한국빅데이터학회지
제2권 제2호, 2017, pp. 75-86

머신러닝을 이용한 빅데이터 품질진단 자동화에 관한 연구*

A Study on Automation of Big Data Quality Diagnosis Using Machine Learning

이 진 형†

위세아이텍 인공지능팀

요 약

본 연구에서는 빅데이터의 품질을 진단하는 방법을 자동화하는 방법을 제안하고 있다. 빅데이터의 품질 진단을 자동화해야 하는 이유는 4차 산업혁명이 이슈화 되면서 과거보다 더 많은 볼륨의 데이터를 발생시키고 이 데이터들을 활용 하려는 요구가 증가하기 때문이다. 데이터는 급증하지만 데이터의 품질을 진단하기 위해 많은 시간이 소비된다면 데이터를 활용하기 위해 많은 시간이 걸리거나 데이터의 품질이 낮아질 수 있다. 그러면 이러한 낮은 품질의 데이터로부터 의사결정이나 예측을 한다면 그 결과 또한 잘못된 방향을 제시할 것이다. 이러한 문제를 해결하기 위해 많은 데이터를 신속하게 진단하고 개선할 수 있는 머신러닝 이용한 빅데이터 품질 향상을 위한 진단을 자동화 할 수 있는 모델을 개발하였다. 머신러닝을 이용하여 도메인 분류 작업을 자동화하여 도메인 분류 작업 시 발생할 수 있는 오류를 예방하고 작업 시간을 단축시켰다. 연구 결과를 토대로 데이터 변환의 중요성, 학습되지 않은 데이터에 대한 학습 시킬 수 있는 방안 모색, 도메인별 분류 모델을 개발에 대한 연구를 지속적으로 진행한다면 빅데이터를 활용하기 위한 데이터 품질 향상에 기여할 수 있을 것이다.

■ 중심어 : 빅데이터, 데이터 품질, 머신러닝, 도메인, 데이터전처리

Abstract

In this study, I propose a method to automate the method to diagnose the quality of big data. The reason for automating the quality diagnosis of Big Data is that as the Fourth Industrial Revolution becomes a issue, there is a growing demand for more volumes of data to be generated and utilized. Data is growing rapidly. However, if it takes a lot of time to diagnose the quality of the data, it can take a long time to utilize the data or the quality of the data may be lowered. If you make decisions or predictions from these low-quality data, then the results will also give you the wrong direction. To solve this problem, I have developed a model that can automate diagnosis for improving the quality of Big Data using machine learning which can quickly diagnose and improve the data. Machine learning is used to automate domain classification tasks to prevent errors that may occur during domain classification and reduce work time. Based on the results of the research, I can contribute to the improvement of data quality to utilize big data by continuing research on the importance of data conversion, learning methods for unlearned data, and development of classification models for each domain.

■ Keyword : Big Data, Data Quality, Machine Learning, Domain, Data Preprocessing

2017년 11월 27일 접수; 2017년 12월 18일 수정본 접수; 2017년 12월 26일 게재 확정

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW컴퓨팅산업원천기술개발 사업의 일환으로 수행하였음 [1711055424, 빅데이터 품질평가 도구 개발].

† 교신저자 jhlee@wise.co.kr

Ⅰ. 서 론

1.1 연구 선정 배경 및 필요성

공공기관, 제조, 금융, 통신, 의료, IT 등 다양한 분야에서 빅데이터를 활용해서 새로운 가치를 창출하려는 니즈가 커지고 있다. 정부와 많은 기업들은 인공지능 연구개발에 많은 투자를 시작하였다[11]. 한국 정부도 AI 기술 관련 R&D를 본격적으로 추진한다고 밝혔다[10]. 정부는 2017년에 들어와서 4차 산업혁명을 활성화하기 위해 공공 빅데이터를 개방하여 인공지능 산업을 키울 것이라고 하였다[9].

빅데이터를 이용하여 새로운 가치를 창출하고, 인공지능 산업을 키우려고 할 때, 데이터가 정확하지 않다면 데이터 기반으로 발생한 분석 결과와 인공지능 기술에 잘못된 영향을 미칠 것이다. 이미 전 세계적으로 민간부분에서만 데이터 신뢰성과 품질 확보를 위해 매년 6,000억 달러를 사용하고 있다[6]. 이미 공공기관을 대상으로 공공데이터의 품질을 높이기 위해 데이터 품질관리 수준을 평가하기 위한 모델 관련 연구도 진행되고 있다[1]. 또한 빅데이터 측면에서 데이터 프로파일링과 정규 표현식을 이용하여 비정형 빅데이터의 품질을 관리하기 위한 연구도 진행되었다[3].

이러한 데이터 품질의 문제로 인한 신규 가치 창출과 인공지능 산업 활성화의 걸림돌을 해소하기 위해 데이터 품질을 향상시킬 수 있는 방법의 연구를 제안한다.

1.2 연구 목적

데이터품질을 진단하는 방법은 크게 도메인 기반 데이터 품질 진단과 업무규칙 기반 데이터 품질 진단이 있다. 이번 연구에서는 도메인 분류 작업을 자동화하여 수작업으로 도메인 분류 시 발생할 수 있는 오류를 예방하고, 작업 시간을 단축시켜 데이터품질을 진단하고 개선하는 핵심 업무에 집중할 수 있도록 도메인 분류를 머신러닝 알고리즘을 사용하여 자동화 하는 것이다.

Ⅱ. 이론적 배경

2.1 연구선정 배경 및 필요성

2.1.1 빅데이터 정의 및 특성

가트너는 빅데이터는 큰 용량, 빠른 속도, 그리고 높은 다양성을 갖는 정보 자산으로써 이를 통해 의사 결정 및 통찰 발견, 프로세스 최적화를 향상시키기 위해서는 새로운 형태의 처리 방식이 필요하다고 정의하였다. 일반적으로 빅데이터의 특성을 이야기 할 때 가트너에서 정의한 3대 요소인 볼륨(Volume), 속도(Velocity), 다양성(variety)을 핵심 특징으로 꼽는다.

볼륨(Volume)은 대용량 데이터를 의미한다. 과거에는 기가바이트 단위도 대용량이라고 했지만 현재는 수십 테라바이트에서 수십 페타바이트 이상 되는 데이터를 대용량 데이터라고 정의한다. 대용량 데이터가 발생하는 원인은 다양하지만 크게 나누자면 데이터를 저장하는 공간을 구매하기 위한 비용이 낮아지고, 스마트폰의 대중화라고 할 수 있다.

