https://doi.org/10.5392/JKCA.2019.19.01.001

정형 데이터와 비정형 데이터를 동시에 고려하는 기계학습 기반의 직업훈련 중도탈락 예측 모형

A Machine Learning-Based Vocational Training Dropout Prediction Model Considering Structured and Unstructured Data

하만석, 안현철
국민대학교 비즈니스IT전문대학원

Manseok Ha(msha@kookmin.ac.kr), Hyunchul Ahn(hcahn@kookmin.ac.kr)

요약

직업훈련 교육 현장에서 느끼는 가장 큰 어려움 중 하나는 중도탈락 문제이다. 훈련과정마다 많은 수의 학생들이 중도탈락을 하게 되어 국가 예산 낭비 및 청년 취업률 개선에 장애 요인이 되고 있다. 본 연구에서는 중도탈락의 원인을 주로 분석한 기존 연구들과 달리, 각종 수강생 정보를 활용하여 사전에 중도탈락을 예측할 수 있는 기계학습 기반 모형을 제안하고자 한다. 특히 본 연구의 제안모형은 수강생 관련 정형 데이터 뿐 아니라 비정형 데이터인 강사의 상담일지 정보까지 동시에 고려하여 모형의 예측정확도를 제고하고자 하였다. 이 때 비정형 데이터에 대한 분석은 최근 주목받고 있는 텍스트 분석 기술인 Word2vec과 합성곱 신경망을 이용해 수행하였다. 국내 한 직업훈련기관의 실제 데이터에 제안모형을 적용해 본 결과, 정형 데이터만을 사용하여 중도탈락을 예측할 때보다 비정형 데이터를 함께 고려했을 때 예측의 정확도가 최대 20%까지 향상됨을 확인할 수 있었다. 아울러, Support Vector Machine을 기반으로 정형 데이터와 비정형 데이터를 결합해 분석했을 때, 검증용 데이터셋 기준으로 90% 후반대의 높은 예측 정확도를 나타냄을 확인하였다.

■ 중심어 : | 직업훈련 교육 | 중도탈락 | 기계학습 | 합성곱 신경망 | Word2vec |

Abstract

One of the biggest difficulties in the vocational training field is the dropout problem. A large number of students drop out during the training process, which hampers the waste of the state budget and the improvement of the youth employment rate. Previous studies have mainly analyzed the cause of dropouts. The purpose of this study is to propose a machine learning based model that predicts dropout in advance by using various information of learners. In particular, this study aimed to improve the accuracy of the prediction model by taking into consideration not only structured data but also unstructured data. Analysis of unstructured data was performed using Word2vec and Convolutional Neural Network(CNN), which are the most popular text analysis technologies. We could find that application of the proposed model to the actual data of a domestic vocational training institute improved the prediction accuracy by up to 20%. In addition, the support vector machine-based prediction model using both structured and unstructured data showed high prediction accuracy of the latter half of 90%.

■ keyword : | Vocational Training | Dropout | Machine Learning | Convolutional Neural Network | Word2vec |

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5A2A03067632).

접수일자 : 2018년 10월 24일
심사완료일 : 2018년 11월 21일
수정일자 : 2018년 11월 20일
교신저자 : 안현철, e-mail : hcahn@kookmin.ac.kr

I. 서 론

청년 실업 문제 해결이 국가적 과제로 떠오르고 있다. 실업 문제 해결을 위해 막대한 국가 예산이 투입되고 있으며 그 일환으로 일자리 창출을 위한 직업훈련 교육에도 많은 예산이 투입되고 있다. 정부에서 지원하고 있는 직업훈련은 미취업자 및 실업자들 뿐만 아니라 노동시장 진입에 불리한 장애인 및 무기능자, 고령자, 3D 직종 및 첨단 분야 직종의 인력을 양성하는데 큰 기여를 하고 있다.

현재 운영되고 있는 직업훈련 교육과정들의 현황을 볼 때 당초의 취지와 달리 다수의 중도탈락자가 발생하고 있고 교육을 수료한 이후에 취업을 하지 못하거나 교육받은 직종과 다른 직종으로 취업하는 경우도 많은 것이 현실이다. 최근의 사례를 살펴보면 2017년 한 해 동안 실업자 직업능력개발지원 사업에 추경을 포함하여 659억 5300만원의 예산을 편성하고 31,288명을 대상으로 교육을 실시했는데 그 중 10.4%의 인원이 중도탈락을 했으며 수료자 중 취업에 성공한 사람도 42.4%로 2016년도의 절반 이하 수준에 불과하였다[1].

그동안 중·고등학교, 대학교 등의 정규교육과정과 이러닝 교육 분야 등에서 중도탈락 문제를 개선하기 위한 선행연구들이 많이 진행되었다. 하지만, 대부분의 선행연구들은 중도탈락을 유발하는 원인 분석을 위주로 진행되었다. 특히 직업훈련 분야에서는 중도탈락의 원인을 분석하는 연구들이 많이 있었지만 중도탈락을 예측하기 위한 연구는 미미한 실정이다.

특히 중도탈락을 예측하는 기존 연구들은 대부분 정형 데이터만을 사용해서 예측을 시도하고 있다. 하지만 일선 교육 현장에서 만들어지는 데이터들은 정형 데이터 외에도 훈련교사들이 입력하는 상담일지, 학생들이 입력하는 만족도 조사 등의 다양한 비정형 데이터들이 존재한다. 최근 수행된 한 선행연구에 따르면 정형 데이터 뿐 아니라 비정형 데이터를 함께 고려하여 분석할 때 정확도가 향상된다는 결과를 제시하고 있다[2]. 이러한 배경에서 본 연구는 정형 데이터와 비정형 데이터를 종합적으로 활용하는 기계학습 기반의 중도탈락 예측 모형을 제안한다.

본 연구에서는 직업훈련의 성과를 개선하기 위한 노력의 일환으로 훈련생의 중도탈락을 예측하기 위하여 정형 데이터를 사용하는 전통적인 데이터마이닝(data mining) 기법에서 한 단계 더 나아가 비정형 데이터 분석을 위해 기계학습을 활용할 것을 제안한다. 구체적으로 본 연구의 제안모형은 비정형 데이터에 이미지 인식과 자연어 처리 분야 등에 특화된 합성곱 신경망(convolutional neural network, CNN)을 적용하여 중도탈락 가능성을 예측하도록 하였다[2]. 이를 통해, 중도탈락 및 취업 가능성이 낮은 훈련생에 대해 조기경보를 가능하게 하여, 직업훈련의 성과를 높이고자 한다. 또한 정형 데이터에만 의존하는 전통적인 데이터마이닝 기법들과 그 성능을 비교하여, 정형 데이터와 비정형 데이터를 동시에 고려할 때 어떤 성능의 향상이 있는지 비교해 보고자 한다.

이후 본 연구의 구성은 다음과 같다. 제2장에서는 이론적 배경에 대해 설명하고, 제3장에서는 중도탈락자 예측을 위한 연구 모형을 제시한다. 제4장에서는 중도탈락자 예측을 위한 모형을 구현하고 실증분석하며, 마지막 제5장에서는 본 연구의 결과 및 향후 연구방향을 제시한다.

II. 이론적 배경

2.1 직업훈련교육

직업훈련이란 「산업교육진흥 및 산학협력촉진에 관한 법률」 및 「근로자 직업능력 개발법」과 그 밖의 다른 법령에 따라 학생과 근로자 등에게 취업 또는 직무수행에 필요한 지식·기술 및 태도를 습득·향상시키기 위하여 실시하는 훈련을 말한다[3]. 직업훈련에 대하여 배경석[4]은 직업훈련은 직업활동을 함에 있어서 직무수행에 필요한 능력을 습득, 향상시킬 목적으로 실시하는 훈련을 말하는 것으로서 흔히 말하는 교육훈련이라는 큰 범주에 포함이 된다고 보았다. 우리나라에서 시행되는 직업훈련은 크게 근로자의 직무능력 향상을 위해 실시되는 재직자 직업훈련과 실업자의 취업능력 및 직업기초능력향상을 지원하는 실업자 직업훈련으로 분

류하고 있다[5].

