http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.23, No.1, February 2018, pp.01-22
https://doi.org/10.7838/jsebs.2018.23.1.001

랜덤포레스트를 이용한 모기업의 하향 거래처 기업의 분류: 자동차 부품산업의 가치사슬을 중심으로

Classification of Parent Company's Downward Business Clients Using Random Forest: Focused on Value Chain at the Industry of Automobile Parts

김태진(Teajin Kim)*, 홍정식(Jeongshik Hong)**, 전윤수(Yunsu Jeon)***,
박종률(Jongryul Park)****, 안태욱(Teayuk An)*****

초 록

가치사슬은 경쟁우위 강화를 위한 전략적 도구로써 주로 기업수준, 산업수준에서 분석되어 왔다. 그런데 기업수준에서 가치사슬 분석을 수행하기 위해서는 분석 기업의 거래처 기업들이 그 기업의 가치 사슬에 속하는지의 여부에 따라 분류되어야 한다. 단일 기업에 대한 가치사슬 분류는 전문가들에 의해 원활히 수행될 수 있지만 다수의 기업을 대상으로 분류할 때는 많은 비용과 시간이 소요되는 등의 한계점이 따른다. 따라서 본 연구에서는 실거래 데이터를 기반으로 특정 기업의 거래처 기업들을 분류해서 가치사슬 기업을 자동적으로 도출해주는 모형을 제안하고자 한다. 총 19개의 거래 속성 변수를 실거래 데이터로부터 도출하여 기계학습의 입력 데이터의 형태로 가공하였고, 랜덤포레스트 알고리즘을 이용하여 가치사슬 분류 모형을 구축하였다. 자동차 부품 기업 사례에 본 연구 모형을 적용한 결과, 정확도 92%, F1-척도 76% 그리고 AUC 94%로 자동적 가치사슬 분류의 가능성을 확인하였다. 또한 거래집중도, 거래금액 그리고 거래처별 총 매출액 등과 같은 거래 속성들이 가치사슬에 속하는 기업들을 대표하는 주요 특성임을 확인하였다.

ABSTRACT

The value chain has been utilized as a strategic tool to improve competitive advantage, mainly at the enterprise level and at the industrial level. However, in order to conduct value chain analysis at the enterprise level, the client companies of the parent company should be classified according to whether they belong to it's value chain. The establishment of a value chain for a single company can be performed smoothly by experts, but it takes a lot of cost and time to build one which consists of multiple companies. Thus, this study

본 연구는 중소기업청의 기술혁신개발사업의 일환으로 수행하였음.

* First Author, Department of Industrial and Information Systems, Public Policy and Information Technology Professional Graduate School, SNUT(taejin@scp.re.kr)
** Corresponding Author, Department of Industry Information System Engineering, SNUT (hong@seoultech.ac.kr)
*** Co-Author, Department of Data Science, SNUT(yunsu3761@seoultech.ac.kr)
**** Co-Author, Department of Data Science, SNUT(jrpark924@seoultech.ac.kr)
***** Co-Author, Business on Communication, Ltd(taeukahn@hanmail.net)
Received: 2017-12-01, Review completed: 2018-02-14, Accepted: 2018-02-22

proposes a model that automatically classifies the companies that form a value chain based on actual transaction data. A total of 19 transaction attribute variables were extracted from the transaction data and processed into the form of input data for machine learning method. The proposed model was constructed using the Random Forest algorithm. The experiment was conducted on a automobile parts company. The experimental results demonstrate that the proposed model can classify the client companies of the parent company automatically with 92% of accuracy, 76% of F1-score and 94% of AUC. Also, the empirical study confirm that a few transaction attributes such as transaction concentration, transaction amount and total sales per customer are the main characteristics representing the companies that form a value chain.

키워드 : 가치사슬 분류, 전자세금계산서, 자동차 부품산업, 기계학습, 랜덤포레스트
Value Chain Classification, Electronic Tax Invoice, Industry of Automobile Parts, Machine Learning, Random Forest

1. 서 론

마이클 포터가 가치사슬 개념을 제안한 이래[32], 최근 가치사슬에 관한 연구는 주로 특정 산업이나 기업을 대상으로 하고 있다[5, 12, 22, 31]. 산업을 대상으로 한 연구의 예시로는 MICE 산업 내에서 시장수요 개발이나 MICE 개최 및 운용 등의 주 활동과 지원활동을 분석한 연구[12]나 문화콘텐츠 산업을 대상으로 기 구축된 가치사슬 구조의 각 기능별 기업들의 유형을 분석하는 연구[31] 등이 있다. 또한 현대자동차의 가치사슬 구조와 협력업체들의 관계를 분석한 기업 대상의 가치사슬 연구도 있다[5].

