The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 19, No. 2, pp.151-159, Apr. 30, 2019. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2019.19.2.151
JIIBC 2019-2-21

시민 데이터과학자를 위한 빅데이터 예측 지원 서비스

Bigdata Prediction Support Service for Citizen Data Scientists

장재영*

Jae-Young Chang*

요 약 4차 산업의 근간이 되는 빅데이터 시대가 도래하면서 대부분의 관련 업계에서는 데이터에 대한 저장, 통계분석 및 시각화 기술 중심으로 관련 솔루션들이 개발되고 있다. 하지만 빅데이터 기술의 근본적인 발전을 위해서는 인공지능을 이용한 예측 분석기술의 발전이 필요하다. 하지만 이러한 고급기술은 현재 데이터과학자라고 불리는 일부 전문가에 의해서만 가능한 수준이다. 이를 극복하기 위해서는 데이터에 대한 통찰력을 지닌 비전문가(시민 데이터과학자)가 빅데이터 분석 과정을 저비용으로 쉽게 접근할 수 있는 기반이 마련되어야한다. 본 논문에서는 고수준의 데이터과학 지식 없이 사용하기 쉬운 빅데이터 도구 및 기술의 도움으로 데이터를 분석하기 위한 서비스 과정을 제안한다. 이를 위해 필요한 예측 분석 시스템에 필요한 구성 요소와 환경을 정의하고 전반적인 서비스 설계 방안을 제시한다.

Abstract As the era of big data, which is the foundation of the fourth industry, has come, most related industries are developing related solutions focusing on the technologies of data storage, statistical analysis and visualization. However, for the diffusion of bigdata technology, it is necessary to develop the prediction analysis technologies using artificial intelligence. But these advanced technologies are only possible by some experts now called data scientists. For big data-related industries to develop, a non-expert, called a citizen data scientist, should be able to easily access the big data analysis process at low cost because they have insight into their own data. In this paper, we propose a system for analyzing bigdata and building business models with the support of easy-to-use analysis system without knowledge of high-level data science. We also define the necessary components and environment for the prediction analysis system and present the overall service plan.

Key Words : Citizen Data Scientist, Bigdata, Prediction Analysis, Data Blending, Feature Engineering

Ⅰ. 서 론

빅데이터 분석은 2008년 미국 오바마 대통령 대선 승리 스토리를 기점으로 하여, 아마존(Amazon)의 빅데이터 기반 추천 기술[1], 구글 트랜드(Google Trends)의 살인 독감 예측[2], 알파고(Alphgo)의 바둑시합 승리[3] 등에 이르기까지 그 가치가 매우 커지고 있어 2010년 이후 4차 산업의 핵심으로 떠오르고 있다. 또한 최근 한국을 포함하여 대부분의 IT 강국에서는 공공 오픈 데이터 서비스를 확대함에 따라, 민간 데이터와 공공 오픈 데이터와의 융합 분석이 가능해지고 있어 산업적 파생력이 급격히 상승하고 있다. 이와 더불어 경제적인 비용으로 언제 어디서나 쉽게 접속하여 다양하고 방대한 빅데이터를 다루면서 유의미한 인사이트를 발견하고 이를 비즈니스에 적

*정회원, 한성대학교 컴퓨터공학부
접수일자 2019년 1월 16일, 수정완료 2019년 2월 16일
게재확정일자 2019년 4월 5일

Received: 16 January, 2019 / Revised: 16 February, 2019 / Accepted: 5 April, 2019
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

용할 수 있는 데이터 분석 서비스에 대한 요구가 커지고 있다.

빅데이터 분석의 영역은 크게 단순 수준의 빅데이터 '요약분석(summary analysis)'과 고수준의 빅데이터 '모델분석(model analysis)'으로 구분할 수 있다[4]. 요약분석은 주어진 데이터를 이해하기 위해 특정 변수(특징)값들의 분포를 시각화하거나 통계량을 산출하는 작업이다. "제품 매출이 전년도 대비 10% 증가", "지역별 제품 판매량"의 시각화 등이 대표적인 요약분석의 예이다. 요약분석은 적절한 변수(속성)들을 선택하여 비교적 단순하게 원하는 결과를 얻을 수 있고 복잡한 사고과정을 필요로 하지 않는다. 반면에 모델분석은 기계학습(machine learning) 기술을 활용하여 방대한 데이터에 숨어 있는 고품질의 패턴(모델)을 추출하는 것으로서, 창의적 사고과정을 요구하는 한층 고도화된 작업이다. 좀 더 구체적으로 모델분석은 빅데이터 자체를 간단히 설명하기 위한 설명모델(description model)과 클래스 및 수치 값을 예측하는 예측모델(prediction model)을 만드는 작업을 포괄한다.