2.2 중도탈락 예측

직업훈련 현장에서의 가장 큰 애로사항을 꼽으라면 훈련생들의 중도탈락 문제라고 말할 수 있다. 지금까지 직업훈련을 포함하여 성인학습자들이 교육과정을 끝까지 마치지 못하고 중도탈락하는 요인들에 대한 연구들이 다수 발표되어 왔다([표 1] 참고). 그런데 기존의 중도탈락에 관련된 연구들은 대부분 중도탈락의 원인을 분석하는 연구를 위주로 진행되어 왔다. 그나마 중·고등학교, 대학교 등의 정규 교육과정과 이러닝 교육과정의 중도탈락 원인에 대한 분석이 대부분이며, 직업훈련 관련 분야에서의 중도탈락 관련 연구는 많지 않은 편이다.

표 1. 성인학습자의 중도탈락 요인

연구자	중도탈락의 주요 요인
권근배[6]	훈련부적응(적성에 맞지 않음), 흥미 부족, 중도탈락 경험의 유무, 결손가정, 생계 곤란
권혜진[7]	개인의 교육에 대한 흥미
김경희[8]	개인의 교육에 대한 흥미와 자긍심
박대권[9]	훈련기간(장기 훈련과정의 경우 중도탈락 비율이 높음), 성별(남성이 여성보다 중도탈락 비율이 높음), 훈련내용에 대한 불만족
배경석[4]	낮은 학업수행능력, 동료 및 강사와의 낮은 친밀도
이준택[10]	진로에 대한 불확실성, 주변의 지지 부족, 훈련부적응, 자기존중감 부족
정선정[11]	낮은 성취욕구
Conte et al.[12]	성별, 결혼여부, 훈련 기간
Yi et al.[13]	낮은 학업 성취도, 가족의 지원 부족

중도탈락을 예측하기 위하여 연관규칙학습, 의사결정나무, 로지스틱 회귀분석, 인공신경망 등의 다양한 전통적인 데이터마이닝 알고리즘들을 활용한 연구들이 많이 진행되어 왔다. 구본용 등[14][15]은 중고생들의 중도탈락을 예측하기 위하여 의사결정나무와 인공신경망 모형을 활용하였다. 전주성[16]은 사이버 대학의 잠재적 중도탈락자를 예측하기 위하여 빈도분석, 요인분석, 로지스틱 회귀분석 등의 기법을 사용하여 총학습시간, 강의접속수, 교육과정 및 내용 등이 유의한 변수임을 보였다. 정소영 등[17]은 고교생의 중도탈락예측을 위하여 학교생활기록부와 대면 면접을 통해 자료를 수집하였고 연관규칙과 의사결정트리의 결합 모형을 제안하였다.

사이버대학 및 이러닝 분야에서의 중도탈락 예측에 관한 연구도 국내·외에서 많이 수행되었다. 유지원[18]은 학습관리시스템(learning management system, LMS)의 로그 데이터를 분석하여 이러닝 학습자의 중도탈락을 예측하는 모형을 제시하여 96%의 높은 예측력을 보였다. 노혜란 등[19]은 대학 이러닝 강좌에서 학습자의 참여 지속 또는 중도탈락을 예측하기 위하여 로지스틱 회귀분석을 활용하여 높은 성과를 보여주었다. 하지만 직업훈련 교육 분야에서 데이터마이닝 기법을 활용하여 중도탈락자를 예측하는 연구는 국내 연구에서는 찾아볼 수 없었다.

해외 연구 또한 대부분 대학생의 중도탈락자 예측을 위주로 연구가 진행되었다. 직업훈련 관련 분야에서의 중도탈락자 예측에 대한 연구로는 로지스틱 회귀분석을 활용한 연구를 찾아볼 수 있었다[20]. 그러나 전통적인 데이터마이닝 기법 외에 텍스트마이닝 등의 최신 기법을 활용한 연구는 거의 찾아보기 어렵다.

이에 중·고교, 대학, 이러닝 등의 분야 위주로 중도탈락 예측에 관한 연구들이 주류를 이루는 상황에서, 본 연구에서는 직업훈련교육 분야에서 활용 가능한 정형 데이터와 비정형 데이터를 동시에 고려하는 기계학습 기반의 예측 모형을 제안하고자 한다.

2.3 정형 데이터와 비정형 데이터

정형데이터는 지정된 필드에 저장되어 컴퓨터 장비에 의해 바로 접근이 가능한 데이터를 의미하고, 비정형데이터는 고객의 이메일이나 웹페이지 같이 문서와 같은 형태로 저장된 자료 또는 음성, 영상 등의 데이터를 의미한다[21].

비정형 데이터는 곧바로 분석에 활용할 수 없으므로 정형화된 형태로 변환하는 전처리 작업이 필수적이다. 최근 비정형 데이터의 분석을 위하여 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 군집 분석 등의 방법론들이 주목을 받고 있다. 특히 텍스트 마이닝은 텍스트 데이터에서 자연어 처리 기술에 기반하여 유용

한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 방대한 텍스트 뭉치에서 의미있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트를 분류하는 등 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 텍스트 마이닝을 위하여 자연어를 분석하고 그 안에 숨겨진 정보를 발굴해 내기 위해 대용량 언어 자원과 통계적, 규칙적 알고리즘을 사용하고 있다. 텍스트 마이닝의 응용 분야로는 문서 분류, 문서 군집, 정보 추출, 문서 요약 등의 응용 분야 등이 있다[22].

직업훈련기관에서 사용되는 데이터를 예로 들면 학생의 나이, 교육기관과 주거지의 거리, 입학경쟁률 등의 수치형 데이터와 성별, 학력, 접수방법, 지원경로 등의 카테고리형 데이터들은 정형 데이터로 분류할 수 있다.

한편 수강생이 입력하는 훈련과정 만족도 조사 정보, 훈련교사가 입력하는 상담일지 등의 텍스트 자료들은 비정형데이터로 분류된다.

2.4 워드 임베딩

2.4.1 워드 임베딩

워드 임베딩(word embedding)이란 텍스트 마이닝이나 자연어 처리 분야에서 광범위하게 활용되고 있는 기법으로 하나의 단어를 인공 신경망을 이용하여 벡터 공간상에 나타낼 수 있는 숫자값으로 변환하는 작업을 의미한다[23]. 기존에는 단어 자체를 아스키코드나 유니코드로 처리해서 사용했지만 이러한 숫자 코드만을 가지고는 단어의 실제적인 의미를 추론하기는 어렵다. 예를 들어 왕과 여왕이 관련이 있는 단어이고 왕의 성별은 남성이고 여왕의 성별은 여성이라는 사실을 기존의 숫자 코드만을 가지고는 알아내기 어렵다.

2.4.2 Word2vec

워드 임베딩 방법 중의 하나인 Word2vec은 신경망 기반의 워드 임베딩 방법으로, 주변 단어를 보면 그 단어를 알 수 있다는 개념에서 착안한 알고리즘으로 기본적으로 유사한 단어는 같은 문맥에서 등장하고 유사한 분포를 가진다는 가정을 기반으로 하고 있다[24-26]. 이러한 가정 하에 Word2vec은 신경망을 이용하여 단어간 유사도를 표현할 수 있도록 각 단어 자체의 의미를 벡터로 표현하는 방법론으로 이를 응용하면 복잡한 개념 표현은 물론 단어들 간의 관계를 추론할 수 있다[27][28].

Word2vec 기법으로 각 단어를 벡터로 표현하는 방법은 CBOW(continuous bag-of-words)와 Skip- gram의 2가지 기법으로 나눌 수 있다. CBOW는 주변의 단어들로부터 가운데 단어의 출현을 예측하는 방법이다. Skip-gram은 가운데 단어로 주변 여러 단어들의 출현을 예측하는 모델을 만들어 학습한 뒤 단어에 할당된 가중치 값을 해당 단어의 연속적 벡터 값으로 사용한다[29][30].

2.4.3 Doc2vec

문서나 문장에 대한 학습 방법으로는 대표적으로 Mikolov가 제안한 Doc2vec 모델이 있다. Doc2vec 알고리즘은 각각의 문서를 하나의 벡터로 표현하며, 각각의 벡터들은 문서나 문장 내에 단어들을 예측하기 위해 훈련되어진다. Word2vec은 개별 단어의 관계를 학습하는데 주안점을 둔 알고리즘이고 Doc2vec은 문장, 단락, 문서와 같은 더 큰 블록에 대한 연속표현을 비지도 학습으로 모델을 생성하는 알고리즘이다[31].