산업수준에서의 가치사슬 연구는 특정 산업의 주요 품목과 이들 품목의 가치사슬 구조를 분석하여 산업구조 분석 및 혁신 방안 수립 등이 주 연구 주제가 된다[25]. 반면, 기업 수준의 가치사슬 연구는 특정 기업의 가치사슬 구조를 통해 모 기업과 하청업체간 거래 및 상호 협력 형태를 분석하고 이를 통해 드러나는 기업의 경쟁력 및 혁신역량 도출이 주 연구 대상이 된다[5].

본 연구는 기업 수준의 가치사슬을 연구 대상으로 하고 있다. 그런데 기존의 기업중심 가치사슬에 대한 대부분의 연구들은 특정 기업의 가치사슬에 속하는 기업들이 알려져 있다는 전제를 가정으로 하고 있다[22, 5]. 물론 하나의 기업을 대상으로 할 경우 그 기업의 전략담당자와 구매담당자의 토의를 거쳐 그 기업의 생산품의 가치에 실질적으로 기여를 하는 주요 납품업체를 선정하면 이들 기업이 바로 그 기업의 가치사슬에 속하는 기업이 될 것이다. 따라서 특정 기업의 가치사슬을 연구대상으로 할 경우, 이 기업의 거래처 기업들이 가치사슬에 속하는 기업인지 아닌지를 판별하여 거래처 기업들을 분류하는 문제는 별도의 연구주제로 다루어지지 않았다. 그런데 때로는 기업 내부 담당자나 관련전문가의 토의를 거치지 않고 특정 기업의 가치사슬에 속하는 기업들을 파악할 필요성이 대두된다. 예를 들어, 중소기업의 신용평가나 투자적격 여부 등을 수행하는 투자자 입장에 서 보자. 이 경우 신용평가사나 투자사가 파악해야 할 대상 기업은 수만 개에 이를 수 있다. 이때 특정 대상 기업의 투자 적격여부를 판단하기 위해 대상 기업에 대한 자료만이 아니라 대상 기업의 가치사슬에

속하는 기업들을 알게 될 경우, 대상 기업의 투자적격 여부를 훨씬 수월하게 판단할 수 있다. 그러나 수만 개에 이르는 대상 기업의 가치사슬에 속하는 기업들을 도출하기 위해 관련 기업 담당자나 관련 전문가의 토의를 거쳐야 한다면 이로 인한 편익보다 비용이 더 클 것이다. 따라서 이 경우 특정 기업의 거래 데이터를 토대로 그 기업의 거래처 기업들이 가치사슬에 속하는 기업인지 여부를 판별하여 거래처 기업들을 분류하는 작업을 수행하는 알고리즘의 가치는 매우 클 것이다. 본 연구는 2010년부터 수집된 전자세금계산서 자료를 토대로 만들어진 특정 기업의 거래 데이터를 기반으로 그 기업과 거래하는 기업들을 그 기업의 가치사슬에 해당하는지의 여부에 따라, 거래처 기업들을 분류하는 모형을 제시하고자 한다. 이 방면의 연구는 필자가 아는 범위 내에서 거의 없는 실정이다. 논문 형태로 제시된 연구는 찾지 못하였고, 보고서 형태로 2016년 국내 연구가 수행된 적이 있다[19]. 또한 가치사슬보다는 폭넓은 개념인 하청업체를 사회연결망 분석도구를 활용하여 도출한 연구[15]와 가치사슬을 품목 간 관계를 통해 분류한 연구[20]가, 데이터에 근거해 정량적으로 가치사슬을 도출한 연구들이다.