속도(Velocity)는 데이터가 발생되는 시점부터 저장되고, 유통, 수집, 분석되는 모든 단계에서 발생하는 시간을 의미한다. 데이터로부터 의미 있는 결과를 도출하기 위해서는 어느 한 순간도 지체되면 안 되고 적시에 제공되어야 한다. 빅데이터에서 속도는 데이터의 발생부터 활용까지 전체 범위에서 중요한 요소이다.

다양성(Variety)은 데이터의 유형이 다양해졌다는 것을 의미한다. 데이터 유형에 따라 정형(Structured) 데이터, 반정형(Semi-structured)데이터, 비정형(Unstructured)데이터로 나뉜다. 정형데이터는 구조화된 데이터로 일반적으로 관

계형 데이터베이스에 저장되는 데이터의 유형이라고 볼 수 있다.

빅데이터는 위와 같이 3가지 특성을 가지고 있는데 이 중 두 가지 이상을 포함한다면 빅데이터로 볼 수 있다.

2.1.2 데이터 품질

데이터 품질이란 다양한 자료에서 발생한 데이터를 이용하기 위해 데이터를 활용할 대상이 데이터에 대한 신뢰를 가질 수 있는 수준을 의미한다. 명재호 는 “데이터 품질관리란 조직이 운영하는 정보시스템과 데이터베이스를 활용하는 이용자의 기대를 만족시키기 위해 지속적으로 수행하는 데이터 관리 활동을 의미한다.”라고 정의하였다[2].

데이터 품질을 관리하기 위해서는 데이터 품질을 믿을 수 있는지 먼저 진단을 해야 한다. 데이터 품질 진단은 데이터를 소유하고 있는 조직이 데이터를 관리하기 위해서 데이터의 품질을 측정하고, 데이터 품질의 신뢰성이 낮은 원인을 파악하고 개선하는 과정을 의미한다[7]. 데이터 품질을 진단하는 종류는 데이터 값 진단, 데이터 구조 진단, 데이터 관리 프로세스 진단이 있다. 데이터 값 진단은 데이터베이스의 테이블, 칼럼, 코드, 관계, 업무 규칙 등을 기준으로 데이터를 분석하여 품질을 진단하는 것이다.

데이터 구조 진단은 데이터 구조 무결성, 구조 표준화, 관리 수준, 변경 관리 등 데이터 설계 관점에서 데이터 품질을 진단하는 것이다. 권성호는 기존 정보를 텍스트 마이닝을 통해 표준화 방법에 관한 연구를 통해 데이터의 신뢰도를 증가 시킬 수 있다고 하였다[8]. 데이터 관리 프로세스 진단은 데이터 관리 프로세스 단계에서 문제점을 찾고 개선하는 것을 의미한다.

2.1.3 도메인 기반 데이터 품질 관리

도메인이란 관계형 데이터베이스에서 테이블에 포함된 각 칼럼들이 모델링 과정에서 정의되는 고유한 특성을 의미한다. 도메인 분석은 칼럼 정보를 기반으로 도메인을 분류하는 방법이다. 도메인 분석을 수행하면 각 칼럼에 알맞은 도메인을 정의하며 데이터는 무결성을 유지할 수 있다.

도메인은 11개로 분류되며 데이터베이스의 각 칼럼에 부합하는 도메인을 정의 시 데이터 품질을 향상에 기여할 수 있다.

번호 도메인은 숫자와 문자의 조합으로 구성되며 사전에 정의된 번호 체계에 따라 정의된다. 금액 도메인은 돈의 액수를 의미하는 값으로 각 국가의 화폐단위에 일치하는 숫자 값으로 구성된다. 명칭 도메인은 하나의 개체를 다른 개체와 구분하기 위해 각 개체에 붙이는 이름이다. 수량 도메인은 숫자로 정의된 항목을 의미하고 수량의 최대, 최솟값을 관리하여 데이터의 품질을 관리할 수 있다. 분류 도메인은 2~3개의 구분 값으로 구성된 도메인으로 단순한 분류를 의미한다. 3개를 초과하는 값으로 구성되어 있다면 코드 도메인을 사용하는 것이 좋다. 날짜 도메인은 날짜 값으로 구성된 데이터를 의미한다. 데이터베이스에서 정의하는 DATE, TIMESTAMP, DATETIME과 같은 날짜 데이터 타입은 기본적으로 날짜 도메인으로 정의가 가능하며 문자형 데이터 타입에서도 날짜를 의미하는 칼럼을 정의할 수 있다. 율 도메인은 금리, 이율, 비율, 환율, 백분율과 같은 용도로 사용되며 각 업무적 의미에 따라 다양하게 사용되므로 다양한 계산식이 있으므로 업무 성격에 맞게 진단해야 한다. 내용 도메인은 어떤 개체나 행동에 대한 설명하는 내용 포함하는 데이터이다. 내용 도메인은 비정형 데이터로 다양한 숫자, 텍스트 등 다양한 값으로 구성될 수 있다. 코드 도메인은 사전에 정의한 기준으로 데이터를 분류하는 것이다. 코드는 코드와 값으로 구성되어 있다. 코드 도메인 진단 시 데이터베이스에서 사용하는 코드가 기존에 등록되어 있는지 여부로 데이터 품질을 확인할 수 있다. 키(Key) 도메인은 두 개 이상의 칼럼의

〈표 1〉 도메인 분류[2]

도메인 분류	도메인 예시	점검내용
번호	주민등록번호, 사업자등록번호, 우편번호, 고객번호, 계좌번호	번호 관련 데이터의 패턴 및 체크비트 진단
금액	금액, 세금, 가격, 단가, 비용, 요금, 잔액, 총액	금액 관련 데이터의 허용범위 진단
명칭	명, 주소, ID, 장소, 고객명, 영문 고객명, URL, E-MAIL, IP	명칭 관련 데이터의 패턴 및 길이 진단
수량	건수, 매수, 회차, 개수, 거리, 규모, 길이, 무게, 속도, 횟수, 평형, 면적, 온도	수량 관련 데이터의 허용범위 진단
분류	여부, 유무, 구분, 상태	분류 관련 데이터의 표준정의 값 진단
날짜	연월, 연, 연월일, 시, 분, 초, 일, 반기, 분기	날짜 관련 데이터의 허용범위 및 유효값 진단
율	금리, 이율, 비율, 환율, 백분율	비율(%) 관련 데이터의 허용범위 진단
내용	내용, 비고, 설명, 정보, 요약	내용 관련 데이터의 적용언어 패턴 진단
코드	개별코드, 통합코드	코드 관련 데이터의 코드값 진단
키	일차키, 외래키	키 관련 데이터의 참조 무결성 진단
공통	데이터 표준화	데이터 표준 준수여부 진단

데이터 품질을 진단 시 사용한다. 데이터베이스에서 기존에 설계된 칼럼 간의 관계를 기준으로 데이터 품질을 진단 할 수 있다. 공통 도메인은 모든 도메인에서 공통으로 사용할 수 있다. 데이터 표준화 여부를 공통 도메인으로 정의하였다[2]. <표 1>과 같이 도메인별 예시와 점검내용을 정의하였다.