Word2vec에서는 각각의 단어의 순서와 의미를 무시하지만, Doc2vec은 단어의 순서와 의미를 포함한 벡터 표현을 생성해 낸다[25]. 즉, 각각의 문서에 같은 단어를 사용하지만 단어의 순서가 다르면 다른 벡터를 만들어 내고, 비슷한 의미를 가진 각각의 문서들을 벡터공간상 근거리에 위치하도록 벡터를 만들게 된다. 이렇게 만들어진 벡터를 이용하여 문서 분류를 하면 Word2vec 모델을 적용했을 때보다 더 좋은 성능을 얻을 수 있다[32].

2.5 합성곱 신경망

본 연구에서는 비정형데이터를 분석하기 위하여 합성곱 신경망(Convolutional Neural Network, CNN)을 활용하였다. 합성곱 신경망은 패턴이나 물체를 인식하는 생물의 시각처리과정을 모방한 모형으로 이미지를 작은 구역으로 나누어 부분적인 특징을 인식하고 이것을 결합하여 전체를 인식하는 방법이다.

합성곱 신경망은 하나의 입력계층과 출력계층, 하나 이상의 합성곱 계층(convolution layer)과 풀링 계층(pooling layer) 그리고 완전 연결 계층(fully connected layer)으로 구성되어 있다. 입력층을 통해 입력하고자 하는 이미지를 입력하고, 합성곱 계층을 통해 필터링되어 적절한 특징을 추출한다. CNN은 이미지 내에서 특징점의 위치가 달라질 때도 효과적으로 대응할 수 있다. 예를 들어 이미지의 각도가 바뀌거나 회전이 되거나 할 경우 일반적인 기계학습으로는 전혀 다른 패턴으로 인식이 되는 경우가 많으나 CNN에서는 부분적인 특징으로부터 풀링 계층을 거쳐서 이미지의 특징들을 잘 추출해낼 수 있으므로 일반적인 기계학습보다 뛰어난 성능을 발휘한다[33].

이미지 인식 분야 뿐 아니라 텍스트 문장을 단어 벡터들의 순열로 표현하여 CNN 모델에 입력하여 효과적으로 분류할 수 있음을 선행 연구들이 보여주고 있다[2][34]. 텍스트는 비정형데이터이므로 CNN 알고리즘에 곧바로 입력하여 분석할 수 없다. 먼저 분석 대상 텍스트들을 활용하여 Word2vec, Doc2vec 등의 모델을 만든다. 그 후 분석 대상 텍스트 중에서 불필요한 숫자, 문장부호, 특수 기호 등을 삭제하고 텍스트를 전단계에서 생성된 Word2vec, Doc2vec 모델에 입력하여 벡터로 변환한 후 CNN 모델에 입력한다. 학습용 데이터셋을 이용하여 학습을 실시한 후 검증용 데이터셋으로 모델의 성능을 최종적으로 평가할 수 있다.

본 연구에서 사용한 텍스트 데이터를 워드 임베딩 기법을 활용하여 기계학습을 위한 입력신호로 변환하기 위한 방법론으로 전술한 Word2vec과 Doc2vec 기법을 활용한 예시를 [그림 1]에 제시하였다.

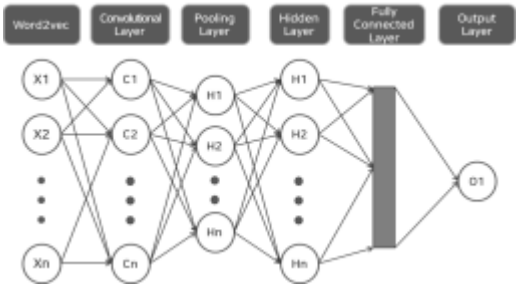

그림 1. Word2vec을 활용한 CNN 모형 예시

III. 제안 모형

본 연구에서 제안하는 정형 데이터와 비정형 데이터를 동시에 고려하는 기계학습 기반 예측 모형의 전반적인 구조는 [그림 2]와 같다.

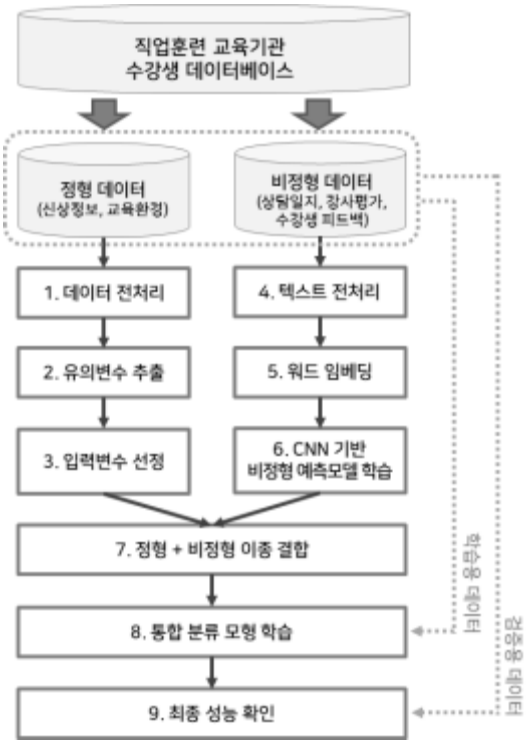

그림 2. 제안 모형의 구조 체계

수집된 데이터 원본은 학생 신상정보, 교육환경 등의 내용으로 구성된 정형 데이터와 훈련교사가 입력하는 상담내용, 수강생이 입력하는 강사평가, 수강생 피드백 등의 내용으로 구성된 비정형 데이터로 구성되어 있다.

1단계로 데이터 전처리 과정에서는 결측값, 이상치 등을 확인하여 삭제하거나 평균값 등으로 대체하는 과정을 거친다. 그리고 변수들의 단위가 다르거나 변수의 분포가 편향되어 있을 경우 scale, log 함수 등을 이용하여 정규화시킨다. 그리고 예측 모델의 정확도를 향상시키기 위해 기존의 변수들로부터 다양한 파생변수를 만들 수 있다.

2단계에서는 전 단계에서 만들어진 변수들을 대상으로 t검정, 카이제곱검정 등의 검정방법을 활용하여 통계적으로 유의한 변수들을 선택한다.

3단계에서는 다양한 기계학습 방법으로 실험하기 위하여 전 단계에서 만들어진 변수들을 독립변수와 종속변수로 나눈다.

4단계는 비정형 데이터인 텍스트를 전처리하는 단계이다. 분석 대상 텍스트 중에서 불필요한 숫자, 문장부호, 특수기호 등을 삭제하고 지도학습을 위하여 라벨링을 수행한다. 라벨링은 샘플의 분류 유형에 따라 기계학습에 적합한 형식으로 one-hot vector로 표현한다.

5단계로 텍스트 데이터를 CNN에서 분석하기 위해서 벡터 형태로 변환하는 과정을 거친다. 이 단계에서는 내부 도메인 또는 외부 도메인에서 수집된 텍스트를 이용하여 사전에 만들어진 Word2vec, Doc2vec 등의 모델을 활용하여 각각의 텍스트 데이터 표본을 워드 임베딩하여 벡터로 변환한다.

6단계에서는 워드 임베딩으로 변환된 데이터를 CNN으로 분석하여 기계학습을 수행하며, 각각의 샘플로부터 0.0 ~ 1.0 사이의 결과값을 추출한다. CNN의 출력값은 1개 클래스의 경우 0과 1로 출력할 수도 있고 0.0에서 1.0 사이의 값으로도 출력이 가능하다. 텍스트 분석의 정확도 향상을 위해 본 연구에서는 0과 1로 이분류된 정보보다는 좀 더 풍부한 정보를 기계학습 모델에 전달하기 위해 0.0에서 1.0 사이의 값을 선택하였다.

7단계에서는 3단계에서 선정된 정형데이터의 입력변수들과 6단계에서 얻어진 비정형데이터의 결과값을 이종결합하여 최종 통합 데이터 셋을 생성한다. 예를 들어 [그림 3]에 제시한 것처럼 정형데이터의 변수가 12개라면 비정형데이터를 CNN으로 분석한 0.0에서 1.0 사이의 예측 결과값을 정형데이터의 필드에 추가하여 총 13개의 필드를 만들게 된다.

8단계에서는 정형데이터와 비정형데이터가 통합된 데이터셋을 대상으로 다양한 알고리즘을 적용하여 기계학습을 수행한다.

마지막 9단계에서는 학습용 데이터로 만들어진 통합 분류 모형의 성능을 확인하기 위하여 검증용 데이터셋으로 통합 분류 모형의 성능을 최종적으로 확인한다.