현재 대부분의 빅데이터 플랫폼들은 요약분석에 초점이 맞춰져있다. 즉, 대용량 데이터에 대한 저장, 통계분석 및 시각화 기술 중심으로 관련 솔루션들이 개발되고 있다. 하지만 대부분의 업체에서 활용하고 있는 요약분석 및 시각화 기술은 실제 부가가치가 크지 않다. 반면에 모델분석은 대용량 데이터에 숨겨진 고급 정보를 추출하는 기술로서 현업에서 의사결정과정에서 매우 중요한 정보를 제공한다. 특히 딥러닝(deep learning)으로 대표되는 인공신경망 기술의 급속한 발전으로 예측기술은 4차 산업의 근간이 되는 빅데이터 분야에서 가장 핵심적인 역할을 하고 있다. 하지만 예측분석은 일련의 과정에서 많은 시간과 비용이 소요되는 부분으로 현재 데이터과학자(data scientist)라고 불리는 일부 전문가에 의해서만 가능한 수준이다. 데이터과학자는 빅데이터 분석과 관련된 기술적 능력뿐만 아니라 분석 대상이 되는 데이터 그 자체를 완벽히 이해하고 이로부터 빅데이터 분석의 실마리가 되는 요소를 도출할 수 있는 능력을 갖추어야 한다. 무엇보다도 수행되는 빅데이터 분석 작업으로부터 적정의 가치를 갖는 결과를 얻기 위해서는 대상 데이터를 이해와 통찰로부터 출발해야 한다는 인식을 가지는 것이 매우 중요하다. 하지만 복합적인 역량을 두루 갖춘 숙련 데이터과학자를 양성하기는 쉽지 않으며, 제대로 된 전문 데이터과학자의 수도 많지 않다. 따라서 유의미한 빅데이터 분석을 시도하기 위해서 전문 데이터과학자 또는 이러한 인력을 갖춘 업체에 대해 고비용을 지불할 수밖에 없다. 현재 대부분의 데이터 분석이 요약분석에 치중하는 것도 이와 같이 이유가 포함되어 있다. 따라서 데이터에 대한 통찰력만 갖춘 비전문가도 쉽게 고수준의 예측분석이 가능하도록 지원해주는 시스템의 요구가 증대되고 있으며, 시민 데이터과학자(Citizen Data Scientist)라는 개념도 이와 같은 상황에서 등장하기 시작하였다[5].

시민 데이터과학자는 고수준의 데이터과학에 대한 지식 없이, 사용하기 쉬운 자동화된 빅데이터 도구 및 기술의 도움으로 예측과 같은 고수준의 데이터 분석과 비즈니스 모델을 만드는 역할을 수행한다. 여기서 자동화된 예측분석 서비스란 데이터 소유자가 데이터를 업로드하고 최소한의 데이터 과학 지식으로 예측 또는 설명 모델을 신속하게 구축 할 수 있게 해주는 서비스를 의미한다. 시민 데이터과학자는 데이터과학, 비즈니스 인텔리전스 전문가일 필요는 없으며, 비즈니스 문제에 초점을 맞추어 자기 영역 내에서 분석하고자 하는 데이터를 수집하고 일련의 자동화 도구를 이용하여 데이터 분석이 가능해야한다. 현재 시민 데이터과학자에 대한 관심이 점차 높아지고 있으나 이를 실현하기 위한 핵심 기술과 시스템에 대해서는 아직까지 많은 연구가 되어있지 않은 상태이다. 본 논문에서는 시민 데이터과학자의 확산을 지원하기 위해 고수준의 예측 분석이 가능 핵심 요소들을 정의하고 이러한 요소들의 자동화 방안을 제시한다. 또한 이러한 기능들이 통합된 시민 데이터과학자를 위한 빅데이터 예측 지원서비스의 설계 결과를 제시한다. 본 논문에서는 시민 데이터과학자를 지원하기 위한 빅데이터 기반 예측 분석 서비스의 자동화 요소를 크게 데이터 블렌딩(data blending)[6], 특징공학(feature engineering)[7], 예측모델 생성으로 정의하였으며, 각 요소들을 자동화하기 위한 방안과 모델을 정의한다. 또한 이러한 서비스를 통해 활용 가능한 분야와 효과에 대해서도 논한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련기술동향에 대해 알아보고 3장에서는 자동화 요소들에 관해 제시한다. 4장에서는 전체적인 서비스 설계 결과를 제시하고 5장에서는 활용분야와 효과에 대해 논한다. 마지막으로 6장에서는 결론을 맺는다.

II. 관련기술동향

2015년 가트너(Gartner)가 시민 데이터과학자라는 용어를 정의하고 이 그룹이 데이터 과학자보다 5배 더 빠르게 성장할 것으로 예상한 이후, 분석 플랫폼 개발자는 시민 데이터과학자를 지원하기 위한 기능 확장에 많은 관심을 기울이고 있다[8]. 시민 데이터과학자의 확산을 위해서는 이들이 쉽게 이용할 수 있는 분석도구인 스마트 데이터탐색(smart data discovery)[9] 기능의 지원이 필수적이다. 가트너는 스마트 데이터탐색을 "비즈니스 사용자나 시민 데이터과학자들에게 고급 분석의 통찰력을 제공하는 차세대 데이터 탐색기능"으로 정의하고 있다. 현재 많은 빅데이터 분석 플랫폼 솔루션들도 이러한 스마트 데이터탐색 가능을 지원하려는 노력이 계속되고 있다. 하지만 현재 이러한 기능을 지원하려는 대부분의 솔루션들은 단순한 요약분석이나 시각화 중심의 자동화에 초점이 맞추어져 있으며, 데이터 분석의 핵심이 되는 자동화된 예측분석에 대해서는 아직 만족할만한 서비스가 이루어지지 않고 있다.