본 논문은 특정 기업의 거래 데이터로 활용될 수 있는 전자세금계산서를 기반으로 그 기업의 거래처 기업들을 가치사슬에 속하는지의 여부에 따라 분류하는 문제를 다루며 이는 전형적인 기계 학습 문제의 하나인 분류 문제이다. 기계학습 문제를 분석하는 과정은 어느 모형을 선택하는지가 핵심 작업 중 하나인데, 본 논문은 기계학습 모형 중 파라미터 추정 작업을 수행하지 않아 활용의 용이성이 높되 성능이 우수한 랜덤포레스트 모형을 활용하고자 한다. 모형의 입력데이터는 전자세금계산서를 기반으로 가치사슬 기업의 주요 거래 속성들을 추출해 구성하였다. 본 연구의 기여도는 다음과 같다. 첫째, 특정 기업의 거래처 기업들을 전자세금계산서 데이터에 기반하여 가치 사슬 여부에 따라 분류하는 문제를 처음으로 다룬 연구이다. 둘째, 이러한 분류문제에 기계학습 알고리즘을 적용하기 위해 다양한 입력변수를 도출하는 작업을 수행하였다. 셋째, 실제 자동차 부품업체를 대상으로 관련 데이터를 토대로 랜덤포레스트 알고리즘을 적용하여 거래처 기업이 가치사슬 기업에 속하는지의 여부를 판별하는 작업이 제대로 수행될 수 있음을 보였다.

본 논문은 다음과 같이 구성된다. 제2장에서는 데이터를 기반으로 가치사슬을 분류한 기존 연구와 본 연구에서 활용된 기계학습 알고리즘에 대해 서술하며, 이어 제3장에서는 본 논문의 연구모형에 대해 기술한다. 제4장에서는 제3장에서 분류한 연구 모형을 자동차 부품 산업에 적용한 사례연구 결과를 제시하고, 제5장에서는 본 논문의 기여도와 한계점 그리고 추후 연구 등에 대해 기술된다.

2. 연구배경

이 절에서는 본 연구와 가장 근접한 연구들을 간략히 기술하고, 본 연구가 제시하는 알고리즘의 기본 모형인 랜덤포레스트 모형을 간략하게 기술한다.

2.1 거래 데이터를 활용한 가치사슬 분류

가치사슬에 대한 연구는 산업 수준이나 기업

수준에서의 가치사슬 분석에 대한 연구가 대부분이다[22, 12, 31, 5]. 본 논문이 연구 대상으로 하는 거래처 기업들을 가치사슬에 속하는지의 여부에 따라 거래처 기업들을 분류하는 문제는 필자가 아는 한 거의 다루어지지 않았다. 가치사슬 분석이 아닌 가치사슬 분류를 다룬 선행 연구들을 간략히 요약하면 다음과 같다.

한 산업의 가치사슬은 산업 내 속하는 품목들과 이 품목들의 관계를 통해 나타낼 수 있다. 이러한 품목들의 연결망을 거래 데이터를 통해 분류하여 한 산업의 가치사슬을 분석한 연구 사례로는 이래은 등[20]의 연구가 대표적이다. 분석 과정을 간략히 기술하자면 먼저 분석하고자하는 완제품을 선택하고 해당 제품을 생산하는 대표기업들을 전문가로부터 추천받아 이들의 거래 데이터를 기반으로 1차 거래업체들의 대다수가 포함된 주 업종들을 추출한다. 이어 선택된 업종별로 그 업종에 속하는 거래기업들의 대표품목들을 빈도 분석하여 가장 많이 출현한 품목만을 남기고 이를 완제품의 하위에 속하는 품목이라 정의한다. 이 논문에서는 스마트기기 산업을 사례로 위와 같은 분석을 진행하였고 분석 결과, 스마트기기는 PCB, LED, LCD, 배터리 등 총 21개의 하위 제품군으로 가치사슬이 구성되고 있음을 결론 내렸다[20].

