연구자들이 데이터를 쉽고 빠르게 처리하고, 분석에 필요한 모든 과정을 웹 환경에서 수행할 수 있도록 한다. 특히, 웹 기반 UI 는 사용자의 입력에 따라 실시간으로 데이터를 처리하고 결과를 즉시 시각화할 수 있어 데이터 분석 과정의 직관성을 크게 향상시킨다. 다음 그림은 웹 기반 데이터 전처리 도구의 화면을 나타낸다.



웹 기반 데이터 전처리 도구 UI

연구자는 전처리 할 데이터를 업로드하여 데이터 전처리에 필요한 다양한 기능들을 UI 기반으로 작업할 수 있으며, 곧바로 화면을 통해 데이터 변환 과정을 확인할 수 있다. AI 기반 기능은 연구자들이 데이터 전처리에서 복잡한 피처 엔지니어링 작업을 자동으로 수행할 수 있도록 지원하는 핵심 기능 중 하나이다. 본 연구에서는 AI 모델을 활용하여 자동으로 중요한 피처를 선택하고, 기존 데이터를 강화하는 기능을 제공한다. 이를 위해 EDISON 플랫폼의 AI 프레임워크에 공유된 최신 모델을 적용할 수 있도록 했으며, 사용자 친화적인 인터페이스에서 이 기능을 쉽게 활용할 수 있다.

Ⅳ. 결론

본 연구에서 제안한 데이터 전처리 도구는 연구자들이 데이터 전처리와 피처 엔지니어링 작업을 보다 쉽게 수행할 수 있도록 돕는 웹 기반 도구이다. 이 도구는 사용자 친화적인 UI 와 다양한 데이터 처리 기능을 통합하여 데이터 전처리 과정에서 발생하는 복잡한 문제들을 해결하는 데 중점을 두었다. 특히, 연구자들은 데이터를 효율적으로 수집하고, 이상치 및 누락된 데이터를 처리하며, 변환 작업을 손쉽게 수행할 수 있다.

향후 연구에서는 이 도구를 실제 연구 프로젝트에 적용하여 그 성능과 효율성을 검증하는 작업이 필요하다. 다양한 연구 도메인에서 실제 데이터를 사용하여 도구의 적용 가능성을 평가하고, 사용자들의 피드백을 바탕으로 도구를 개선할 계획이다. 또한, 데이터 처리 과정의 자동화를 더욱 발전시켜 연구자들이 더 많은 시간을 핵심 분석 작업에 집중할 수 있도록 지원할 것이다.

ACKNOWLEDGMENT

This research was supported by the EDISON Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. NRF-2022M3C1A6090416).

참 고 문 헌

[1] H. T. Duong and T. A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," Computational Social Networks, vol. 8, no. 1, p. 1, 2021.

[2] M. Feurer and F. Hutter, "Automated Machine Learning," Springer Handbook of Computational Intelligence, pp. 1-30, 2023.

[3] H. Liu, et al., "Automated Data Preprocessing: State-of-the-Art and Open Challenges," ACM SIGKDD Explorations Newsletter, vol. 25, no. 1, pp. 17-30, 2023.

[4] S. Han, et al., "Data Framework Design of EDISON 2.0 Digital Platform for Convergence Research," KSII Transactions on Internet and Information Systems (TIIS), vol. 17, no. 8, pp. 2292-2313, 2023.