IBM Watson Analytics[10], Microsoft Azure ML[11], Amazon ML[12]과 같은 주요 클라우드 서비스 제공업체에서 사용자 친화적 인터페이스 제공함으로써 시민 데이터과학자를 지원하기 위한 서비스를 고려하고 있으나 아직 데이터 융합, 특징공학의 자동화 측면에서 새로운 돌파구가 필요한 실정이다. 구체적으로 IBM Watson Analytics의 경우에는 결정트리(decision tree)와 같은 일부 예측분석 기능을 제공하고 있으나 엔터프라이즈급의 데이터 처리에 한계가 있다. 또한 자연어 처리에는 강점이 있으나 데이터 시각화나 탐험분석 중심으로 기능을 제공하고 있다. Miscrosoft Azure ML의 경우에는 사용자 친화적인 인터페이스를 제공하고 있으나 기계학습에 대한 지식이 있는 개발자나 데이터과학자 위주의 사용환경을 제공하고 있다. 마지막으로 Amazon ML은 데이터 수집이 용이하고 별도의 관리가 필요하지 않는 장점이 있다. 또한 시민 데이터과학자가 사용하기 좋은 환경을 제공하고 있으나 적용분야가 좁은 단점을 갖고 있다. 이외에도 많은 빅데이터 분석 솔루션에서도 시민 데이터과학자를 위한 스마트 데이터탐색 기능을 제공하려는 움직임이 있으나 아직까지는 만족할만한 성과를 내지는 못하고 있는 실정이다. 그 이유는 여러 가지로 분석될 수 있으나, 시민 데이터과학자가 데이터분석 과정에서 어려움을 겪는 데이터 블렌딩이나 특징공학 그리고 예측 분석과 같은 고수준의 작업들이 일부 혹은 전체적으로 자동화되지 못한 것에 기인한다. 본 논문에서 제안하는 시스템은 단순히 편리한 인터페이스 기능을 제공하는 수준을 넘어, 데이터과학자들에 의해 진행되어 온 고급 기능들을 자동화하기 위한 방안과 설계 결과를 제안한다.

III. 데이터탐색을 위한 자동화 요소

1. 데이터 과학자 역량

숙련된 데이터과학자는 매우 폭넓은 역량을 갖춰야 하는데, 실제 산업계에서 요구되는 역량을 갖춘 데이터 과학자를 양성하는 것은 매우 어려운 일이다[13]. 데이터 과학자가 가져야 하는 빅데이터 분석 능력은 다음과 같이 크게 3가지로 나눌 수 있다.(그림 1)

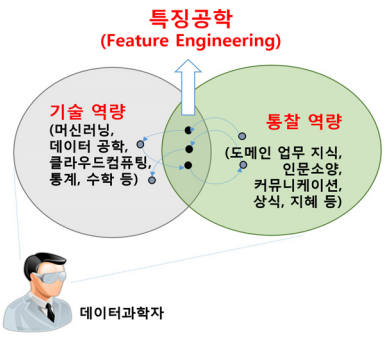


그림 1. 데이터과학자의 역량
Fig. 1. Abilities of Data Scientist

○ 업무 및 데이터에 대한 '통찰 역량 (Insight Competence)'
○ 빅데이터 분석을 위한 제반 '기술적 역량(Technical Competence)'
○ 최적의 모델 인자를 구성할 수 있는 '특징공학 역량 (Feature Engineering Competence)'

우선 통찰 역량은 분석 대상이 되는 도메인에 대한 전문적 지식의 습득 능력, 인문학적 소양, 커뮤니케이션 능력, 상식 능력 등을 포괄한다. 구체적으로 '왜 분석을 해야 하는지', '어떤 목적의 분석을 수행하는 것이 가치와 의미가 있는 것인지', '무엇을 예측해야 하는지', '정의된 분석 목적을 달성하기 위해서 어떠한 데이터 속성들이

필요한지' 등의 문제를 체계적이고 명확하게 접근하는 능력을 말한다. 데이터과학자는 보다 일반화된 영역에 대한 탐구력을 가져야 하며, 친숙하지 않은 미지의 도메인이라 할지라도 분석의 동기와 목적을 신속 정확하게 파악해야 할 수 있어야 한다.