그림 3. 정형 데이터와 비정형 데이터의 결합 방식

IV. 실증분석

4.1 자료의 수집 및 처리

본 연구에서는 제안 모형의 성능과 유용성을 검증하기 위해 서울 소재 A직업훈련기관의 IT분야 국가기간전략훈련 수강생들의 실제 사례에 적용시켜 보았다. 실험용 데이터는 2015.1.1.부터 2018.9.30.까지 9개월, 1,440시간의 장기교육과정 수강생 767명을 대상으로 수집하였다. 이 중 수료자는 605명이며 중도탈락자는 162명이었다.

분석을 위해 훈련기관의 인트라넷에 입력된 훈련생들의 인구통계학적인 정보 및 훈련생들이 입력하는 훈련생 만족도 조사 자료, 훈련교사가 입력하는 상담일지 등의 데이터를 활용하였다.

데이터 수집을 위해 Python의 BeautifulSoup 패키지를 활용하여 작성된 코드로 자동화된 웹스크래이핑(Web scraping) 과정을 거쳐 mysql 서버에 저장하였다. 수집된 데이터를 가공하고 전처리하는 작업을 위해 자체 개발한 자바(java) 프로그램을 사용하였다.

훈련교사가 입력하는 상담일지는 9개월 교육과정 중 매월 1회 이상씩 입력되어 학생당 평균 10회 정도 입력되어 있다. 사전 예측이 목적이므로 전체 9개월의 교육기간 중에서 입학 후 5개월 시점까지의 상담일지만을

분석에 사용하였다. 그리하여 분석 대상이 된 상담일지는 수강생 767명으로부터 수집된 정형, 비정형 데이터 총 8,608건이며 그 중에서 중도탈락자의 샘플 수는 1,767건, 수료자의 샘플 수는 6,841건이다. 정확한 예측을 위하여 수료자와 중도탈락자의 샘플비율을 1:1로 맞추어 각각 1,755건씩, 총 3,510건을 대상으로 분석을 수행하였다. 이어 전체 실험용 데이터셋을 학습용 80%, 검증용 20%의 비율로 구분하여, 학습용 데이터셋을 통해 구축된 모형의 성능을 검증용 데이터셋을 통해 검증하고자 하였다. 전체적인 데이터셋의 구성은 다음의 [표 2]와 같다.

표 2. 실험 데이터셋 구성

구분	전체 데이터셋	실험에 사용한 데이터셋	학습용 데이터셋	검증용 데이터셋
중도탈락	1,767	1,755	1,404	351
수료	6,841	1,755	1,404	351
합계	8,608	3,510	2,808	702

그 밖에 학생이 입력하는 데이터로는 훈련과정 만족도 조사 데이터가 있다. 9개월간의 교육 기간 중 매월 1회 이상 실시하는 훈련과정 만족도 조사에 훈련생들이 입력한다. 주된 내용은 강사의 교육 내용에 대한 평가, 교육과정에 대한 만족도와 훈련생의 애로사항 등을 파악하기 위한 질문들로 구성되었다. 1~15번 문항은 5점 척도의 정형 데이터이고 16~18번은 텍스트 데이터인데 이 중에서 1~15번 문항의 응답 내용을 수치로 변환하여 각 훈련과정의 평가점수에 반영하였다.

훈련과정 만족도 조사 자료는 훈련과정별로 월 1회 실시되는데 훈련생의 이름은 무기명으로 처리되는 자료이다. 1번에서 15번의 정형데이터는 5점 척도 자료이므로 각 번호에 숫자를 곱하여 100점 만점으로 계산하여 훈련생이 소속된 훈련과정의 만족도 점수를 계산하는데 사용하였다. 본 연구에서는 16번에서 18번의 텍스트 자료는 분석에 사용하지 않았고 1번에서 15번의 정형데이터만 사용하였다.

본 연구는 미래를 예측하는 연구이므로 훈련과정 중반 이후의 데이터는 제외하고 총 9개월 중 5개월까지의 데이터만 반영하였다.

수강생들의 학력은 고졸, 전문학사, 학사로 구분했으며 학력 정보가 미입력된 데이터들은 고졸로 처리하였다. 중졸, 고교재학, 전문대중퇴, 4년제중퇴는 고졸로 입력하였고, 석사 3명은 학사로 전처리하였다. 나이가 입력되지 않은 레코드는 평균값으로 대체하였다.

거주지 주소로부터 교육기관과의 거리는 네이버 지도 API와 구글 지오코딩 라이브러리를 이용하여 거주지 주소 텍스트를 읽은 후 위도, 경도 좌표를 얻은 후 교육기관의 좌표와 직선 거리를 계산하도록 자체적으로 프로그래밍한 자바 코드를 활용하였다.

훈련과정만족도는 훈련생이 매월 입력하는 훈련과정 만족도 조사 내용 중 5점 척도로 입력되는 1~15 문항의 입력 내용을 100점 만점으로 수치화한 점수이다. 본 연구에서는 미래를 예측하는 연구이므로 전체 9개월 과정 중 5개월 시점까지의 만족도 조사 입력 내용만을 대상으로 처리하였다.

최종적으로 실험에 사용된 후보 변수는 총 40개로 이 중 중도탈락그룹과 수료생을 구별하는 변수(중도탈락 여부)가 종속변수로 사용되었고 나머지 38개의 정형 변수와 훈련교사가 매월 입력하는 텍스트 데이터인 상담일지 내용은 비정형 변수로 사용되었다. 다음의 [표 3]에 본 연구에서 사용된 39개 정형 변수들의 목록이 제시되어 있다.

데이터마이닝에 사용되는 데이터에는 많은 입력변수가 존재하는 것이 일반적이다. 이러한 경우 예측의 정확도를 높이기 위해 사전에 목표변수와 무관하다고 파악되는 입력변수들의 제거 작업이 선행되어야 한다. 이를 위해 본 연구에서는 독립표본 t-검정과 카이제곱검정, 그리고 로지스틱 회귀분석의 변수선택 기능을 활용하여 총 38개의 정형변수 가운데 통계적으로 유의한 총 12개 변수를 최종적인 모형의 입력변수로 선정하였다. 최종 선정된 정형 입력변수는 다음의 [표 4]와 같다.

표 3. 전체 정형 데이터 변수 목록

연번	변수	설명
1	나이	나이(세)
2	성별	0 여성, 1 남성
3	20대 남성여부	0 비해당, 1 해당
4	30대 남성여부	0 비해당, 1 해당
5	20대 여성여부	0 비해당, 1 해당
6	30대 여성여부	0 비해당, 1 해당
7	고졸여부	0 비해당, 1 해당
8	전문대졸여부	0 비해당, 1 해당
9	대졸여부	0 비해당, 1 해당
10	고졸남성여부	0 비해당, 1 해당
11	전문대졸남성여부	0 비해당, 1 해당
12	대졸남성여부	0 비해당, 1 해당
13	고졸여성여부	0 비해당, 1 해당
14	전문대졸여성여부	0 비해당, 1 해당
15	대졸여성여부	0 비해당, 1 해당
16	훈련직종코드	0 디자인직종 1 프로그래밍 직종
17	교육기관과 주거지의 거리	거리(km)
18	서울거주여부	0 비해당, 1 해당
19	아파트거주여부	0 비해당, 1 해당
20	대도시거주여부	0 비해당, 1 해당
21	거주지주택평균가격	금액(천원)
22	접수방법(인터넷)	0 비해당, 1 해당
23	접수방법(모바일)	0 비해당, 1 해당
24	접수방법(방문)	0 비해당, 1 해당
25	지원경로(인터넷)	0 비해당, 1 해당
26	지원경로(페이스북)	0 비해당, 1 해당
27	지원경로(인스타그램)	0 비해당, 1 해당
28	지원경로(유튜브)	0 비해당, 1 해당
29	지원경로(고용지원센터)	0 비해당, 1 해당
30	지원경로(지인소개)	0 비해당, 1 해당
31	지원경로(취업성공패키지)	0 비해당, 1 해당
32	지원경로(HRD사이트)	0 비해당, 1 해당
33	지원경로(키워드검색)	0 비해당, 1 해당
34	지원경로(지식인)	0 비해당, 1 해당
35	지원경로(이벤트)	0 비해당, 1 해당
36	강의실창문여부	0 창문없음, 1 창문있음
37	입학경쟁률	입학정원 대비 지원자 비율(%)
38	훈련과정만족도	5점 척도 입력 내용을 100점 만점으로 수치화한 값
39	중도탈락여부	0 수료, 1 중도탈락

표 4. 최종적으로 선택된 12개의 정형 입력변수

연번	변수명	연번	변수명
1	성별	7	인터넷접수여부
2	고졸여부	8	모바일접수여부
3	고졸여성여부	9	지원경로(키워드검색)
4	전문대졸여성여부	10	강의실창문유무
5	훈련직종코드	11	입학경쟁률
6	교육기관과 주거지의 거리	12	훈련과정만족도

4.2 실험 설계

본 연구에서는 데이터 종류에 따라 텍스트를 제외한 정형 데이터만으로 실험, 비정형 데이터(상담일지 텍스트)만을 대상으로 CNN 분석, 정형데이터와 비정형데이터를 이종결합하여 분석하는 3가지 실험으로 진행되었다. 다음의 [표 5]는 본 연구의 실험에서 사용된 소프트웨어들의 목록을 제시하고 있다.