이래은 등[20]의 연구와는 달리, 품목들의 관계로 표현된 가치사슬 구조를 사전에 정의해 놓고 각 품목별 대표 기업들을 도출한 이호신 등[19]의 연구가 있다. 연구 결과는 가치사슬을 구성하는 품목별 핵심기업과 유사/잠재 기업들의 집합이다. 먼저 핵심기업을 도출하기 위해 가치사슬을 구성하는 각 품목별 대표 키워드를 선정하고 이 키워드를 기업의 생산품목 데이터베이스(DataBase, 이하 DB로 칭함)로부터 검색하여 핵심 키워드 기업을 추출한다. 이어 전문가로부터 각 품목별 대표 생산기업을 선정받아 앞서 추출된 핵심 키워드 기업과 종합하여, 핵심기업들을 도출한다. 또한 대표 품목 키워드 외에 해당 품목을 지칭할 수 있는 다른 표현들을 모아 기업의 생산품목 DB와 상표권 DB로부터 생산 업체를 검색하고 이 기업들의 표준산업분류코드와 상표권분류 코드가 핵심기업들의 코드와 일치하는지 여부를 따져 유사/잠재 기업들을 도출한다. 이 연구에서는 이상의 과정을 통해 총 30개의 완제품에 대한 가치사슬을 도출하였다[19].

위의 두 연구에서 사용한 데이터는 한국기업데이터가 보유하고 있는 기업정보, 기업 간 거래정보를 바탕으로 만들어졌는데 해당 DB는 일정 규모 이상인 기업들의 데이터만 보유하고 있고 실시간 축적되는 거래 데이터가 아니라는 점에서 본 연구에서 다루는 전자세금계산서 데이터와는 차별점을 갖는다. 전자세금계산서 데이터를 활용해 가치사슬 분류 문제를 다룬 사례로는 이재후 등[15]의 연구가 있다. 이 연구는 기업 간 거래 연결망을 네트워크 지표들을 활용해 분석하고 특정 산업의 관여하는 주요 하청업체들에 대한 특징을 서술하였다. 주요 하청 업체들을 도출하기 위해, 거래 연결망으로부터 총 거래액을 기준으로 일정 규모 이하의 하청 업체들을 제외시키는 방법을 사용하였다[15].

본 논문은 특정 기업의 실시간 거래 데이터를 토대로 이 기업의 가치사슬에 속하는 기업들을 도출해내는 것이다. 그런데 이래은 등[20]의 연구는 특정 산업 내 가치사슬을 품목 간 관계로 구조화한 것이고, 이호신 등은 이러한 품목 간 가치사슬하에서 각 품목에 관련된 기업을 도출하는 것이다. 이호신 등[19]는 전문가

의 의견을 반영해 가치사슬이 분류된다는 점에서 자동화된 알고리즘 측면으로는 미흡하다는 한계를 지닌다. 한편 이재후 등[15]은 특정 기업의 거래연결망을 토대로 관련 하청업체를 직접적으로 도출하는데, 단순히 총 거래금액이나 거래횟수만을 사용하여, 가치사슬에 속하는 기업들만을 도출하는 데에는 한계가 존재한다.

2.2 랜덤포레스트(Random Forest)

2.2.1 의사결정나무

의사결정나무(Decision Tree)는 훈련 데이터를 가지고 반응변수(Response Variable)를 가장 잘 설명하는 설명변수(Explanatory Variable)로 가지(Branch)를 뻗어나가는 알고리즘이다[27]. 이 과정의 내부에는 데이터들을 동일 범주끼리 묶이도록 하는 가장 좋은 설명변수가 무엇인지 수치적으로 비교하는 과정이 존재하는데, 그 수치적 표현들 중 대표적인 값들이 지니지수(Gini Index)와 엔트로피 지수(Entropy Index)이다. 두 지수는 특정 설명변수로 인해 분리된 자식노드들의 불순도를 측정하는데 사용되며, 불순도가 낮을수록 두 지수 값이 낮아지고 이는 동일 범주끼리 한 자식노드에 잘 묶여있다는 것을 의미한다. 때문에 두 지수 값이 낮을수록 데이터를 잘 분류하는 변수가 되고 모든 설명변수들에 이 값들을 도출 시킨 후 가장 낮은 값을 갖는 설명변수를 선택하는 것이 의사결정나무의 내부에서 일어나는 가지 형성 과정이다.