다음으로 빅데이터 분석 관련 기술적 능력은 데이터 분석을 위해 활용되는 소프트웨어에 대한 활용 능력을 의미하며, 세부적으로 빅데이터 기술 능력은 기계학습, 데이터공학, 클라우드 컴퓨팅 기술 능력으로 구분할 수 있다. 기계학습은 주어진 빅데이터의 개별성을 아우르는 일반화된 설명 및 예측모델을 생성하는 기술이다. 데이터 공학(data engineering)은 데이터베이스 모델링, 데이터 저장/검색/관리 등과 관련된 기술 영역을 포괄한다. 최근 빅데이터 저장/관리를 위해 빠르게 확산되고 있는 NoSQL 관련 기술 또한 데이터공학의 범주에 들어갈 수 있다. 마지막으로 클라우드 컴퓨팅(cloud computing) 기술은 네트워크 상에서 데이터 저장의 영속성/신뢰성을 높이고 데이터 처리 및 계산 속도를 높이기 위한 제반 기술로서, 최근 Hadoop, Spark가 분산컴퓨팅과 관련된 기술로서 각광을 받고 있다.

마지막으로 특징공학 능력은 빅데이터를 일반화하여 생성하게 되는 모델의 인자를 구성하는 것을 의미한다. 특징 공학은 세부적으로 특징 생성(feature generation), 특징 변형(feature transformation), 특징 선택(feature selection)과 같은 작업으로 나눌 수 있다. 우선 특징 생성은 분석 목적과 관련된 새로운 속성을 도출하거나 유도 속성을 정의하는 작업을 의미하며, 특징 변형은 기계학습 알고리즘의 특성에 맞게 또는 분석 모델의 성능을 높이기 위해 속성 값을 변환하는 작업을 말한다. 마지막으로 특징 선택은 차원의 저주(curse of dimensionality) 문제 및 과도학습(overfitting) 현상을 회피하기 위해 설명/예측모델의 성능에 기여할 수 있는 속성을 자동 또는 수동으로 선택하는 작업을 의미한다. 특히 특징 생성 작업은 관련 업무 및 데이터에 대한 통찰력과 창의적 사고가 요구하며, 비교하여 특징 선택과 특징 변환 작업은 활용되는 기계학습 알고리즘에 대한 충분한 이해가 필요하다. 최근 인공신경망 기술의 급속한 발전으로 딥러닝 기술의 약진하고 있으며, 이와 더불어 특징공학 관련 작업을 완전 자동화하기 위한 연구로서 특징 학습(feature learning)[14], End-to-End Learning[15] 등의 기술도 주목받고 있다.

2. 예측분석을 위한 자동화

본 논문에서 제안하는 시민 데이터과학자를 위한 예측분석 서비스를 지원하기 위해서는 기존의 데이터과학자에 의해 수행되어진 일련의 과정을 자동화하는 것이 필수적이다. 데이터분석 전과정을 자동화하는 것은 거의 불가능하므로 데이터과학자만이 수행할 수 있었던 비교적 고난도의 작업에 대해 자동화가 필요한데, 본 논문에서는 구체적으로 데이터 블렌딩(융합), 특징공학, 예측모델 생성 과정의 자동화에 대해 제안한다. 그림 1은 본 논문에서 제안하는 자동화 과정을 보여준다. 이 그림에서 "Data In-Model Out"은 최소한의 과정으로 분석 데이터셋을 입력하여 원하는 예측모델을 얻을 수 있는 개념이며, 이 용어는 이는 최근 빅데이터 기반 예측 분석의 자동화에 대한 요구가 커지면서 William Vorhies에 의해 2016년에 처음으로 등장하였다[16].

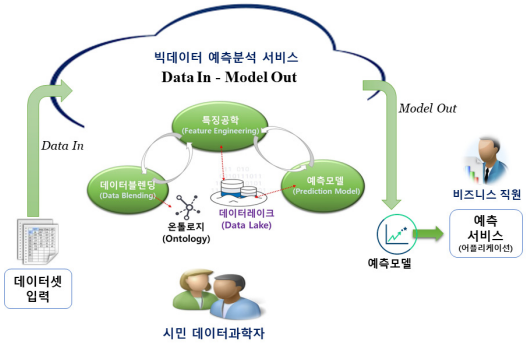


그림 2. Data In-Model Out 서비스 개념
Fig. 2. Service Concept of Data In-Model Out

우선 데이터 블렌딩이란 관련 데이터를 분석 목적에 맞도록 매쉬업(mashup)하는 작업을 의미한다. 특히 예측분석의 경우 예측모델을 생성하기 위한 학습데이터를 생성하는 것을 목적으로 한다. 이와 비교하여 데이터 융합(data integration)은 다양한 목적을 가진 사람들이 서로 공유하기 위해 여러 소스의 데이터를 하나의 저장소에 체계적으로 축적하는 작업을 의미한다. 따라서 데이터 융합을 통해 얻어진 하나의 저장소는 지속성을 갖는 것이 필수이지만, 데이터 블렌딩을 통해 생성된 매쉬업 데이터는 선택적이다. 따라서 데이터 블렌딩을 자동화한다는 것은 사용자가 입력한 로컬 데이터들의 매쉬업 연산을 자동화하는 것을 의미하며, 사용자의 로컬 데이터와

외부의 오픈 데이터 그리고 SNS 텍스트 간의 매쉬업 연산에 대한 자동화가 필요하다. 또한 이를 실현하기 위해서는 의미적 유사관계를 가지는 속성(컬럼)간의 매쉬업을 위해 관련 온톨로지를 구축하는 것이 필수적이다.