표 5. 실험에 사용된 소프트웨어 목록

구분	소프트웨어
데이터수집	BeutifulSoup 4.4.0
데이터전처리, 지오코딩	mysql 5.7.21, Java10, Eclipse 4.8.0, 네이버 지도 API, Google Geocoder
정형 데이터 분석	SPSS
Word2vec 비정형 데이터 분석	Python 3.6.6, Tensorflow 1.10.0, Keras 2.2.2, KoNLPy, gensim

비정형데이터 분석은 CNN을 사용하여 실험을 진행했다. 본 연구에서 적용된 CNN 실험의 설정값들은 다음의 [표 6]과 같다.

표 6. CNN 설정값

Layer	Size	Activation Function
Input Layer	N×262×500×1	
Convolution Layer	3×3×1 Convolution filter	ReLU
Pooling layer	2×2 Max pooling filter	
Hidden layer	31,993 Nodes	
Fully Connected layer	128 Nodes	
Output layer	1 Node	Sigmoid

본 연구에 적용한 CNN 실험의 설정값에 대하여 자세히 설명하면 아래와 같다.

첫번째로 Input Layer의 사이즈는 N×262×500×1로 설정하였는데 이 의미는 샘플의갯수×문장의단어갯수×벡터의차원×채널수이다. 본 연구에서 사용된 텍스트 데이터는 문장별로 벡터로 변환하였는데 한 문장에 최대 262개의 단어가 포함되어 있다. 단어의 수가 많은 문

장도 있고 단어의 수가 적은 문장도 있는데 기계학습을 위해서는 가로,세로 사이즈를 정확히 맞추어야 한다. 따라서 가장 단어의 수가 많은 문장을 기준으로 262로 설정하고 단어의 수가 적은 문장은 0으로 채우는 패딩(padding) 작업을 하게 된다. 벡터의 차원은 본 연구에서 실험한 100차원, 300차원, 500차원 중에서 가장 좋은 결과를 도출한 500차원으로 설정하였다. 채널수는 이미지 분석의 경우 흑백 이미지는 1로 컬러 이미지는 빛의 3원색인 RGB(Red, Green, Blue)를 고려하여 3으로 설정하게 되는데 본 연구에서는 텍스트 분석이므로 1로 설정하였다.

두번째로 Convolution Layer에서는 가로 3, 세로 3 사이즈의 필터 1개를 사용하였다. 활성화 함수에는 탄젠트(thanh) 함수, 시그모이드(sigmoid) 함수, ReLU 함수 등이 있다. 본 연구에서는 탄젠트 함수나 시그모이드 함수보다 학습시간이 빨라서 최근 많이 사용되고 있는 ReLU(Rectified Linear Unit)를 사용하였다[35].

세번째로 Pooling Layer에서는 가로 2, 세로 2 사이즈의 Max pooling filter를 사용하였다. Max pooling은 현재 대부분의 CNN 모델에서 사용하고 있는 방법으로서 Convolution Layer에서 나온 값 중에서 가장 큰 값을 선택하는 방법이다[36].

네번째로 Hidden Layer에서는 총 31,993개의 노드를 사용하였다.

마지막으로 Output Layer에서는 시그모이드 함수를 이용하여 0.0에서 1.0 사이의 값을 중도탈락에 대한 예측값으로 출력하였다.

비정형 데이터 분석은 코퍼스(Corpus)와 기법, 도메인, 차원에 따라 구성을 달리하여 14가지 경우로 진행되었다. 코퍼스는 조사, 어미, 문장부호를 제외한 경우(Corpus-부분)와 모두 포함한 경우(Corpus-전체)의 2가지 케이스로 나누어 실험했다. 조사, 어미, 문장부호 등은 중요도가 낮기 때문에 제외하고 분석하는 것이 일반적인데 모든 어휘를 다 포함한 경우와 얼마나 성능 차이가 나는지 확인하기 위하여 2가지 케이스로 나누었다.

워드 임베딩 기법은 벡터 모델에 따라 Word2vec, Doc2vec의 2가지로 나누어 진행하였다. 내부 도메인으로는 본 연구에서 사용한 상담일지 데이터를 활용하여 각각 100차원, 300차원, 500차원으로 3가지 벡터 모델을 만들어서 실험을 진행하였다.

Word2vec 모델은 파이썬의 konlpy 패키지를 이용하여 만들었는데 그 과정을 설명하면 다음과 같다. 상담일지 전체 텍스트 자료를 문장별로 나눈 후 문장별로 조사, 어미, 문장부호를 제외한 경우와 모두 포함한 경우 2가지로 나누어 단어 사전을 만들었다. Word2vec의 차원은 선행연구를 참조하여 100차원, 300차원, 500차원으로 구분하였으며[2] 앞뒤 10개의 주변 단어를 문맥으로 보고 예측 모델에 반영하였다. 벡터의 차원은 벡터 모델에서 세로 사이즈를 결정하게 되며 벡터가 커지면 그만큼 실험 속도가 느려지게 된다. 출현 빈도가 2회 미만인 단어는 제외하였고 Word2vec 알고리즘 중 Skip-gram 알고리즘을 사용하였다. [그림 4]에 word2vec으로 만들어진 실제 벡터의 예를 표시하였다.

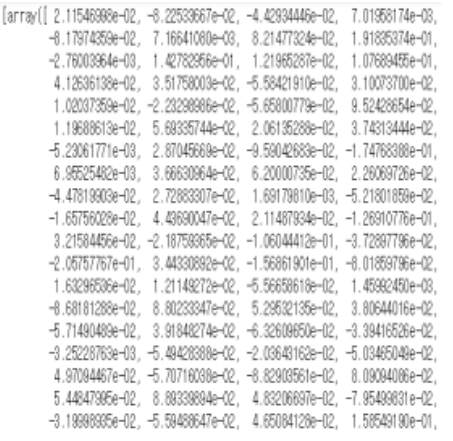

그림 4. word2vec으로 만들어진 벡터 예시

Doc2vec 모델도 100차원, 300차원, 500차원으로 나누어 만들었으며 주변 단어를 앞뒤 300개 문맥을 살펴서 모델에 반영하였다. 출현 빈도가 1회 미만인 단어는 제외하였다.

벡터 모델을 만들 때 내부 도메인의 데이터인 상담일지 텍스트만으로는 어휘가 제한적일 수 있으므로 실험결과 비교를 위하여 외부 도메인의 텍스트로 만든 모델을 사용하였다.

외부도메인으로 사용한 벡터 모델은 나무위키와 위키피디아의 문장을 기반으로 학습된 깃허브에 공개되어 있는 모델이며, 파일 사이즈는 약 86MB이다[37]. 공개되어 있는 모델은 다운로드 받아 파이썬 코드에서 바로 불러와서 텍스트 데이터의 벡터를 만들 수 있는 형태로 배포되어 있다.

4.3 실험 결과

4.3.1 정형 데이터만 활용한 모형의 실험 결과

본 연구에서는 로지스틱 회귀 분석(Logistic Regression, LOGIT), 다중판별분석(Multiple Discriminant Analysis, MDA), 회귀 및 분석 트리(Classification And Regression Tree, CART), 인공신경망(Artificial Neural Network, ANN), 서포트 벡터 머신(Support Vector Machine, SVM), 랜덤 포레스트(Random Forest, RF)등 총 6가지 방법의 데이터마이닝 분석 기법을 활용하여 정형 데이터를 분석했다.