이와 같은 의사결정나무는 한 분리지점에서 하나의 변수만을 가지고 가지를 형성하고 한번 가지를 나누는 기준으로 선택된 변수는 모형에서 제거되지 않는다. 이러한 의사결정나무의 장점은 모형 설명력이 높다는 것이고 설명변수가 연속형 또는 범주형 모두 문제없이 쓸 수 있다는 점이다. 또한 직관적인 결과를 제공함으로써 분석자가 결과를 이해하는데 큰 어려움이 없다. 그러나 의사결정나무는 연속형 데이터를 처리하는 능력이 신경망이나 전통적인 통계기법에 비해 떨어지고 예측력이 감소하는 단점이 있다[10]. 또한 분기 시 한 변수만을 사용하는 알고리즘의 특성 때문에, 비슷한 설명력을 가진 다른 변수들은 모델 구축 시 고려되지 못하고 오로지 학습 데이터만을 잘 분류하는 분기 변수들로 모형이 구축되게 되고 이는 학습 데이터를 따라 모형 구성이 크게 상이해지는 모형의 불안정성 문제를 야기하게 된다[23].

2.2.2 랜덤포레스트

랜덤포레스트(Random Forest)는 의사결정나무를 여러 개 구축하여 예측 값을 내는 기계학습 기법이다[3]. 랜덤포레스트를 이루고 있는 각 의사결정나무는 전체 학습 데이터 셋으로부터 일부 무작위 복원 추출된 학습 데이터와 설명변수에 의해 구축된다. 각 의사결정나무마다 학습한 데이터 셋이 다르기 때문에 구축된 나무의 모형과 그로부터 예측된 값이 모두 다르게 된다. 랜덤포레스트에서 하나의 의사결정나무의 학습 데이터가 매번 다르게 형성되듯 학습되지 않은 데이터 역시 매번 다르게 추출된다. 의사결정나무 구축에 사용되지 않은 데이터는 모델의 검증용으로 사용되는데, 이를 OBB (Out-Of-Bag)라 한다. 랜덤포레스트의 전체 의사결정나무에서 OOB로 선택된 횟수는 개별 개체마다 다르고 선택되었을 때 분류되는 값도 나무마다 다르게 예측된다. 이러한 특성은 개별 개체에 대한 예측 확률을 산출할 수

있게 해주는데, 예를 들어 이진 분류 문제 중 Y = 1로 예측할 확률은 (OOB로 사용됐을 때 1로 예측한 횟수)/(OOB로 사용된 횟수)로 계산된다. 이렇게 모든 개체는 확률로 최종 예측 값을 가지게 되고 임계값(Cut-off-Value)에 따라 Y = 1 또는 Y = 0으로 분류된다. 본 연구에서는 이러한 확률 값을 모기업의 특정 거래처 기업이, 모기업의 가치사슬 기업일 가능성으로 사용하였다. 이 모델의 장점은 첫째, 나무 하나의 정확도는 학습된 데이터 셋이 적고 온전하지 않기 때문에 떨어질 수 있으나, 이들을 종합하여 예측한 최종 정확도는 단순 의사결정나무 알고리즘 보다 우수하다[9]. 둘째, 대수의 법칙에 의해 숲의 크기(나무 수)가 커질수록 일반화 오류, 흔히 알려진 이름으로는 오분류율이 특정 한계 값으로 수렴하게 되고 과적합(over-fit)현상을 보이지 않는 안정적인 모델로 구축된다[9]. 셋째, 개별 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음이나 이상치로부터 크게 영향을 받지 않는다[9].

랜덤포레스트는 빈도가 불균형한 이항분류의 예측에 있어 가장 우수한 예측력을 보인 것으로 보고되고 있다[4]. 실제 전자세금계산서 데이터를 보면 한 기업의 매입/매출에 출현한 거래처 중 가치사슬 내 속하는 거래처들은 일부이고, 일시적인 거래를 하는 기업이거나 소모품 등 쉽게 대체 가능한 업종의 기업들이 대부분 차지하고 있다. 이러한 불균형 현상을 고려했을 때 이와 같은 환경에서 우수한 예측력을 보이는 랜덤포레스트는 가치사슬 기업의 분류에 효과적으로 적용될 수 있을 것으로 보인다.