특징공학의 자동화란 예측모델의 기반이 되는 입력 데이터에 대한 정제, 보정 연산의 자동화를 의미한다. 예측모델의 생성을 위해서는 기존 특징의 변환, 새로운 특징의 생성, 우수 특징의 선별하는 과정이 필요한데 이러한 일련의 작업을 자동화함으로써 데이터 분별력이 높고 정보량이 많은 특징을 선별하는데 도움을 줄 수 있다. 특징 변환/생성의 자동화를 위해서는 특징 타입/개념과 변환/생성간의 관계를 가지는 온톨로지의 구축이 선행되어야 한다.

마지막으로 예측모델 생성의 자동화는 분석 데이터의 형태 및 내용을 고려하여 기계학습 알고리즘을 사용자의 도움 없이 자동으로 선택하여 모델을 생성하는 것을 의미한다. 예측모델은 기계학습 알고리즘에 의해 학습되어 출력된 결과물로서, 미지의 입력 데이터에 대하여 사용자가 원하는 타겟 특징값(target feature value)을 수치 또는 클래스 값으로 출력하게 되는데, 현재까지 많은 알고리즘이 개발되었으며 각 알고리즘마다 장단점이 있어 어떤 알고리즘을 선택하느냐에 따라 예측 성능이 좌우될 수 있다. 따라서 최적의 예측모델 생성을 위해 기계학습 알고리즘을 자동 선택하는 것은 주요한 자동화 기술 중의 하나라고 볼 수 있다.

Ⅳ. 예측분석 지원 서비스 과정

1. 예측분석 지원 서비스를 위한 구성요소

빅데이터를 이용한 전형적인 예측 분석 과정은 그림 1(a)와 같다. 이 그림에서 보는 바와 같이 수집 과정에 얻어진 데이터를 가공한 후에 융합 및 확장 등의 데이터 블렌딩 과정을 거치게 되면 분석을 위한 기반 데이터가 생성된다. 다음으로 원하는 예측모델이 생성될 때까지 특징공학과 예측모델 생성 및 테스트를 반복한다. 앞서 언급한 바와 같이 이러한 과정은 주로 데이터과학자에 의해 수행되어 왔다. 이를 기반으로 시민 데이터과학자를 위한 예측분석 지원을 위해서는 주요 과정에 대한 자동화가 필요한데 이 과정은 그림 1(b)에 표현되어 있다. 이 그림에서 보는 바와 같이 사용자가 로컬 데이터를 입력한 이후 해당 예측모델이 생성되기까지 일련의 과정을 자동화하기 위해서는 데이터 블렌딩, 특징공학, 예측모델 생성을 자동화해야 한다. 이 과정에서 온톨로지와 데이터 레이크(data lake)[17] 의 구축이 필수적이다. 온톨로지는 데이터 블렌딩, 특징공학, 예측모델 생성의 자동화를 위해 필요한 제반 지식 체계를 담고 있는 저장소로서 로컬 데이터와 외부 오픈 데이터와의 블렌딩, 새로운 특징 생성 등을 위해 그 역할이 매우 중요하다. 또한 데이터 레이크는 예측분석의 대상이 되는 데이터를 저장할 수 있는 데이터 저장소로서 데이터 블렌딩을 통해 얻어진 데이터 집합체가 저장되는 공간을 의미한다.

2. 데이터 블렌딩의 자동화

데이터 블렌딩은 다양한 소스로 부터 특정 분석 목적에 부합하는 융합 뷰(view)를 생성하는 과정이다. 데이터 분석 과정에서 원하는 데이터가 수집 데이터 집합에 산재해있는 경우가 자주 발생하게 되는데, 이를 자동적으로 하나의 뷰로 통합하여 제공할 수 있다면 데이터 분석 과정을 획기적으로 개선할 수 있다. 예를 들어 데이터베이스 관점에서 테이블간의 외래키 관계 또는 메타데이터(meta data)를 이용한 컬럼 간의 의미적 유사성을 이용하여 여러 테이블에 나뉘어져 있는 관련 컬럼들을 하나의 뷰로 융합하여 제공할 수 있다.