정형데이터의 실험 결과가 다음의 [표 7]에 제시되어 있다. 이 표에서 볼 수 있듯이, 기계학습 기법 중 SVM과 RF에서 가장 우수한 예측 정확도가 산출됨을 확인할 수 있다. 특히 SVM에서 RF보다 검증용 데이터셋 기준으로 소폭 더 우수한 예측 정확도를 산출함을 확인할 수 있다.

표 7. 정형 데이터 모형의 예측 정확도

모형	학습용	검증용	최적 모형의 실험설정
LOGIT	67.50%	66.40%	Forward Selection (총 12개 변수 선정)
MDA	67.13%	66.38%	Enter (총 12개 변수 사용)
CART	69.10%	70.40%	Gini index
ANN	78.40%	78.06%	은닉층의 노드 수=26
SVM	94.91%	93.30%	RBF 커널, C=78, σ 2= 1
RF	94.91%	93.16%	max_depth=15, mtry=3, # of trees=1,000

LOGIT의 경우 Forward Selection 옵션으로 설정하여 최종 선정된 12개의 정형변수들을 하나씩 추가하면서 학습을 수행하였다. MDA의 경우 Enter 방식으로 12개의 변수들을 사용하여 분석하였다. CART의 경우 모든 경우 중에서 특정 class에 속하는 관측치의 비율을 모두 제외하는 방식인 지니계수(Gini index) 방식을 선택하여 분석하였다. ANN에서는 은닉층의 노드수를 26개로 설정하였다. SVM에서는 가우시안 커널로도 불리우는 RBF(Radial Basis Function) 커널을 사용하여 차원이 무한한 특성 공간에 매핑하는 방식을 선택했으며 C(Cost)는 78로 설정했으며 RBF 커널에서 사용되는 매개변수 gamma는 1로 설정하였다. RF에서는 1,000개의 트리를 사용했으며 최대 깊이(max depth)는 15, 각 샘플에서 이용한 변수의 갯수인 mtry는 3으로 설정하였다.

정형 데이터의 학습은 LOGIT, MDA, CART, ANN, SVM의 경우 SPSS를 사용하여 수행하였으며 RF의 경우 파이썬의 사이킷런 패키지를 이용하여 수행하였다. 정형데이터의 분석 절차는 다음과 같다. 먼저 분석 대상 데이터 3,510건에 대하여 랜덤 샘플링을 하여 학습용 데이터셋과 검증용 데이터셋을 80:20으로 구분했다. 학습용 데이터셋으로 학습을 실시하여 모델을 만들고 검증용 데이터셋으로 모델의 예측 정확도를 측정하였다. 모든 실험에는 정확한 비교 분석을 위하여 동일한 학습용 데이터와 동일한 검증용 데이터를 사용하였다.

표 8. 로지스틱 회귀분석 상세 결과

변수	가중치	유의수준	의미
성별	-.432	.018	
고졸여부	.392	.000	고졸자는 중도탈락률이 높다.
고졸여성여부	.423	.080	
전문대졸여성여부	-.997	.000	전문대를 졸업한 여성의 수료율이 높다.
훈련직종코드	-.992	.000	프로그래밍 직종이 디자인직종보다 수료율이 높다.
교육기관과 주거지의 거리	.007	.002	교육기관과 주거지의 거리가 멀수록 중도탈락률이 높다.
인터넷접수여부	20.767	.999	
모바일접수여부	21.076	.999	
지원경로(키워드검색)	2.957	.000	키워드검색으로 지원한 학생은 중도탈락률이 높다.
강의실창문유무	-.886	.000	강의실에 창문이 있으면 수료율이 높다.
입학경쟁률	.067	.006	
훈련과정만족도	-.035	.000	훈련과정만족도가 높으면 수료율이 높다.

상기 모형들 중에서 LOGIT/MDA 등 통계 기반 모형의 경우, 각 입력변수가 종속변수(즉, 중도탈락여부)에 어떤 영향을 미쳤는지 사후 분석이 가능하다. 다음의 [표 8]은 이 중 LOGIT 분석의 상세 결과와 그 함의를 제시하고 있다. 이 결과를 통해 미루어 볼 때, 전반적으로 LOGIT 분석의 결과는 합리적으로 해석 가능한 결과가 산출되었음을 알 수 있다.

4.3.2 비정형 데이터만 활용한 모형의 실험 결과

비정형 데이터인 상담일지 텍스트를 Word2vec과 Doc2vec을 이용하여 벡터로 변환하여 CNN 모델에 입력하여 예측 정확도를 측정해 보았다. 사전에 분류한 80%의 학습용 데이터셋을 이용하여 학습을 실시하였으며 학습률(learning rate)은 0.001로 설정하였으며 실험 반복 횟수(epoch)는 50회로 설정하였다. 한번에 처리하는 벡터의 갯수인 batch size는 15로 설정하였는데 본 연구에서는 학습용 데이터셋의 샘플수가 적은 관계로 작게 설정하였다. 실험 조기종료 콜백함수를 사용하여 이전 epoch 때와 비교하여 학습오차가 3회 이상 증가하면 반복 횟수가 50회가 되기 이전에도 학습을 조기에 중단하도록 설정하였다. 나머지 20%의 검증용 데이터셋을 학습이 완료된 모델에 입력하여 손실률과 정확도를 측정하였다.

비정형 데이터만 활용한 모형의 실험에서도 정형 데이터 실험과의 정확한 비교 분석을 위하여 정형 데이터 실험에서 사용한 동일한 학습용 데이터와 동일한 검증용 데이터를 사용하였다.

[표 9]는 비정형 데이터만 활용하여 CNN으로 학습한 모형의 전반적인 실험결과를 제시하고 있다. 이 표에서 볼 수 있듯이 전반적으로 Word2vec보다는 Doc2vec 모델이 더 우수한 성능을 보였다. 특히 내부도메인, 500차원, 전체 코퍼스를 활용한 모형이 학습용 90.03%, 검증용 86.61%로 가장 좋은 결과를 산출하였다. 참고로 내부 도메인은 100차원, 300차원, 500차원으로 실험을 하였으나 외부 도메인의 벡터는 직접 데이터를 수집한 것이 아니라 인터넷에서 구한 데이터 자체가 300차원으로 된 모델만 제공되는 상태였기에 300차원만을 대상으로 실험을 수행하였다.

표 9. 비정형 데이터 모형의 예측 정확도

벡터모델	도메인	차원	Corpus	학습용	검증용
Word2vec	내부	100	부분	87.68%	85.04%
			전체	89.21%	83.90%
		300	부분	83.73%	82.05%
			전체	91.24%	84.47%
		500	부분	87.61%	83.33%
			전체	89.14%	84.19%
	외부	300		84.08%	83.05%
Doc2vec	내부	100	부분	94.52%	86.47%
			전체	95.01%	85.75%
		300	부분	87.82%	85.19%
			전체	87.39%	85.61%
		500	부분	87.68%	86.47%
			전체	90.03%	86.61%
	외부	300		91.95%	84.19%

4.3.3 제안 모형의 실험결과

제안 모형은 앞의 4.3.2의 결과로 추출된 CNN 모형의 결과값(0.0~1.0 사이의 값)을 기존 12개의 정형 입력변수에 추가하여, 총 13개의 입력변수로 실험을 실시하였다. 그렇게 함으로써, 정형 데이터와 비정형 데이터를 하나의 모형에서 모두 반영하여, 중도탈락에 대한 예측 정확도를 제고하고자 하였다.

다음의 [표 10]은 제안 모형의 실험결과를 제시하고 있다. [표 7]에 제시된 정형 데이터(12개 변수)만 활용한 실험결과와 기법 별로 비교해 보면, 비정형 데이터를 함께 고려했을 때 최대 약 20%(LOGIT 모형의 경우, 66.4% → 85.9%로 개선)까지 예측정확도가 향상됨을 확인할 수 있다. 아울러, 전반적인 예측정확도는 정형 데이터와 비정형 데이터를 모두 고려하여 SVM을 적용했을 때, 검증용 데이터셋 기준으로 95.73%의 예측 정확도를 나타내 가장 우수한 것으로 나타났다.