랜덤포레스트는 분류와 회귀 문제 모두에 적용될 수 있으나, 범주형 예측 값을 다루는 분류문제에 주로 활용되고 있다[21]. 랜덤포레스트를 분류 문제에 활용한 국내 연구로는 39개의 재무제표 변수를 사용하여 기업의 채권 등급 분류를 진행한 연구가 존재하는데, 이 연구는 인공신경망, 서포트벡터머신, 다변량판별분석 그리고 랜덤포레스트에 모두 적용하여 모델별 성능을 비교하였고 그중 랜덤포레스트에서 가장 우수한 성능을 보였다[14]. 해외 연구사례에도 인공신경망, 서포트벡터머신 알고리즘에 비해 랜덤포레스트가 우수한 분류 성능을 보인 사례가 있는데, 이 연구의 경우 전자 혀(Electronic Tongue) 데이터를 통해 오렌지 음료 및 중국 식초를 구별해내는 연구였다[24]. 랜덤포레스트가 항상 다른 알고리즘들에 비해 더 우수한 분류 성능을 보이는 것은 아니지만[26, 35] 인공신경망 또는 서포트벡터머신과 비교할 때 명확한 이점은 존재한다. 상대적으로 적은 파라미터만을 사용자가 선택하면 된다는 점에서 서포트벡터머신, 인공신경망보다 학습의 편의성이 높다고 할 수 있다[28]. 따라서 성능이 우수하면서 학습이 간편한 랜덤포레스트를 본 연구의 분석 알고리즘으로 적용하고자 한다.

3. 연구모형

이 장에서는 가치사슬의 개념과 이에 따른 가치사슬 분류 문제를 기계학습 문제로 모형화하고, 모형에 사용될 데이터의 정형화 과정을 서술한다.

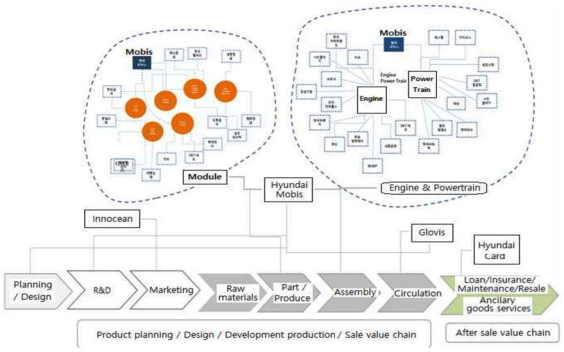
3.1 가치사슬 정의

한 기업의 가치사슬에 속하는 기업을 분류하

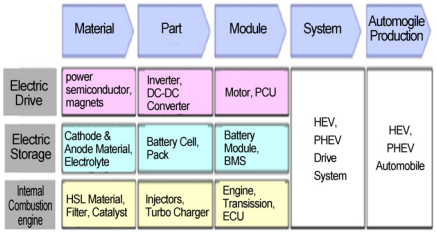
고 나아가 전체적인 가치사슬 기업을 분류하기 위해서는 가치사슬에 대한 명확한 정의가 요구된다. 마이클 포터에 따르면 가치사슬이란, 한 기업의 부가가치 창출과 관련된 일련의 활동, 기능, 프로세스의 연계이다[32]. 이를 토대로 가치사슬에 대한 다양한 학술적 정의들이 제시되어 왔으며[5, 29, 33], 이를 요약하면 크게 다음과 같다.

첫째 기업중심 가치사슬로, 마이클 포터가 정의한 가치사슬 분석 모형에 따라 기업의 주활동과 지원활동에 해당하는 가치 활동들을 분석하고 각 활동에 속하는 기업을 도출해내는 것이다[5, 11]. 분석 결과는 가치 활동과 각 가치활동에 속하는 기업들이 되며 이를 통해 자사의 경쟁적 위치를 파악하고 향후 경영전략을 세우는 데에 사용된다[22]. 둘째 품목중심 가치사슬로, 완제품을 생산하는데 사용되는 하위부품 및 서비스들의 연계흐름을 분석한 것이다. 주로 특정 제품 및 서비스의 전반적인 생산 흐름이 확인되지 않은 신사업이나[29, 34] 제품의 특성 및 사용 환경이 달라지면서 변화되는 가치사슬의 구조를 비교하기 위해 사용된다[8, 18, 38]. 셋째 공정중심 가치사슬로, 제품 및 서비스가 완성되기까지의 과정을 공정단위로 분해하고 구조화한 것이다[13]. 공정중심 가치사슬의 분석 결과로 생산 공정들이 도출되고 이는 공정별 생산소요시간, 불량 발생률 등 측정 가능한 성과 지표들을 통해 비교분석하여 공정 개선에 주로 활용한다[33].