2.2 데이터 전처리

데이터전처리란 분석 목적에 맞는 데이터를 수집한 후 분석이 가능하도록 데이터를 축소, 제거, 수정 등과 같은 단계를 거쳐 최상의 분석 결과를 도출하기 위한 과정을 의미한다.

〈표 2〉 데이터 전처리의 5가지 단계

단계	작업
데이터 클리닝 (Data cleaning)	결측값을 채우고, 잡음(Noise)을 제거하고, 이상치를 탐지하여 제거함
데이터 통합 (Data integration)	다양한 소스에서 입력된 데이터를 합침
데이터 변환 (Data transformation)	기존에 존재하는 변수를 이용하여 새로운 변수를 생성함
데이터 축소 (Data reduction)	분석 결과가 이전과 비슷한 범위 안에서 데이터 축소함
데이터 이산화 (Data discretization)	성적과 같은 연속형 변수를 수, 우, 미, 양, 가와 같은 이산형 변수로 변환함

데이터전처리는 위의 표와 같이 5단계로 진행하는데 첫 번째로 데이터 클리닝(Data Cleaning) 단계이다. 먼저 결측값을 제거하는 방법은 7가지로 나눌 수 있다. 첫 번째 방법은 결측값이 발생한 모든 레코드를 무시하거나 분석에 사용할 변수 중 결측값이 있는 레코드만 무시하는 방법이 있다. 두 번째 방법은 결측값을 수작업으로 채우는 방법이 있다. 세 번째 방법은 전역상수(global constant)를 채우는 방법이 있다. 네 번째 방법은 해당 변수의 평균값을 사용해서 입력할 수 있다. 다섯 번째 방법은 해당 레코드와 동일한 그룹에 속하는 모든 데이터의 평균값을 적용할 수 있다. 여섯 번째 방법은 가장 많이 존재하는 값으로 결측값을 대체한다. 일곱 번째 방법은 Regression과 Logistic Regression과 같은 예측모델을 사용하는 방법이다. 결측값을 제외한 레코드를 기반으로 학습한 모델을 결측값이 존재하는 레코드에 적용하여 예측 값을 구할 수 있다.

잡음이란 측정된 변수에서 발생하는 오류나 오차 값을 의미한다. 오류나 오차로 인해 데이터의 의미가 훼손되는 것을 막기 위해서 데이터 평활화 기법을 사용할 수 있다.

데이터 평활화의 기법은 구간화(Binning), 회귀(Regression), 군집화(Clustering)와 같이 세 가지가 있다. 구간화는 데이터를 정렬한 후 몇 개의 버킷(bucket)또는 빈(bin)으로 분할하여 평활화(Smoothing)하는 방법이다. 회귀는 회귀함수를 이용하는 것으로써 선형관계가 있는 두 개의 변수로 예측 모델을 만들고 한 개의 기준이 되는 변수를 이용하여 평활화를 하려는 변수의 값을 찾는 것이다. 다중 회귀분석은 선형 회귀분석의 확장의 개념으로 두 개 이상의 변수를 사용하여 다른 한 개의 변수 값을 찾는 것이다. 군집화는 유사한 특성을 가진 데이터를 묶어서 그룹을 만드는 것이다. 이때 그룹에 묶이지 않는 값들이 발생한다면 이상치(Outlier)라고 볼 수 있으므로 조치가 필요하다.

두 번째는 데이터 통합(Data Integration) 단계이다. 에서는 개체 식별 문제(Entity Identification Problem)를 해결하는 것이 핵심이다. 이러한 문제를 해결하기 위해서는 데이터 값의 특성과 메타데이터를 이용하여 동일한 의미인지 파악해야 한다. 단순히 메타데이터를 이용할 뿐 아니라 메타데이터 요소별로 가중치를 적용하여 품질측정의 정확도를 높일 수도 있다[4]. 중복(Redundancy)도 데이터 통합에 중요한 고려사항이다. 유도 속성(derived attribute)은 다른 속성으로부터 유도된 것을 의미한다.

상관분석은 두 개의 변수에 대한 상관관계를 분석하여 상관도가 높은 경우 한 개의 변수를 제거할 수 있다. 상관분석은 두 변수가 관련이 있는지 분석을 하는 것을 의미하며 연관성의 측도는 아래 식 (1)과 같은 피어슨 상관계수(Pearson correlation coefficient)를 이용하여 측정할 수 있다.

$$r_{X,Y} = \frac{cov(X,\ Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

피어슨 상관계수는 X와 Y의 공분산을 X와 Y의 표준편차로 나눈 값으로 구할 수 있다. X와 Y의 공분산은 아래 식 (2)와 같이 X와 Y의 값을 각각의 평균값과의 차이를 구하여 합한 값을 모집단의 수인 m으로 나누어 구할 수 있다.

$$E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum_{i=1}^{m}(X_i-\mu_X)(Y_i-\mu_Y)}{m} \quad (2)$$

세 번째는 데이터 축소(Data reduction)로 크게 차원 축소(Dimensionality reduction)와 수치 축소(Numerosity reduction)로 나눌 수 있다. 많은 차원이 있다는 것은 그만큼 많은 양의 데이터가 있다는 것이고 데이터도 기하급수적으로 증가할 수 있다는 것을 의미한다. 또한 고차원적인 데이터에서는 분석의 결과를 얻기가 쉽지가 않다 이러한 문제를 제거하기 위해서 의미있는 차원들만 남기도 나머지 차원들을 제거하는 것을 차원 축소라고 한다. 수치 축소는 데이터 값을 대치 값으로 변경하는 것을 의미한다. 수치 축소는 모수적 모형(Parametric model)과 비모수적 모형(non-parametric model)로 나누어진다. 데이터 축소의 핵심은 데이터 축소 후의 데이터 분석 결과와 데이터 축소 전의 분석 결과가 최대한 유사하게 나와야 한다는 것이다.