표 10. 제안 모형 실험 결과

모형	학습용	검증용	최적 모형의 실험설정
CNN+ LOGIT	91.42%	85.90%	Enter(총 13개 변수 사용)
CNN+ MDA	90.49%	86.18%	Enter(총 13개 변수 사용)
CNN+ CART	90.95%	84.90%	Gini index
CNN+ ANN	92.50%	85.90%	은닉층 노드의 수=28
CNN+ SVM	98.82%	95.73%	RBF 커널, C=10, σ 2= 1
CNN+ RF	99.96%	88.89%	max_depth=15, mtry=3, # of trees=1,000

4.3.1~4.3.3절의 실험결과 중, 각 경우별 가장 우수한 성과를 보인 대표 모형의 예측정확도를 정리한 결과가 [표 12]에 제시되어 있다. 전반적으로 봤을 때, 정형 데이터와 비정형 데이터를 모두 사용한 CNN+SVM의 성능이 제일 우수하고, 두 번째로 정형데이터만 사용한 SVM 모형이, 마지막으로 비정형데이터만 사용한 CNN 모형이 가장 낮은 예측 정확도를 보이는 것으로 확인되었다.

본 연구에서는 일반적인 정형데이터 분석 외에 비정형데이터를 동시에 고려하는 것이 중도탈락 예측의 정확도를 높일 수 있음을 입증하고자 하였다. 본 연구에서 사용한 데이터 중에서 정형 데이터만을 볼 때 훈련생의 중도탈락을 예측하기에는 정보가 부족한 편이다. 예를 들면 학생의 출석관련 정보들, 과목별 점수 등의 정량적인 필드가 추가로 필요하다. 또한 학습진도에 대한 이해도, 교육에 대한 태도 등 정성적인 부분들도 본 연구에서 사용한 정형 데이터에 포함되어 있지 않다. [표 10]에 나타난 실험 결과를 볼 때 정형데이터에서 제공하는 정보가 제한적인 경우 비정형 데이터를 추가로 분석하는 것이 예측 정확도를 제고할 수 있음을 시사하는 것이라고 생각된다.

교육 현장에서는 정형 데이터 외에도 많은 비정형데이터들을 사용하고 있다. 직업 훈련 교육의 경우를 예로 들면 훈련일지에 훈련생들의 지각, 결석 사유들을 입력하게 된다. 또한 과목별 평가 결과에 대한 교사의 피드백, 상담일지 등 텍스트가 포함된 많은 행정 서류들을 만들고 관리하게 된다. 이러한 텍스트 데이터에는 정형 데이터로 표현하기 어려운 부분들을 보완할 수 있는 중요한 정보들을 포함하고 있는 경우가 많다. 이러한 많은 문서들을 활용하여 훈련생의 중도탈락을 예측하는 용도로 활용할 수 있다면 직업훈련의 성과를 개선하는데 도움이 될 수 있으리라 생각된다.

표 11. 실험결과 요약

모형	예측정확도 (검증용 데이터 기준)	특징
SVM	93.30%	정형데이터만 사용
CNN	86.61%	비정형데이터만 사용
CNN+SVM	95.73%	정형+비정형데이터 이종결합

마지막으로 [표 11]에 제시된 모형 간 예측정확도 차이가 통계적으로 유의한 지 검증하기 위해 다수의 기존 연구에서 이미 활용된 바 있는 비모수 통계검정 기법인 McNemar 검정을 적용해 보았다[38][39]. 다음의 [표 12]는 McNemar 검정의 결과를 제시하고 있다. 이 표에 제시된 바와 같이, 본 연구의 제안 모형인 CNN+SVM은 비교모형인 CNN과 대하여 99% 신뢰수준 하에서, SVM에 대하여 95% 신뢰수준 하에서 통계적으로 유의한 성과 차이를 보이고 있음을 확인할 수 있다.

표 12. McNemar 검정 결과

	CNN	CNN+SVM
SVM	16.66**	5.02*
CNN		48.402**

*95% 신뢰수준 하에서 유의, **99% 신뢰수준 하에서 유의

5. 결 론

직업훈련 현장에서 느끼는 가장 큰 어려움 중 하나는 중도탈락 문제이다. 해마다 많은 수의 학생들이 중도탈락을 하게 되어 국가 예산 낭비를 초래하고 있으며 청년 취업률을 개선하고자 하는 노력에 장애 요인이 되고 있다. 본 연구에서는 중도탈락의 원인을 주로 분석해 온 기존 연구들과 달리, 각종 수강생 정보를 활용하여 사전에 중도탈락을 예측할 수 있는 기계학습 기반 모형을 제안하였다. 특히 본 연구의 제안모형은 수강생 관련 정형 데이터 뿐 아니라 비정형 데이터인 훈련교사가 작성한 상담일지 정보까지 동시에 고려하여 모형의 예측정확도를 제고하고자 하였다. 이 때 비정형 데이터에 대한 분석은 최근 주목받고 있는 텍스트 분석 기술인 Word2vec과 CNN을 이용해 수행하였다. 서울 소재 한 직업훈련기관의 실제 데이터에 제안 모형을 적용해 본 결과, 정형 데이터만을 사용하여 중도탈락을 예측할 때보다 비정형 데이터를 함께 고려했을 때 예측의 정확도가 최대 약 20%까지 향상됨을 확인할 수 있었다. 아울러, SVM을 기반으로 정형 데이터와 비정형 데이터를 결합해 분석했을 때, 검증용 데이터셋 기준으로 약 96%

의 높은 예측 정확도를 나타내는 모형이 산출됨을 확인할 수 있었다.

본 연구의 학술적인 의의는 다음과 같다.

첫째, 중도탈락의 요인만을 위주로 제시한 기존 연구들과 달리 본 연구는 직업훈련분야에서 기계학습에 기반한 중도탈락 예측을 시도한 국내 첫 연구이다. 실제 운영 중인 직업훈련기관의 실제 데이터를 사용하여 예측 성과를 실증적으로 입증하여, 제안 모형의 유용성을 과학적으로 검증하였다.

둘째, 본 연구에서는 중도탈락의 예측 정확도를 제고하기 위해 정형데이터 뿐 아니라 비정형데이터도 동시에 고려하는 예측 모형을 제시하였다. 그 결과 정형데이터만 사용했을 때와 비교하여 비정형데이터를 동시에 고려했을 때 최대 20%나 성능이 개선되는 것을 실증적으로 확인하였다. 비정형데이터는 정형데이터와 달리 일정한 규칙이 없어 그동안 기계학습으로 처리하기에 난해한 부분들이 있었으나, 선행연구들에서 제시된 텍스트를 벡터로 변환한 후 CNN 기법을 적용하는 접근법을 채택함으로써 이 같은 문제를 해결할 수 있었다. 실제로 교육 현장에서는 정형데이터 외에도 훈련교사의 교수 활동 및 훈련생들의 학습 활동에 의해 만들어지는 다양한 비정형 데이터들이 생산되고 있다. 이러한 데이터들을 의미없이 사장시키지 않고 분석에 활용하여 중도탈락을 예측하는데 도움이 되는 모형을 제시함으로써 직업훈련의 교육 성과를 향상시키는데 기여할 수 있을 것으로 기대된다.

셋째, 본 연구에서는 CNN을 활용하여 비정형 데이터를 분석하였다. 이미지 분석에 성능이 탁월하다고 알려져 있는 CNN이 텍스트 분석에도 효과적인 것과 비즈니스 문제 해결에도 적용할 수 있음을 본 연구를 통해 실증적으로 확인할 수 있었다. 특히 직업훈련의 중도탈락자 예측을 통한 교육 성과 개선이라는 주제에 기계학습 알고리즘을 국내에서 처음으로 적용함으로써 새로운 응용분야 개척에 공헌하였다고 할 수 있다.

한편 본 연구는 실무적인 측면에서 다음과 같은 시사점을 갖는다. 일자리 창출 및 청년 실업문제 해결을 위해서는 질적인 직업훈련의 실시와 직업훈련의 성과를 향상시키는 것이 매우 중요하다. 하지만 많은 수의 수강생들이 여러 사유로 인하여 중도탈락하게 되어 국민세금으로 운영되는 직업훈련의 효용성에 의문이 제기되기도 한다. 본 연구에서는 훈련교사가 입력하는 상담일지를 텍스트 마이닝으로 분석하였는데, 그 결과 훈련교사가 생각하는 훈련생의 수료 및 취업가능성이 실제로 많이 반영이 되었다. 훈련교사가 입력하는 것들 중 활용되지 못하고 사장되어지는 부분들이 많은 현실에서 상담일지를 형식적으로 작성하지 않고 객관적으로 자세하게 작성하는 것이 중요하며 상담일지를 분석한 결과를 잘 활용하면 직업훈련 분야에서 중도탈락 예방을 위한 적극적인 지도방안이 마련될 수 있음을 본 연구의 결과가 시사하고 있다. 아울러, 본 연구에서 제안한 모델을 활용하여 교육 현장에서 중도탈락 위험이 높은 훈련생들을 미리 체크하여 집중적으로 관리함으로써 직업 훈련의 성과를 제고하는 것이 가능하리라 예상된다.