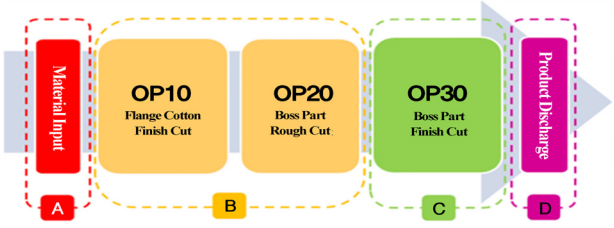
분석 중심별 분류 사례 예시는 <Figure 1>에 제시되어 있다. <Figure 1-(a)>는 현대자동차의 가치 활동별 가치사슬 기업을 도출한 결



(a) Enterprise-Centric: Hyundai Motor[5]



(b) Product-Centric: Low Carbon Car[37]



(c) Process-Centric: Automotive Hub Assembly [7]

〈Figure 1〉 Change of Value Chain Analysis Result by Subjects: Automobile

과로 분석 대상이 기업중심일 때의 결과 예시이고 <Figure 1-(b)>는 완성차를 구성하는 원자재 및 부품을 구조화 한 것으로 품목중심 가치사슬의 예시가 된다. 마지막으로 <Figure 1-(c)>는 자동차 엑셀에 장착되는 허브 어셈블리(Hub Assembly)의 부가가치를 창출시키는 공정들을 분석하여 구조화한 것으로 공정중심 가치사슬의 분석의 예시이다. <Figure 1>의 분석 대상이 되는 산업은 모두 자동차로 동일하지만 분석 중심을 어떻게 정의하느냐에 따라 결과가 상이해짐을 확인할 수 있다.

그런데, 이러한 정의들은 모기업과 하청기업 간의 거래형태분석이나 산업생태계 분석 혹은 공정의 효율성 분석을 위해 제시된 정의들이다. 본 연구는 수 만 개 기업들의 가치사슬에 속하는 기업들을 기업 간 거래 데이터를 기반으로 효율적으로 도출하는 것이므로, 위의 정의들에 비해 보다 구체적인 정의가 필요하다.

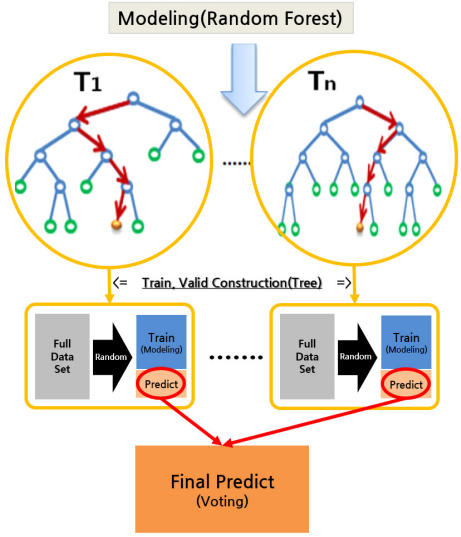
특정 기업의 가치사슬에 속하는 기업은 결국 모기업이 생산하는 특정 품목의 부가가치에 기여하는 기업이다. 따라서 구체적인 가치사슬 정의를 위해서는 우선적으로 기업과 품목이 정해져야 한다. 다음으로는 가치사슬의 방향을 정해야 한다. 예를 들어 브레이크를 생산하는 기업은 매출처인 제동장치를 생성하는 기업을 상향 가치사슬 기업으로 두며, 브레이크 생산에 들어가는 부품인 패드를 생산하는 기업을 하향 가치사슬 기업으로 둘 것이다. 마지막으로 가치사슬은 하나의 연결고리이므로, 특정기업 입장에서 하향 가치사슬 기업을 분류할 때, 연결 고리의 바로 다음 단계에 속하는 기업은 1차 가치사슬 기업이고, 그 다음 단계에 속하는 기업은 2차 가치사슬 기업이 될 것이다. 이상의 논의를 종합하면, ① 기업 - 품목, ② 방향 그리고 ③ 차수, 3가지 요소를 고려하여 가치사슬이 정의되어야 한다. 본 논문에서는 (기업-품목, 후방향, 1차)에 대한 가치사슬기업 분류 모형을 제시하고자 한다.