차원축소(Dimensionality Reduction)는 원본 데이터의 손실 없이 축소하는 무손실 차원 축소와 데이터가 손실되는 손실 차원축소로 나눠진다. 그 중 웨이블릿 변환(Wavelet transform)과 주성분분석(Principal components)이 가장 효과적인 손실 차원축소 방법이다.

웨이블릿 변환 중 이산 '웨이블릿 변환(Discrete wavelet transform)은 벡터 X를 다른 수치적 벡터(Numerically Vector)X'로 변환하는 것을 의미하고 두 벡터의 길이는 같다[5].

주성분분석은 고차원의 데이터 중 서로 상관성이 높은 여러 변수들을 선형 조합하여 새로운 변수를 생성하는 것을 의미한다.

수량축소(numerosity reduction)는 데이터의 양

을 줄이는 기법이다. 수량축소에는 표본추출, 히스토그램, 클러스터링을 이용하는 방법이 있다. 먼저 표본추출은 많은 데이터에서 일부 데이터를 샘플링하는 방식이다.

히스토그램은 구간화(Binning)을 사용하여 데이터 분포의 근사치를 구하는 데이터 축소의 전형적인 형태이다.

히스토그램은 희소 데이터나 밀집 데이터 모두에 효과적이며, 비대칭적 데이터와 균일한 데이터 모두에 매우 효과적이다.

군집화(Clustering)는 데이터 레코드를 객체로 간주하고, 각 객체들을 군집(cluster)이라는 그룹으로 나눈다. 한 군집 내 객체들과는 유사하면서도 다른 군집 내 객체들과는 유사하지 않도록 군집화한다. 유사성은 공간 내에서 객체들이 어떻게 가까운지의 관점에 따라 거리 함수에 기반을 두어 정의된다[5].

네 번째로 데이터 변환(Data transformation)은 분석에 적절한 포멧으로 데이터를 바꾸는 것을 의미한다. 데이터 변환 방법은 구간화, 회귀, 군집화와 같이 데이터의 잡음을 제거하는 평활(Smoothing), 데이터를 월별, 연도별로 그룹핑하는 집계(Aggregation), 기존의 속성 집합에서 새로운 속성을 만드는 속성구성(Attribute construction), 정해진 구간 안에 데이터를 존재하도록 변환하는 정규화(Normalization), 수치형 데이터를 구간 라벨로 대체하는 이산화(Discretization), 도로명과 같은 속성을 상위(예: 시, 국가) 레벨 개념으로 일반화 시키는 개념 계층(Conceptual hierarchy)이 있다.

2.3 머신러닝

머신러닝은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 나누어진다. 지도학습은 얻고자 하는 답을 포함하는 학습 데이터(Training Data)를 사용하여 모델을 만드는 것이고, 비지도학습은 이러한 학습데이터를 제공하지 않고 모델을 만드는 것을 의미한다.

지도학습은 학습데이터를 이용하여 모델이 생성되는데 이 모델에 대한 평가를 통해서 최적의 파라미터를 찾는다. 이와 같은 평가하기 위해서 교차 검증을 이용하는데 이를 위해 세 가지의 데이터 셋이 필요하다. 첫 번째는 훈련 셋(Training Set)으로 학습에 사용할 데이터이고, 두 번째로 검증 셋(Validation Set)으로 앞에서 만든 모델이 최적의 모델인지 검증에 사용된다. 마지막으로 테스트 셋(Test Set)으로 실제 발생하는 데이터라고 가정하고 정확도를 확인하는 단계이다. 생성된 모델을 이용하여 아직 발생하지 않은 데이터에 대해서 검증을 할 수 없기 때문에 학습 데이터의 일부를 사용하여 미래에 발생할 데이터에 대한 예측을 진행한다. 훈련 셋과 테스트 셋만 있는 경우 데이터 셋의 비율은 70% 대 30% 정도이며, 훈련 셋, 검증 셋, 테스트 셋으로 나눌 경우 60% : 20% : 20% 정도로 하는 것이 일반적이다. 하지만 도메인 특성이라 데이터양에 따라 적합한 비율로 조정이 가능하다.

이렇게 <그림 1>와 같이 교차 검증을 통해서 모델의 정확도(Accuracy)와 정밀도(Precision) 그리고 재현율(Recall)을 측정할 수 있다.

정확도는 전체 데이터 중 실제 참과 거짓을 맞춘 비율이고 정밀도는 참이라고 예측한 데이터 중 실제 참인 데이터를 맞춘 경우를 의미한다. 마지막으로 재현율은 실제 참인 데이터 중 참이라고 예측한 경우를 의미한다.

		실제 결과	
		True	False
예측 결과	True	True positive	False positive
	False	False negative	True negative

〈그림 1〉 Confusion Matrix 정의

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3)$$

$$\Pr ecision = \frac{tp}{tp + fp} \quad (4)$$

$$Recall = \frac{tp}{tp + fn} \quad (5)$$

다시 지도학습의 알고리즘은 크게 분류(Classification)와 예측(Prediction) 모델이 있고 비지도 학습은 군집(Clustering) 모델이 있다.

분류 모델에는 여러 가지 알고리즘이 있는데 대표적으로 의사결정나무(Decision Tree), k-최근접 이웃 알고리즘, 서포트 벡터 머신(Support Vector Machine)이 있다. 예측 모델에는 대표적으로 회귀(Regression)모델이 있다.

분류와 예측은 학습데이터가 필요한 지도학습 모델이지만 모델에서 제공하는 결과 값에는 차이가 있다. 분류 모델은 결과 값이 범주형(Categorical) 데이터이지만 예측 모델의 결과는 수치형(Numerical) 데이터이다.

Ⅲ. 연구 및 결과

3.1 연구모델 설계 및 방법론

데이터 품질을 진단하는 다양한 방법이 존재한다. 그 중 데이터 도메인을 기반으로 하는 데이터 품질 진단 방법을 선정하였다. 그 이유는 다음과 같다. 첫 번째, 데이터 도메인의 중요성이다. 데이터 도메인이 정확하게 정의되었다면 전혀 다른 의미의 데이터가 입력되는 문제가 발생하지 않을 것이다. 두 번째, 도메인은 11개로 구성되어 있으며 머신러닝의 알고리즘을 이용하여 분류가 가능할 것으로 예상했다. 세 번째, 도메인의 분류를 판단하기 위해서 다양한 속성을 분석해야 하는데 많은 시간이 소요된다. 머신러닝 알고리즘을 이용하여 빠르게 도메인을 분류한다면 데이터 품질을 진단하는 시간을 단축할 수 있을 것으로 예상된다. 연구는 크게 4단계로 나누어 진행한다.