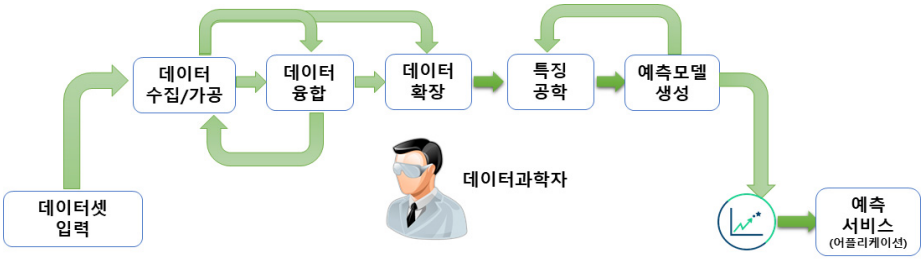
특히 공공 데이터포털과 같은 오픈 데이터의 경우에는 메타데이터를 활용하여 데이터셋들 간의 관계를 맺게 함으로써 사후 특정 데이터셋을 중심으로 한 매핑과정을 통해 융합 작업을 수행할 수 있다. 데이터의 융합을 위해서는 데이터간의 연관관계를 사전에 정의할 필요가 있는데 대표적으로 융합에 필요한 5가지의 연관관계는 표 1과 같다. 이러한 연관관계는 대부분 데이터에 부여된 메타데이터를 통해 확인 가능하다. 따라서 메타데이터를 이용하여 각 데이터셋 간의 연관관계를 자동으로 추출하는 것이 데이터 블렌딩 자동화의 핵심기술이다. 특히 연관관계 중에 비교적 명확한 스키마 연관의 경우에는 융합뷰에 대한 자동 생성이 가능하다. 하지만 나머지 연관에 대해서는 아직까지 자동적인 융합 뷰를 제공하는 것이 기술적으로 매우 어렵다. 따라서 이 경우에는 데이터맵(datamap)과 같은 다양한 시각화를 통해 시민 데이터과학에게 제공하여 융합 뷰를 생성하는데 도움을 줄 수 있다.

표 1. 데이터 융합을 위한 연관관계 종류
Table 1. Association Types for Data Convergence

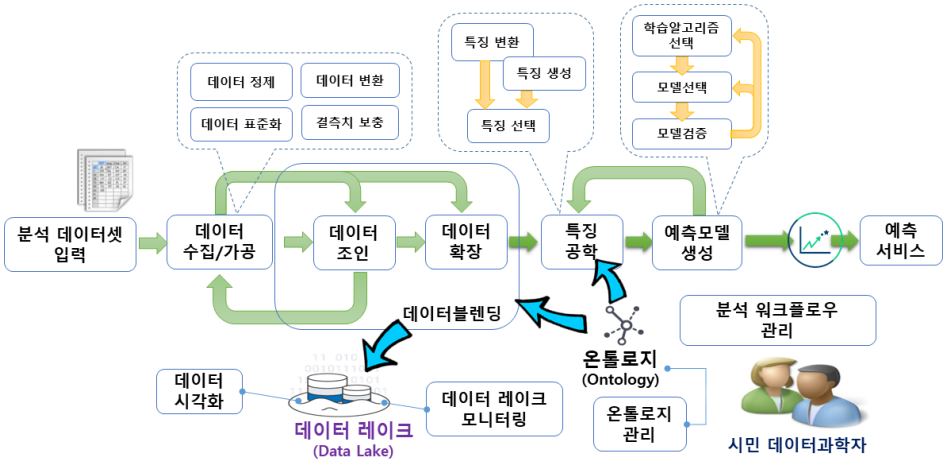
연관관계	설 명
스키마 연관	스키마에 포함된 필드 (또는 속성) 간 동일(또는 유사) 관계
의미적 연관	데이터셋 내부의 콘텐츠간 의미적 유사성을 평가 (Related 관계)
포함 연관	데이터셋간의 개념적 포함 관계 (IS_A관계)
지역 커버리지 연관	데이터 제목과 스키마는 유사하지만, 관련 지역이 상이한 경우
시간 커버리지 연관	데이터 제목과 스키마는 유사하지만, 관련 지역이 상이한 경우

3. 특징공학의 자동화

특징공학은 기존 특징변수로부터 예측에 도움이 되는 새로운 특징변수를 생성 또는 변형하는 과정이다. 따라서 다음 단계인 특징선택 과정을 위해 유의미한 특징 집합을 생성해야한다. 기본적으로 특징공학을 위한 온톨로지는 주어진 특징 타입에 대하여 적용할 수 있는 가능한 변환 연산자에 정보를 갖고 있어야한다. 예를 들어 수치형 특징에 대한 변환 연산자로는 Log Transformation, Binarization, Binning, Scaling (MinMax Scaling, Z-score Scaling) 등이 있으며, 범주형 특징에 대한 변환 연산로는 One-Hot Encoding, Feature Hashing, Bin-counting, Label Count Encoding, Category Embedding 등을 정의할 수 있다[18]. 시공간 특징에 관한 연산자로는 Time Binning, Duration, Distance 추출, 좌표 기반 외부 데이터 블렌딩 등의 연산자를 예로 들 수 있다. 이와 같이 사전에 정의된 연산자를 기반으로 특징공학의 자동화를 위해서는 기계학습을 이용하여 후보 특징을 선정하기 위한 랭킹 모델을 생성해야한다.



(a) 기존 예측지원 서비스 과정
(a) Legacy Predictive Support Services Process



(b) 자동화된 예측지원 서비스 과정
(b) Automated Predictive Support Services Process

그림 3. 시민 데이터과학자를 위한 예측지원 서비스
Fig. 3. Prediction Support Services for Citizen Data Scientists

그림 4는 특징 랭킹모델을 이용하여 후보특징들을 생성하는 과정을 보여준다. 이 그림에서 보는 바와 같이 우선 과거의 특징생성 이력을 이용하여 후보특징을 평가하기 위한 기계학습 기반 랭킹모델을 생성한다. 이 모델의 입력으로 사용될 후보 특징들은 데이터셋으로부터 생산되는데, 위에서 언급한 다양한 연산자를 통해 생성하게 된다. 다음으로 이 후보특징들에 대해서 랭킹모델을 이용하여 후보 특징들을 평가하여 랭킹한다. 그 결과를 바탕으로 선정된 특징들을 새로운 후보특징으로 분류하게 된다. 이러한 과정을 반복하게 되면 다음 단계인 특징선택 과정에서 사용하기 적합한 유의미한 특징들만을 선별할 수 있게 된다.

