비 전문가를 위한 시각화 솔루션에서의 분류 모델링 서비스 구현

엄태창, 박민호
숭실대학교

francis.eom@gmail.com, mhp@ssu.ac.kr

Implementation of classification modeling services in visualization solutions for non-professionals

TaeChang Eom, Minho Park
Soongsil Univ.

요 약

정보기술 인프라와 분석 기술의 지속적인 발전으로 데이터 분석이 현실 업무에 적용되어 업무 효율성을 높이고 있다. 요즘과 같이 변동성이 큰 환경에서 즉시적인 데이터 분석을 통한 현실 업무에의 적용이 조직이나 기업의 성과를 개선하는데 큰 요인이 된다. 그러나, 데이터 분석 기술이라는 높은 진입 장벽으로 인해 업무 이해도가 높은 비 전문가의 직접적인 데이터 분석 수행은 어려운 현실이다. 이에 시각화 솔루션 기반으로 데이터 분석 기술을 융합한 서비스 구현으로 비 전문가도 분류 모델링을 수행할 수 있는 기반을 제공하는데 있다.

Ⅰ. 연구배경

많은 조직과 기업에서 데이터 분석을 통해 업무 의사결정에 도움을 받고 있으며, 빠른 환경 변화의 대응이 필요한 데이터 분석을 필요로 하고 있다. 일반적으로 데이터 과학자와 업무 담당자의 협업을 통해 데이터 분석 업무가 수행 된다. 하지만, 협업으로 인한 분석 목적 이해 및 분석 업무 수행에 많은 소요 시간과 고비용 문제가 발생한다.

2015 년 Gartner 에서는 "수학이나 통계에 대한 깊은 지식 없이도 자신의 전문지식과 데이터 과학의 원리를 결합할 수 있는 비즈니스 사용자"를 시민 데이터 과학자(Citizen Data Scientist)라고 정의했다.

관련 선행 연구[1][2]에서는

“데이터 과학자는 분석 이론 및 분석 기술은 있으나, 분석 데이터를 명확히 이해하고 분석을 수행하는 데의 한계”

“분석 데이터에 대한 이해력만 갖춘 비전문가도 쉽게 분석이 가능한 시스템의 요구”

"시민 데이터 과학자는 사용하기 쉬운 분석 도구 및 기술의 도움으로 예측과 같은 데이터 분석과 비즈니스 모델을 만드는 역할”로 시민 데이터 과학자의 필요성과 역할을 수행하기 위한 방법론이 제시 되었지만, 실제 서비스 구현은 후속 연구 과제로 정의 하였다.

이에 비 전문가도 쉽게 사용할 수 있는 시각화 솔루션 기반의 분류 모델링 서비스 구현 방안을 제시한다.

Ⅱ. 관련연구

서비스 구현을 위해 특정되지 않은 데이터의 사용과 성능이 확보된 모델 학습 방안을 연구한다.

시각화는 데이터를 시각적으로 표현하여 인사이트 도출 및 효과적인 전달을 가능하게 한다. 시각화는 데이터 분석 영역에서도 분석 데이터의 특성 및 분석 결과를 표현 하는데 많이 사용된다. TIBCO Spotfire[3] 는 대표적인 데이터 시각화 솔루션으로 다양한 데이터 소스의 사용과 데이터 전처리 기능이 제공되고, 내장된 다양한 시각화 차트와 시각화 차트 간의 연계 분석을 지원한다. 최근 버전부터 Python 인터프리터가 내장되어 솔루션 내에서도 Python 라이브러리를 이용한 데이터 분석이 가능한 환경이 구성되었다.

머신러닝의 분류(Classification)는 지도학습의 한 종류로 특징 데이터를 통해 분류 Class 예측을 목적으로 한다. 현실 업무에서는 업무 처리 방안(Class)을 정의하고, 처리할 데이터를 통해 최적화 된 방안(Class)으로 분류하고 처리한다. 가령, 위험등급을 상, 중, 하 Class 로 정의 하고, 요청 된 특징 데이터로 위험등급을 예측하고 결과를 업무에 활용한다. 분류 학습은 종속변수의 이진 또는 다중 범주에 따라 학습 방법이 다르며, 일반적으로 많이 사용되는 모델로는 Logistic Regression, Random Forest, KNN 이 있다.