3.1.1 학습 데이터 확보

첫 번째 단계는 데이터 확보 단계이다. 머신러닝 알고리즘 중 지도학습을 이용하여 도메인 분류 모델을 생성할 것이다. 지도학습을 진행하기 위해서는 학습에 필요한 데이터가 필요하다. 기존에 데이터 품질 진단을 완료한 데이터 중 학습에 적절한 데이터를 확보한다. 본 연구에서는 <그림 2>와 같이 L사의 도메인 기반 품질 진단 결과 데이터 7만 개를 학습에 필요한 데이터로 이용하였다.

테이블한글명	컬럼명	컬럼한글명	데이타타입	도메인
재무재무제표	FIRM_CD	기업코드	VARCHAR2(7)	코드
재무재무제표	FSCL_YM	회계년월	CHAR(6)	날짜
재무재무제표	SETTLE_TY	결산구분	CHAR(1)	코드
재무재무제표	ACCT	계정과목	CHAR(10)	코드
재무재무제표	ACC_AMT	계정금액	NUMBER(13)	금액
재무재무제표	AMD_YMD	수정일	VARCHAR2(8)	날짜
재무재무제표	INPUT_YMD	입력일	VARCHAR2(8)	날짜
재무재무제표	SGNPGS_SEQ	결재일련번호	CHAR(15)	번호
재무재무제표	SGN_TYPE	결재상태	CHAR(1)	코드
재무재무제표	CRE_USER	생성USER	CHAR(7)	ID
재무재무제표	CRE_TIME	생성시간	CHAR(14)	날짜
재무재무제표	JOB_USER	작업USER	CHAR(7)	ID
재무재무제표	JOB_USERIP	작업자IP	CHAR(15)	내용
재무재무제표	JOB_TIME	작업시간	CHAR(14)	날짜
국내운용역	MNGM_TLG_HMRS_CODE	운용전문인력코드	VARCHAR2(20)	코드
국내운용역	BRTH	생년월일	CHAR(8)	날짜
국내운용역	KRN_NM	한글명	VARCHAR2(300)	명칭
국내운용역	ENG_NM	영문명	VARCHAR2(300)	명칭
국내운용역	IMC_CODE	운용사코드	CHAR(3)	코드
국내운용역	USE_YN	사용여부	CHAR(1)	플래그
국내운용역	SGNT_SEQ	결재일련번호	CHAR(15)	번호
국내운용역	SGNT_STS_DSCD	결재상태구분코드	CHAR(1)	코드
국내운용역	CRPR	생성자	VARCHAR2(20)	ID
국내운용역	CRT_DTTM	생성일시	CHAR(14)	날짜
국내운용역	WRKR	작업자	VARCHAR2(20)	ID
국내운용역	WRKR_IP	작업자IP	CHAR(15)	내용
국내운용역	WRK_DTTM	작업일시	CHAR(14)	날짜

〈그림 2〉 학습에 이용할 L사의 칼럼별 도메인 정의

3.1.2 데이터 전처리

두 번째 단계는 데이터 전처리(Data Preprocessing)를 수행한다. 앞의 관련 연구에서 언급한 것처럼 데이터 전처리에는 다양한 단계가 있다. 이 중에서 데이터 클리닝, 데이터 통합, 데이터 변환, 데이터 축소 단계를 수행한다. <그림 3>과 같이 필수적으로 존재해야 하는 데이터는 예를 들어, 물리 테이블 명, 물리 칼럼명, 데이터 타입, 데이터 사이즈 등이 있으며 선택적 데이터로는 논리 테이블명, 논리 칼럼명, 코멘트 등이 있다.

데이터 클리닝 단계에서는 선택적 입력 데이터 중 논리 칼럼명의 결측치에 대해서 “기타” 값으로 변경하였다. 데이터 통합 단계에서는 다양한 데이터베이스에서 데이터 품질 진단을 수행한 데이터를 통합하였다. 데이터 통합 시 가능

테이블한글명	컬럼명	컬럼한글명	데이타타입
재무재무제표	FIRM_CD	기업코드	VARCHAR2(7)
재무재무제표	FSCL_YM	회계년월	CHAR(6)
재무재무제표	SETTLE_TY	결산구분	CHAR(1)
재무재무제표	ACCT	계정과목	CHAR(10)
재무재무제표	ACC_AMT	계정금액	NUMBER(13)
재무재무제표	AMD_YMD	수정일	VARCHAR2(8)
재무재무제표	INPUT_YMD	입력일	VARCHAR2(8)
재무재무제표	SGNPGS_SEQ	결재일련번호	CHAR(15)
재무재무제표	SGN_TYPE	결재상태	CHAR(1)
재무재무제표	CRE_USER	생성USER	CHAR(7)
재무재무제표	CRE_TIME	생성시간	CHAR(14)
재무재무제표	JOB_USER	작업USER	CHAR(7)
재무재무제표	JOB_USERIP	작업자IP	CHAR(15)
재무재무제표	JOB_TIME	작업시간	CHAR(14)
국내운용역	MNGM_TLG_HMRS_CODE	운용전문인력코드	VARCHAR2(20)
국내운용역	BRTH	생년월일	CHAR(8)
국내운용역	KRN_NM	한글명	VARCHAR2(300)
국내운용역	ENG_NM	영문명	VARCHAR2(300)
국내운용역	IMC_CODE	운용사코드	CHAR(3)
국내운용역	USE_YN	사용여부	CHAR(1)
국내운용역	SGNT_SEQ	결재일련번호	CHAR(15)
국내운용역	SGNT_STS_DSCD	결재상태구분코드	CHAR(1)
국내운용역	CRPR	생성자	VARCHAR2(20)
국내운용역	CRT_DTTM	생성일시	CHAR(14)
국내운용역	WRKR	작업자	VARCHAR2(20)
국내운용역	WRKR_IP	작업자IP	CHAR(15)
국내운용역	WRK_DTTM	작업일시	CHAR(14)