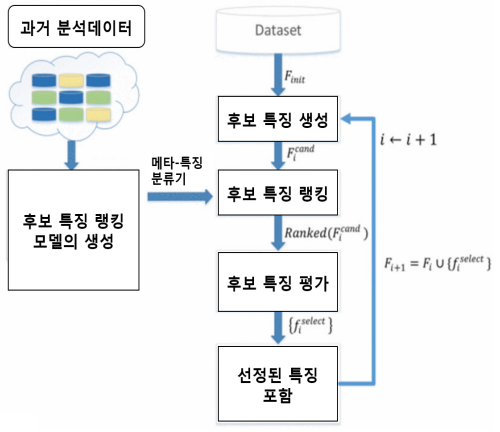


그림 4. 기계학습 기반 자동 특징생성 프로세스
Fig. 4. Machine Learning-Based Automatic Feature Generation Process

4. 예측분석의 자동화

예측분석을 위해 우선적으로 고려해야할 사항은 학습알고리즘의 자동 선택이다. 주어진 데이터에 따라 적합한 학습 알고리즘을 선택하기 위해서는 예측 변수의 타입을 먼저 선정해야한다. 기본적으로 예측 변수의 타입이 범주형인 경우 분류 알고리즘을 사용하고, 수치형인 경우 회귀분석 알고리즘을 사용한다. 또한 세부적인 학습 알고리즘을 선택하기 위해서는 입력 데이터의 정보(각 후보특징 타입과 예측변수, 학습 데이터의 양 등)를 입력하고 이를 기반으로 후보 알고리즘 추천해야한다. 후보 알고리즘은 각 알고리즘의 장단점과 기존의 성능평가 정보를 바탕으로 추천할 수 있다. 사용자는 이러한 과정을 통해 추천된 알고리즘을 이용하여 예측분석을 실시하고 그 예측결과의 시각화를 통해 그 결과를 확인하게 된다. 위와 같은 과정을 통해 예측분석을 수행하게 되면 추천 알고리즘의 수와 입력 특징들의 조합에 따라 다수의 예측모델이 생성된다. 이 중에서 최적의 모델을 선택하기 위해서 교차검증(cross validation)과 같은 검증 단계를 거쳐 높은 예측 정확도를 가지는 모델을 최종적으로 선택하게 된다. 그림 5는 지금까지의 과정을 보여주는데 각 단계마다 튜닝 및 시각화를 위해 적정 수준에서 제어 및 편집 기능을 추가할 수도 있다.

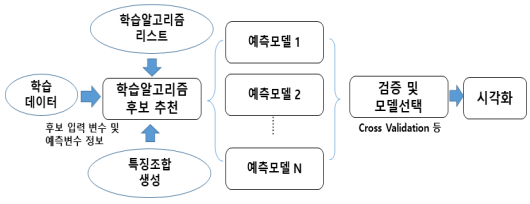


그림 5. 예측분석 자동화 과정
Fig. 5. Automatic Prediction Analysis Process

V. 예측분석 서비스의 활용

현재 기업 및 기관 등에서 예측분석 등의 고급분석 업무는 일부 데이터과학자에 의해 독점적으로 수행되고 있다. 데이터과학자는 데이터분석 그 자체에는 전문가이나 현장의 업무를 제대로 파악하지 못한 외부 인력인 경우가 대부분이다. 따라서 현장의 업무에 대한 통찰력이 없어 분석결과를 실시간으로 평가하는 데 한계가 있다. 반면에 본 논문에서 제안한 예측분석 서비스의 자동화는 고수준의 분석 업무를 현업사용자 수준에서 수행할 수 있는 환경이 마련해 줄 수 있다. 따라서 각 기관의 신속한 의사결정으로 인한 업무의 효율화 및 매출 증대를 기대할 수 있다. 표 2는 시민 데이터과학자의 예측지원서비스가 활용될 수 있는 분야의 예를 보여준다. 이 서비스들은 현 시점에서 사회적으로 예측분석의 요구가 많으나 대중적으로 활용되기에는 시간과 비용측면에서 여전히 많은 한계를 갖고 있다. 따라서 본 논문에서 제안한 서비스가 적용될 경우 시민 데이터과학자의 위치에서 직접적인 분석으로 인한 업무효율화를 가져올 수 있을 것으로 평가된다.