모델 학습 과정에서 종속변수의 클래스 비율에 의해 모델 예측 성능을 저해 할 수 있다. 이러한 문제를 비대칭 데이터 문제(imbalanced data problem)라고 한다. 모델의 정확도는 높아도, 작은 비율의 클래스는 재현율이 상대적으로 많이 낮아진다. 이 문제는 학습 데이터의 소실이 없는 Over Sampling 기법 이나 클래스의

가중치를 적용하는 Cost-sensitive Learning 기법으로 보완 가능하다. [4]

모델 성능 확보를 위해 Hyper parameter 최적화를 고려해야 한다. 학습 알고리즘마다 Hyper parameter 가 있고, 최적 parameter 의 적용만으로도 성능 향상을 할 수 있다. 기법으로는 Grid search, Random search, Bayesian Optimization 이 있다. Grid search 는 수행 시간이 오래 걸리는 단점이 있어, 학습 데이터가 작은 경우에 사용되고, Random search 는 Grid search 에 비해 정확도가 떨어지나 수행 시간은 상대적으로 작다. Bayesian Optimization 는 이전 두 기법보다 더 나은 결과를 제공하나, Bayesian Optimization 을 적용하기 위한 또 다른 자체 Hyper parameter 가 있으며, 이를 tuning 해야 한다는 복잡성이 있다.

Scikit-Learn 패키지의 make_classification 함수로 다양한 가상 데이터를 생성하여 구현될 서비스의 실험에 사용한다. 이 함수는 종속변수의 클래스, 종속변수와 상관성이 있는 독립변수, 독립변수간의 선형 조합 형성이 가능하여 다양한 유형의 가상 학습 데이터를 활용하여 구현 서비스의 완성도를 높일 수 있다.

Ⅲ. 제안 서비스

3.1 서비스 환경

본 서비스는 시각화 솔루션 기반으로 제공 되므로, 저작 도구인 TIBCO Spotfire Analyst 로 구성 된다. 아래 표 1 은 서비스 환경 정보 이다.

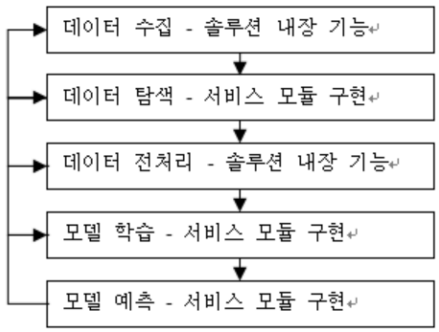
표 1. 서비스 환경 정보

구분	사양
OS	Windows 10 이상
CPU	2.0Ghz, 64bit 4 코어 이상
RAM	16GB 이상
HDD	여유공간 10GB 이상

3.2 서비스 흐름

서비스 흐름은 일반적인 분류 모델링 학습 절차와 동일하고, 아래 그림 1 과 같다.

그림 1. 서비스 흐름



"데이터 수집" 단계부터 "모델 예측" 단계로 진행되며, 각 단계별 피드백에 의해 이전 단계에서의 재 진행 한다.

3.3 서비스 구성

구성은 총 7 개의 메뉴로 사용 가이드, 데이터 명세, 데이터탐색 - 단일 변수, 데이터탐색 - 다중 변수, 모델학습 - KNN, 모델학습 - Random forest, 모델 예측으로 구성된다. 그림 2 는 서비스 화면 이다.

그림 2. 서비스 화면

3.4 서비스 구현

정의된 서비스 흐름 및 구성을 솔루션 API 로 시각화 연동 개발하고, 모델 학습 및 예측 모듈은 Python 라이브러리를 활용하여 개발하고 솔루션에 이식하여 구현된다.

3.4.1 데이터 수집

솔루션에 내장된 데이터 조회 및 병합 기능이 사용된다. 사용 가능한 데이터 소스는 Database 및 파일 데이터가 있다. 분석에 필요한 가공된 새로운 특징 변수 생성도 가능하다.

3.4.2 데이터 탐색 및 전처리

솔루션 API 로 분석 데이터와 시각화 표현 속성을 연계할 수 있다. 이를 활용한 데이터 탐색 템플릿 모듈 구현으로 변수 선택 등의 조작으로 시각화된 데이터 탐색 및 데이터 이상 여부의 확인을 가능하게 한다.

데이터 전처리는 솔루션 데이터 변환 기능을 사용한다. 데이터 탐색을 통해 확인된 이상치, 결측치에 대한 제거, 보정 처리를 한다.

3.4.3 모델 학습

모델 학습을 위한 학습 템플릿 모듈을 구현한다. 분류 모델링 알고리즘은 사이킷런 Random Forest 와 KNN 을 적용하였다.