〈그림 3〉 데이터 전처리 전 테이블 정보

한 다양한 산업별 데이터 품질 진단 결과 데이터를 사용하면 모델의 분류 성능을 향상 시킬 수 있다. 그 이유는 산업별로 사용하는 용어와 업무가 다르기 때문에 한 가지 산업군의 데이터를 이용하여 모델 생성 시 타 산업 군에서 사용하는 칼럼에 대해서 인식을 못할 수 있다. 예를 들어 금융업의 데이터로 학습할 경우 제조업 데이터에 발생하는 “너비”, “높이” 같은 칼럼에 대해서 “수”라는 도메인으로 분류 할 수 없다. 이러한 문제점을 해결하기 위해 두 가지 방법을 사용할 수 있는데 먼저 다양한 산업 군별 데이터를 통합하여 한 개의 분류 모델을 생성할 수 있다 다른 방법으로는 데이터베이스에서 공통으로 사용하는 칼럼을 통합하여 공통 분류 모델로 생성하고, 각 산업 군별 모델을 생성할 수 있다. 첫 번째 방법인 통합 모델을 사용할 경우 별도의 관리가 필요하지 않기 때문에 편할 수 있지만, 많은 데이터로 인해 모델링을 위한 시간이 많이 소요되고 산업별로 다른 의미로 분리되는 도메인에 대해서 분리하여 분류 할 수 없다. 산업별 모델을 생성하는 경우 별도로 관리하는 불편함은 있지만 산업별 특성을 고려하여 모델을 만들 수 있는 장점이 있다. 데이터 변환은 특징 추출(Feature Engineering)이라고도 하며 기존에 존재하는 변수를 사용하여 새로운 변수를 생성하는 것이다. 본 연구에서는 9개의 변수를 사용하는데 이 중 8개의 파생변수를 사용한다. 이번 연구에서 변수를 만들기 위해 물리 칼럼명, 논리 칼럼명, 데이터 타입, 데이터 값, 데이터 수를 사용하였다. 최종 선택한 9개의 변수는 아래와 같다.

〈표 3〉 기존 변수와 파생변수 정의

변수	설명
데이터 타입	Integer, Character, Text, Timestamp와 같은 데이터 값을 구분할 수 있는 변수
논리 칼럼명 접미사	고객명, 상품명, 주민등록번호, 매출액과 같이 사람이 인식할 수 있는 칼럼의 정의하는 명칭이 논리 칼럼명이고 접미사는 논리 칼럼명의 마지막 형태소를 의미
물리 칼럼명 접미사	CUST_NAME, PRODUCT_NAME, SALES_AMOUNT와 같이 컴퓨터가 이해할 수 있는 칼럼의 이름을 의미하며 접미사는 NAME, AMOUNT와 같이 마지막에 사용된 단어 또는 “_”와 같은 특수 기호로 구분되는 마지만 단어를 의미
소수점 포함 여부	숫자형 데이터 타입인 칼럼 중 데이터에 소수점 포함 여부에 따라 분류
날짜 데이터 여부	DATE, TIMESTAMP와 같이 날짜 데이터 타입은 날짜 데이터 여부가 Y이며, 문자형 데이터 칼럼 중에도 데이터가 날짜형으로 구성되어 있으면 Y값으로 정의
숫자 여부	INT, FLOAT와 같은 숫자형 데이터 타입으로 정의된 칼럼은 Y값을 가짐
데이터 중복 제외 건수	고유한 데이터 건수를 의미하며 예를 들어 100건의 데이터 중 모든 데이터가 Y와 N으로만 구성되어 있다면 2건이 중복 제외 건수 임
텍스트 200자 초과 여부	명칭 도메인과 내용 도메인을 구분하기 위한 변수로 문자형 데이터 타입을 가진 칼럼의 데이터가 최대 200자를 초과하면 Y값을 가짐
텍스트 자릿수 변동 여부	코드 도메인을 판별하기 위한 변수로 칼럼에 존재하는 데이터의 자리수가 변동되는지 확인함

9개의 변수를 선택 이유는 다음과 같다. 데이터 타입은 데이터가 숫자, 날짜, 텍스트인지 분류하기 위해 가장 기본이 되는 특징이다. 예를 들어 데이터 타입이 DATE와 같은 날짜형이라면 “수”, “금액”과 같은 도메인으로 분류하지 않는다. 논리 칼럼명을 정의할 경우 일반적으로 “단어”+“단어”와 같은 패턴을 이용한다. 예를 들어 매출원가라는 “금액”이라는 도메인은 “매출”+“원가”로 구성되어 있고, “고객명”이라는 “명칭” 도메인은 “고객”+“명”으로 구성되어 있다. 이와 같이 단어와 단어의 조합으로 구성된 칼럼명의 경우 마지막 칼럼의 단어를 이용하여 도메인을 판단하는 변수를 만들 수 있다. 논리 칼럼명 접미사는 이와 같이 칼럼명의 마지막 단어를 의미한다. 물리 칼럼명은 일반적으로 영어 단어를 사용하여 만든다. 논리 칼럼명과 차이점은 물리 칼럼명을 영어 단어를 사용하는데 영어 단어를 그대로 사용하기도 하고 준말을 사용하는 특징이 있다. 예를 들어 논리 칼럼명이 “고객명”이라고 하면 물리 칼럼명은 “CUSTOMER_NAME”, “CUST_NAME”, “CUST_NM”과 같이 생성할 수 있다. “CUST”는 “CUSTOMER”의 준말이고, “NM”은 “NAME”의 준말을 의미한다. 여기서 중요한 점은 물리 칼럼명도 논리 칼럼명과 같이 단어와 단어의 조합을 사용하지만 단어와 단어 사이에 “_”와 같은 특수문자를 이용하여 단어를 구분하고, 준말을 사용하는 경우에도 다른 단어와 중복되지 않도록 한다. 앞에서 예를 들었던 “NM”같은 경우 “NAME”의 준말로 정의된다면 다른 단어의 준말로 사용하지 않는다. 물리 칼럼명 접미사는 이와 같이 물리 칼럼명이 단어와 단어로 조합된 경우 마지막 단어를 의미한다. “CUST_NM”의 경우 “NM”이 논리 칼럼명의 접미사이다. “NM”이라는 접미사는 “명칭”이라는 도메인으로 <그림 4>과 같이 정의할 수 있다.