하지만 본 연구의 학술적, 실무적 의의는 다음과 같은 한계점을 인식한 상태에서 채택되어야 할 것으로 보인다. 우선 본 연구에서는 서울 지역의 1개 교육기관만을 대상으로 데이터를 수집하여 표본 수가 적고, 본 연구에서 연구한 결과를 일반화하기에는 다소 어려움이 있는 것이 사실이다. 향후 연구에서는 다양한 지역과 기관을 대상으로 데이터를 수집하고, 훈련생의 학습 및 평가 데이터 등 좀 더 질적인 자료 수집 방법을 도입하며, 대량의 데이터를 수집하여 좀 더 안정적인 분석 모델을 설계할 필요성이 있다.

참 고 문 헌

[1] 아시아경제 뉴스, http://www.asiae.co.kr/news/view.htm?idxno=2 018082113382462570

[2] 김승수, *비정형정보와 CNN기법을 활용한 고객행태예측: 전자상거래 사례를 중심으로*, 한양대학교 경영학과, 박사학위논문, 2018.

[3] 직업교육훈련 촉진법 제2조 제1호, http://www.law.go.kr/lsInfoP.do?urlMode=lsInfoP&lsId=000864#0000

[4] 배경석, *직업교육훈련에 참여한 성인학습자의 중도탈락 요인 분석*, 한국기술교육대학교 대학원, 석사학위논문, 2004

[5] 고용노동부, *직업능력개발사업현황*, 2017.

[6] 권근배, *직업전문학교 수료자와 중도탈락자의 특성 비교연구 - 직업훈련 청소년의 중도탈락예방 프로그램 개발을 위한 기초연구*, 성균관대학교 행정대학원, 석사학위논문, 2001.

[7] 권혜진, "개인, 교육기관, 사회적 변인이 사이버대 재학생의 중도탈락의도 결정에 미치는 영향," 한국콘텐츠학회논문지, 제10권, 제3호, pp.404-412, 2010.

[8] 김경희, "지방대학생들의 학업중단 영향요인과 대학생활만족도 분석," 한국콘텐츠학회논문지, 제11권, 제8호, pp.378-387, 2011.

[9] 박대권, *실업자직업훈련의 중도탈락 원인분석*, 연세대학교 대학원, 박사학위논문, 1999.

[10] 이준택, *직업학교장면에서 중도탈락에 영향을 미치는 요인 탐색*, 호서대학교 대학원, 석사학위논문, 2004.

[11] 정선정, *직업교육 이러닝 e-Learning의 중도탈락 원인 분석*, 이화여자대학교 정보과학대학원, 석사학위논문, 2005.

[12] M. L. Conte, F. M. Rottino, and L. Salvati, "Dropping out from a Training Course after the High School in Italy," Proceedings of SIS2007, pp.503-504, 2007.

[13] H. Yi, L. Zhang, Y. Yao, A. Wang, Y. Ma, Y. Shi, J. Chu, P. Loyalka, and S. Rozelle, "Exploring the dropout rates and causes of dropout in upper-secondary technical and vocational education and training (TVET) schools in China," International Journal of Educational Development, Vol.42, pp.115-123, 2015.

[14] 구본용, 신현숙, 유제민, "데이터마이닝을 이용한 중퇴 모형에 관한 연구," 청소년상담연구, 제10권, 제2호, pp.35-57, 2002.

[15] 구본용, 유제민, "중퇴에 관한 위험 및 보호요인의 신경망 모형," 한국심리학회지, 제8권, 제1호, pp.133-146, 2003.

[16] 전주성, "사이버 대학의 잠재적 중도탈락자 예측에 관한 연구," Andragogy Today, 제13권, 제1호, pp.121-139, 2010.

[17] 정소영, 권수태, "연관규칙과 의사결정트리를 이용한 중도탈락자 예측모형 개발," 한국정보기술학회논문지, 제6권, 제5호, pp.202-210, 2018.

[18] 유지원, "일반대학에서 교양 e-러닝 강좌의 중도탈락 예측모형 개발과 조기 판별 가능성 탐색," 한국컴퓨터교육학회 논문지, 제17권, 제1호, 2014.

[19] 노혜란, 최미나, "대학 이러닝에서 학습자의 참여지속에 관한 로지스틱 회귀분석, 교육정보미디어연구," 제17권, 제4호, pp.593-614, 2011.

[20] B. S. Acharya and S. Neupane, "Determinants of vocational training drop out: A Logit Model Analysis," Annamalai International Journal Of Business Studies & Research, Vol.4, No.1, pp.75-80, 2012.

[21] H. Baars and H. G. Kemper, "Management Support with Structured and Unstructured Data -an Integrated Business Intelligence Framework," Information Systems Management, Vol.25, No.2, pp.132-148, 2008.

[22] 조성우, *Big Data 시대의 기술*, KT 종합기술원, pp.5-7, 2011.

[23] Y. Li and L. Xu, "Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective," In Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015.

[24] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," Journal of Artificial Intelligence Research, Vol.37, pp.141-188, 2010.

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations

in Vector Space," In Proceedings of Workshop at ICLR, pp.1-12, 2013.

[26] J. Firth, *A Synopsis of Linguistic Theory, 1930-1955*, Studies in Linguistic Analysis, pp.1-32, 1957.

[27] 김우주, 김동희, 장희원, "Word2vec을 활용한 문서의 의미 확장 검색방법," 한국콘텐츠학회논문지, 제16권, 제10호, pp.687-692, 2016.

[28] 박성수, 이건창, "워드 임베딩과 반감독 학습을 사용한 효율적 한국어 감성 표지 생성 방안," 한국지능시스템학회 논문지, 제28권, 제2호, pp.185-191, 2018.

[29] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," In International Conference on Neural Information Processing Systems (NIPS'13), pp.3111-3119, 2013.

[30] http://operatingsystems.tistory.com/entry/Data-Mining-Word2vec-CBOW

[31] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," In International Conference on Machine Learning, pp.1188-1196, 2014.

[32] 유용민, *Doc2vec과 문서 군집기법을 적용한 카테고리 자동생성*, 인하대학교 공학대학원, 석사학위논문, 2018.

[33] 이모세, 안현철, "효과적인 입력변수 패턴 학습을 위한 시계열 그래프 기반 합성곱 신경망 모형: 주식시장 예측에의 응용," 지능정보연구, 제24권, 제1호, pp.167-181, 2018.

[34] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[35] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," International Conference on Neural Information Processing Systems (NIPS'12), pp.1097-1105, 2012.

[36] 한정수, *컨벌루션 신경회로망과 ELM 분류기를 이용한 영상 분류*, 조선대학교 대학원, 석사학위논문, 2017.

[37] https://github.com/kkb2849/Word2vec-kor/blob/master/dict_data/w2v_model_wiki_kor

[38] 안현철, "유전자 알고리즘을 이용한 다분류 SVM의 최적화: 기업신용등급 예측에의 응용," Information Systems Review, 제16권, 제3호, pp.161-177, 2014.

[39] 이종식, 안현철, "입력변수 및 학습사례 선정을 동시에 최적화하는 GA-MSVM 기반 주가지수 추세 예측 모형에 관한 연구," 지능정보연구, 제23권, 제4호, pp.147-168, 2017.

저 자 소 개

하 만 석(Manseok Ha) 정회원

- 2002년 8월 : 국민대학교 정보관리학과(석사)
- 2010년 3월 ~ 현재 : 국민대학교 비즈니스IT전문대학원 비즈니스IT전공(박사과정)

<관심분야> : 비즈니스 애널리틱스

안 현 철(Hyunchul Ahn) 정회원

- 2006년 8월 : KAIST 테크노경영대학원 경영공학(박사)
- 2008년 3월 ~ 2009년 2월 : 성신여자대학교 경영학과
- 2009년 3월 ~ 현재 : 국민대학교 경영대학 부교수

<관심분야> : 비즈니스 애널리틱스, 추천시스템