모델 성능 확보를 위해 Random search 를 이용한 hyper parameter 최적화 방안을 적용한다. 성능 평가는 Classification report, Confusion matrix 를 시각화하여 제공한다. 학습 된 모델은 저장하여 모델 예측 시 사용된다.

3.4.4 모델 예측

모델 예측 템플릿 모듈은 저장된 학습 모델로 예측을 실행하고, 각 알고리즘 별 예측 결과를 비교 분석할 수 있는 시각화를 제공한다.

3.4 서비스 입출력

서비스 구현 방안으로 정의된 서비스 메뉴별 입/출력은 아래 표 2 와 같다.

표 2. 서비스 메뉴별 입출력 명세

입력	출력
데이터 명세	
분석 데이터 선택	데이터 명세 기초 통계량
종속변수 선택	종속변수 설정 반영
데이터 탐색-단일변수	
연속형 변수 선택	Box Plot Histogram Data Table
범주형 변수 선택	Cross Table Data Table
데이터 탐색-다중변수	
X 축, Y 축 변수 선택	Box Plot Scatter Plot Cross Table Heat Map Data Table
모델 학습	
모델명 입력 분석 제외 변수 선택 학습 실행	Train / Test Score Confusion Matrix Table Classification Report Table Trained Model
모델 예측	
예측 데이터 선택 예측 실행	Predict / Proba. Cross Table Predict / Proba. Histogram Predict / Proba. Table

3.5 서비스 활용

서비스 구성 및 메뉴 흐름대로 입력 옵션의 설정 및 실행으로 활용 된다.

Ⅳ. 서비스 실험

실험을 위하여 3 개의 분류를 가지는 가상 데이터 2 종(10,000*16)을 생성하였다. 첫번째 셋은 종속변수의 분류 비율이 균등하고, 두번째 셋은 각각 10%, 30%, 60% 비율로 구성되어 있다. 공통적으로 15 개의 독립변수로 구성되고, 이중 4 개의 독립변수는 종속변수와 상관성이 있도록 설정 하였고, 가상 데이터 중 500 건을 분할하여 모델 예측 모듈 실험에 활용 하였다.

Random forest 와 KNN 으로 균등/비균등 데이터로 모델 학습을 실험하였다. 균등 / 비균등 조건에서 정확도와 정밀도는 0.94 이상으로 학습 되었으나, 비균등 조건에서의 재현율이 상대적으로 낮음이 확인 되었다. 실험 결과는 아래 표 3 과 같이 정리할 수 있다. 정밀도, 재현율, f1-score 는 분류 클래스별 최소값 이다.

표 3. 학습 실험 분류 결과

학습 알고리즘	분류 비율	정확도	정밀도	재현율	f1-score
Random Forest	균등	0.97	0.96	0.94	0.96
	비균등	0.97	0.97	0.83	0.91
KNN	균등	0.95	0.95	0.92	0.94
	비균등	0.96	0.94	0.79	0.86

Ⅴ. 보완점 및 기대효과

비 전문가를 위한 특정되지 않은 데이터에 대한 분류 모델링 서비스 구현으로 클래스 비균등 클래스 데이터 학습의 재현율 보완 및 적용 가능한 분류 학습 알고리즘의 추가 적용 보완이 필요하다.

본 연구를 통해 업무 지식과 업무 데이터의 이해도는 가지고 있으나 데이터 분석 기술의 진입장벽에 의해 직접적인 데이터 분석을 수행하지 못하는 분석 비 전문가로 하여금 일정 수준의 분석을 진행할 수 있는 서비스 기반을 제공할 수 있다는 기대 효과를 가지고 있다.

참 고 문 헌

[1] "시민 데이터 과학자를 위한 빅데이터 예측 지원 서비스" 장재영 - 한국인터넷방송통신학회 논문지, 19(2), pp.151-159 Apr, 2019

[2] "디지털 혁신의 시대에 '시민 데이터 과학자(CDS)'로 성장하기 위해 필요한 지식과 도구들" 민철희 - 대한산업공학회 춘계공동학술대회 논문집. 2022-06 2022(6):1685-1699

[3] TIBCO Spotfire 공식 사이트: https://www.tibco.com/ko/products/tibco-spotfire

[4] 지도학습 기반 암상 분류 시 클래스 간 자료 불균형을 고려한 평가지표 개발, 김도완, 최준환, 변중무 - 지구물리와 물리탐사 Geophysics and Geophysical Exploration Vol. 23, No. 3, 2020, p. 131-140 https://doi.org/10.7582/GGE.2020.23.3.00131