데이타타입	논리 칼럼명 접미사	소수점여부	물리 칼럼명 접미사	날짜여부	숫자여부	유니크건수	글자200자 이상 초과 여부	자릿수 변동 여부	도메인
CHAR	일	N	YMD	Y			N		날짜
CHAR	코드	N	CD				N	N	코드
NUMBER	BS	N	ETC		Y		N		금액
NUMBER	액	Y	AMT		Y		N		금액
NUMBER	증가	Y	AMT		Y		N		금액
CHAR	수준	N	ETC				N	N	코드
CHAR	상태	N	ETC				N	N	코드
CHAR	번호	N	SEQ				N		번호
CHAR	상태	N	TYPE				N	N	코드
CHAR	ER	N	ETC				N	Y	ID
CHAR	시간	N	TIME	Y			N		날짜
CHAR	IP	N	ETC				Y		내용
CHAR	코드	N	ETC				N	N	코드
NUMBER	순번	N	SEQ				N		번호
CHAR	순번	N	SEQ				N		번호
CHAR	구분	N	TYPE				N	N	코드
CHAR	외화	N	CD				N	N	코드
CHAR	래처	N	FIRM				N	N	코드
NUMBER	액면	Y	QTY		Y		N		수
NUMBER	액면	N	QTY		Y		N		수

〈그림 4〉 데이터 전처리 후 테이블 정보

3.1.3 분류 모델 개발

세 번째 단계는 머신러닝을 이용한 분류 모델 생성이다. 데이터 전처리 단계를 거친 데이터를 이용하여 도메인 분류 모델을 생성한다. 분류 모델을 만들기 위해서 본 연구에서는 최근접 이웃 알고리즘, 의사결정 나무, 랜덤 포레스트, 서포트 벡터 머신과 같은 4가지 지도학습 알고리즘을 사용하여 분류 모델을 개발하고 최종적으로 정확도가 높은 모델을 선정하였다.

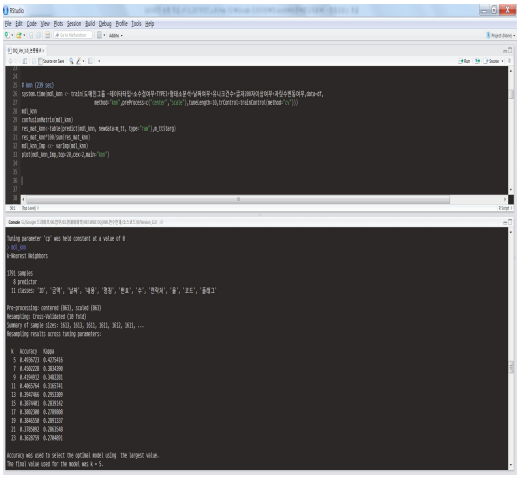
3.1.4 모델 검증 및 보완

마지막 단계에서는 모델 검증 및 보완을 수행한다. 모델 검증 및 보완이란 모델에 대한 결과가 나온 원인을 분석하고, 결과에 대한 원인이 확인되면 하이퍼파라미터를 튜닝하거나 학습 데이터를 추가하여 정확도를 높이거나 새로운 파생변수를 찾아 모델에 적용하는 방법이 있다.

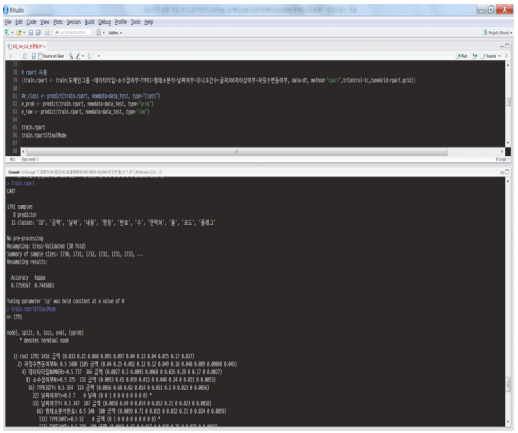
3.2 연구 결과

첫 번째로, <그림 5>와 같이 최근접 이웃 알고리즘을 사용한 결과 k 값은 5개이고 Accuracy는 49.3%가 최고의 정확도를 나타냈다.

두 번째로, <그림 6>과 같이 의사결정 나무 알고리즘을 사용한 결과 77.6%의 Accuracy를 나타낸다. 의사결정 나무 알고리즘의 장점은 학습 데이터가 입력되어 어떤 과정을 걸쳐서 결과가 나오는지에 대한 검증이 가능하다.



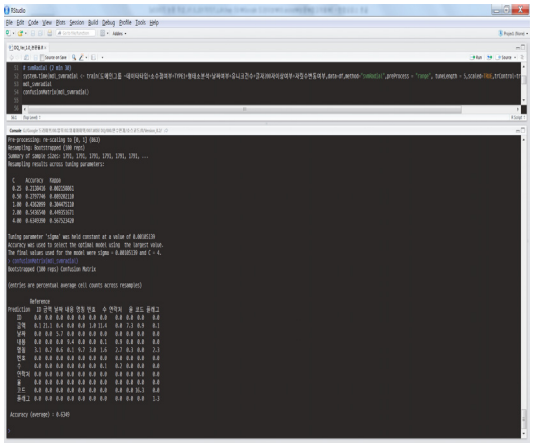
〈그림 5〉 최근접 이웃 알고리즘 모델 정확도



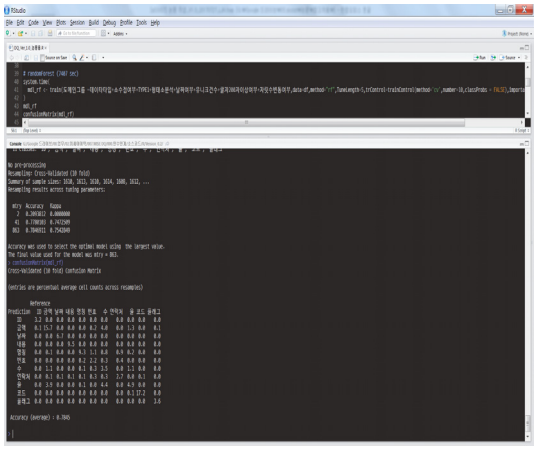
〈그림 6〉 의사결정 나무 알고리즘 모델 정확도

세 번째로, <그림 7>과 같이 서포트 벡터 머신 알고리즘을 사용한 결과 63.5%의 Accuracy를 나타낸다. 본 알고리즘의 예측 결과를 Confusion matrix로 확인해보면 ID 값을 금액과 명칭으로 잘못 예측하는 것으로 나타난다. 이에 따라 ID 값에 대한 예측 정확도를 높이기 위해 데이터를 추가적으로 학습하거나 새로운 파생변수를 생성해야 할 것으로 판단된다.