표 2. 예측분석의 적용분야
Table 2. Applications of Prediction Analysis

분야	적용 시스템	예측모델
금융·통신·판매	우수고객 분석 시스템	우수고객 판별 예측 모델
	고객 반응 분석 시스템	고객 반응 모델
	상품 추천 시스템	선호 상품 예측 모델
	해외 배송 예측 시스템	해외 직접구매 배송 예측 모델
제조	제조 품질관리 시스템	품질 평가 모델
공공	범죄 분석 시스템	범죄 발생 예측 모델
	세금포탈 추적 시스템	세금 포탈 여부 판별 모델
	민원 평판 분석 시스템	극성 분석 모델

Ⅵ. 결 론

본 논문에서는 시민 데이터과학자를 위해 빅데이터 예측 지원서비스가 갖춰야할 조건과 기능을 제안하였다. 자동화 요소로는 데이터 블렌딩, 특징공학, 예측모델 생성을 정의하였으며, 각 요소들을 자동화하기 위한 방안과 모델을 정의하였다. 본 논문에서 제안한 서비스는 현재 대부분 외부 위탁으로 진행되고 있는 고수준의 분석 업무를 현업사용자 수준에서 수행할 수 있는 기반을 마련해 줄 수 있다. 따라서 데이터 분석 및 예측 업무의 대중화에 기여할 수 있으며, 각 기관에서 데이터 분석 업무가 활성화되어 관련 업무를 수행하기 위한 인력의 수요도 증가할 것을 기대된다. 향후에는 후속 연구로서 본 논문에서 제안한 자동화 요소의 실현 가능성에 대한 검증과 세부 구현 방법에 대해 진행할 예정이며, 클라우드 환경에서 이러한 서비스를 제공하기 위한 연구도 병행할 예정이다.

References

[1] Greg Linden, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." IEEE Internet computing 1 (2003): 76-80.
DOI: 10.1109/MIC.2003.1167344

[2] Jörg Rech, "Discovering trends in software engineering with google trend." ACM SIGSOFT Software Engineering Notes 32.2 (2007): 1-2.
DOI: 10.1145/1234741.1234765

[3] Jim X. Chen. "The evolution of computing: AlphaGo." Computing in Science & Engineering 18.4 (2016): 4-7.

[4] Gary Miner, John Elder IV, and Thomas Hill. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012.

[5] Steve Banker, “The Citizen Data Scientist”, Article of Forbes, Available via https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist

[6] https://www.datawatch.com/ what-is-data-blending/

[7] Sam Scott and Stan Matwin. "Feature engineering for text classification." ICML. Vol. 99. 1999.

[8] David W. Cearley and Carl Claunch. "The top 10 strategic technology trends for 2013." The Top 10 (2012).

[9] Jacky Akoka and Isabelle Comyn-Wattiau. "A Method for Emerging Technology Evaluation. Application to Blockchain and Smart Data Discovery." Conceptual Modeling Perspectives. Springer, Cham, 2017. 247-258.
DOI: https://doi.org/10.1007/978-3-319-67271-7_17

[10] Robert Hoyt, et al. "IBM Watson analytics: automating visualization, descriptive, and predictive statistics." JMIR public health and surveillance 2.2 (2016).

[11] Roger Barga, et al. Predictive analytics with Microsoft Azure machine learning. Apress, 2015.
DOI: 10.1007/978-1-4842-0445-0

[12] Amazon. Available via https://aws.amazon.com/aml/

[13] Wikipedia. Available via https://en.wikipedia.org/wiki/Data_science

[14] Y-lan Boureau and Yann L. Cun. "Sparse feature learning for deep belief networks." Advances in neural information processing systems. 2008.

[15] Jie Zhou and Wei Xu. "End-to-end learning of semantic role labeling using recurrent neural

networks." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing . Vol. 1. 2015.
DOI: 10.3115/v1/P15-1109

[16] William Vorhies. "Data Scientists Automated and Unemployed by 2025!", Available via https://www.datasciencecentral.com/profile/WilliamVorhies

[17] Natalia Miloslavskaya and Alexander Tolstoy. "Big data, fast data and data lake concepts." Procedia Computer Science 88 (2016): 300-305. DOI: https://linkinghub.elsevier.com/retrieve/pii/S1877050916316957

[18] Amanda Casari, Alice Zheng, Feature Engineering for Machine Learning, O'Reilly Media, Inc.

[19] J. Chang. "Feature-Based Summarization for a Large Opinion Documents Collection", Journal of the Institute of Internet, Broadcasting and Communication, Vol. 16, No. 1, 2016.

[20] K. Lee and J. Lee, "A Classification of Medical and Advertising Blogs Using Machine Learning", Journal of the Korea Academia-Industrial cooperation Society(JKAIS), Vol. 19, No. 11, 2018.

[21] K, Ko, D. Whang, S. Park, and K. Moon, "Electrical fire prediction model study using machine learning", The Journal of KIIECT, Vol. 11, No. 6, 2018.

저자 소개

장 재 영(정회원)

- 1992년 : 서울대학교 계산통계학과 (이학사)
- 1994년 : 서울대학교 계산통계학과 (이학석사)
- 1999년 : 서울대학교 계산통계학과 (이학박사)
- 2000년~현재 : 한성대학교 컴퓨터공학부 교수
- 관심분야 : 데이터베이스, 데이터마이닝

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.