네 번째로, 랜덤 포레스트 알고리즘을 사용한 결과 78.5%의 Accuracy를 나타냈다. 본 연구에서 사용한 알고리즘의 수행 결과 가장 높은 정확도를 보여준다. <그림 8>과 같이 랜덤 포레스트 모델 결과의 Confusion matrix를 확인해본 결과 서포트 벡터 머신 알고리즘에서 분류를 잘하지 못하던 ID에 대해서는 높은 정확도를 나타냈으나 “율” 도메인에 대해서 정확하지 못한 결과를 나타낸다. 이러한 결과가 나오는 원인은 잘못된 분류 결과를 확인해본 결과 “금액”, “수”와 같이 숫자로 구성된 도메인으로 분류된다는 특징을 발견했다.



〈그림 7〉 서포트 벡터 머신 알고리즘 모델 정확도



〈그림 8〉 랜덤 포레스트 알고리즘 모델 정확도

최종적으로 모델을 선정한 후 모델 검증 및 보완을 수행한다. 앞 단계에서 만들어진 모델을 검증하고 보완하여 정확도를 향상 시킨다. 예를

데이타 타입	논리 칼럼명 접미사	소수점 여부	물리 칼럼명 접미사	날짜 여부	숫자 여부	유니크 건수	글자 200자 이상 여부	자릿수 변동 여부	도메인	예측결과
NUMBER	포함	Y	ETC	N	Y	9.854	N	N	을	금액
VARCHAR2	기관	N	ETC	N	Y	1.254	N	N	을	명칭
NUMBER	일전	Y	ETC	N	Y	6.874	N	N	을	금액

〈그림 9〉 랜덤 포레스트 알고리즘 예측 결과 분석

들어 랜덤 포레스트 알고리즘에서 “율” 도메인을 “금액”, “명칭” 도메인으로 잘못 예측한 결과를 검증한 결과 가장 큰 오 분류의 원인은 논리 칼럼명 접미사가 도메인을 판단하기 어려운 데이터로 정의되어 있기 때문이다. <그림 9>에서 확인할 수 있는 것처럼 논리 칼럼 접미사가 “포함”, “기관”, “일전”으로 되어 있는 “율” 도메인 데이터를 확인할 수 있다. 이와 같이 예측 변수에서 부적절한 데이터로 인하여 잘못된 분류를 하는 경우 예외 처리를 통해 수작업으로 분류할 수 있도록 방안을 제시해야 한다.

Ⅳ. 결 론

본 논문은 빅데이터의 품질 진단을 자동화 모델 개발에 대하여 연구하였다. 자동화 모델 개발 절차는 데이터 확보, 데이터 전처리, 자동화 모델 개발, 모델 검증 및 보완과 같이 네 단계로 구성된다.

첫 번째, 모델 생성에 필요한 데이터를 확보해야 한다. 지도학습 알고리즘을 사용하기 때문에 모델링 시 학습에 필요한 올바른 데이터가 있어야 모델의 정확도가 향상된다. 두 번째는 모델링에 적합한 구조로 데이터를 변형하는 데이터 전처리 단계이다. 데이터 전처리는 데이터 클리닝, 데이터 통합, 데이터 변환, 데이터 축소, 데이터 이산화의 단계를 거쳐야 한다. 세 번째 단계는 모델링 단계이다. 모델링이란 데이터 전처리를 거친 변수를 사용하여 머신러닝 알고리즘을 이용한 모델을 생성하는 것이다. 본 연구에서는 k-NN, 의사결정나무, 서포트 벡터 머신, 랜덤 포레스트 알고리즘을 수행하여 가장 정확도가 높은 알고리즘을 최종 모델로 선정하였다. 마지막 단계는 모델 검증 및 보완 단계이다. 모델링을 통해 개발한 모델을 검증하여 정확도를 확인하고 파라미터를 수정하거나 학습 데이터를 보정하여 데이터의 정확도를 높이는 단계이다. 특이 본 연구는 빅데이터의 품질을 진단하는 것이므로 정확도가 낮은 결과를 신뢰할 수 없기 때문에 예측 결과의 확률을 확인하여 결과의 채택 여부를 결정하였다.

추후 연구 과제에서는 산업 군에 따른 데이터 표준화 정의 문제, 학습되지 않은 데이터 발생 시 나타나는 문제, 도메인별 다양한 값의 이상치를 찾는 문제에 대한 연구가 진행 되어야 할 것이다.

참 고 문 헌

[1] 김선호, 이창수, 이진우, “공공데이터 품질관리 성숙 수준에 대한 연구”, 대한산업공학회 추계학술대회논문집, pp.159-165, 2016.

[2] 명재호, 안희진, 이창수, 김성현, 임동진, 오경조, 이종규, 김선영, 최용준, 데이터 품질 가이드라인, 한국데이터진흥원, 2011.

[3] 이상기, 채철주, 홍의경, “데이터프로파일링과 정규 표현식 활용 비정형 과학기술 빅데이터 품질관리 방안”, 한국콘텐츠학회논문지, 제14권, 제12호, pp.486-493, 2014.

[4] 이용구, 김병구, “학술지 기사에 대한 메타데이터 품질의 계량화 방법에 관한 연구”, 정보관리학회지, 제28권, 제1호, pp.309-326, 2011.

[5] 이현호, R과 SQL을 활용한 실전 데이터전처리, 카오스북, 2016.

[6] 차경엽, 심광호, “공공부문 정보시스템 데이터의 신뢰성 점검기법 개발”, 한국통계학회 논문집, 제17권, 제5호, pp.745-753, 2010.
[7] 호진원, 이미영, “IT활용 감사의 효과성 향상을 위한 데이터 품질관리 방안 연구”, 한국사회와 행정연구, 제23권, 제4호, pp.31-53, 2013.
[8] Sungho Kwon, A Study on the Standardization Method of Inventory Item Master Data Using Text Mining and Standardization Approaches, Masters dissertation. University of Seoul, Seoul 2010.
[9] http://news.hankyung.com/article/2017060160031.
[10] http://biz.chosun.com/site/data/html_dir/2017/03/08/2017030801954.html.
[11] http://www.yonhapnews.co.kr/bulletin/2017/05/24/0200000000AKR20170524044700089.HTML.

저 자 소 개

이 진 형(Jin-Hyoung Lee)
・2016년～현재 : 위세아이텍 인공지능팀 수석
・관심분야 : 빅데이터 활용, 추천 시스템, 인공지능