3.2 가치사슬 분류 모형

기계학습은 알고리즘을 통해 주어진 데이터를 학습하고 예측 값을 산출하는 분석 과정을 의미한다. 여기서 예측변수가 범주형이면 분류문제, 연속형이면 회귀문제로 분류되고 또한 데이터에서 예측 값을 산출하는데 사용되는 변수를 설명변수라 한다. 본 논문에서 다루는 문제는 특정기업의 실시간 거래 데이터를 토대로 그 기업의 매입처 중 가치사슬에 속하는지 여부에 따라 매입처 기업들을 분류하는 문제이므로, 이를 기계학습 문제로 표현하면, 설명변수는 거래 데이터를 토대로 형성되는 변수들이며, 예측변수는 특정기업의 매입처에 해당하는 기업이 가치사슬 기업인지의 여부가 될 것이다. 주어진 설명 변수를 토대로 기업들을 분류하는 작업을 수행하는 기계 학습 알고리즘 중 가장 널리 활용되는 것은 의사결정나무, 신경망기법, 그리고 서포트 벡터 머신 등이다[16]. 그런데 NN과 SVM은 파라미터의 추정 작업에 있어 전문적인 지식을 요구하며 또한 학습 데이터가 상당히 클 때 성능이 뛰어나다[39]. 이에 반해 의사결정 나무는 추정해야 할 파라미터가 없으며, 사용이 매우 간단하다. 다만 나무를 형성해 가는 과정에서 한 분기시점에 하나의 변수만을 선택하여 분할하기 때문에 비슷한 설명력을 가진 나머지 변수들이 고려되지 못하며, 또한 변수선택 기준에 따라 분기시 선택되는 변수가 달라지며, 이렇게 해서 형성된 나무에 따라 최

종 성능이 달라지는 불안정성이 단점으로 지적되고 있다[23]. 또한 상대적으로 NN이나 SVM보다 성능이 떨어지는 단점도 갖고 있다[10]. 이러한 단점을 보완하기 위해 랜덤포레스트모형이 제시되었고[3, 9], 따라서 본 논문은 사용의 편이성과 높은 성능을 보이는 랜덤포레스트를 분류모형으로 활용하고자 한다[4, 14, 28].

랜덤포레스트는 의사결정나무를 여러 개 형성하여 결과 값을 예측하는 앙상블 알고리즘이다. 랜덤포레스트를 이루고 있는 하나의 나무는 모델이 형성될 때 전체 학습 데이터로부터 설명변수와 개체를 무작위 복원 추출한 일부 데이터 셋만을 가지고 학습을 수행해나간다. 이런 개별 의사결정나무는 저마다 다른 결과 값을 가지게 되는데 랜덤포레스트는 각 나무마다 다르게 예측된 값을 모두 모아 평균을 내거나 또는 투표를 통해 최종 예측 값을 산출한다[36]. 이 과정을 도식화 하면 <Figure 2>과 같다.



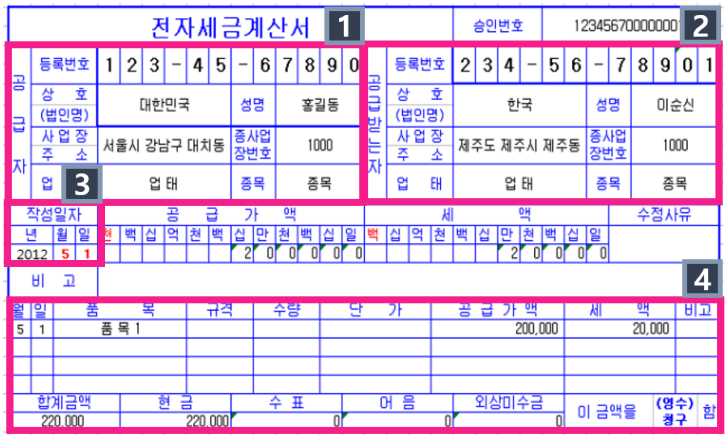
〈Figure 2〉 Modeling Process of Random Forest

랜덤포레스트에서 결정해야 될 파라미터는 두 가지가 있는데, “구성할 나무의 수”와 “나무를 만들 때 쓸 설명변수의 수”다[21]. 나무의 수는 모델이 과적합을 피하고 정확도를 높일 수 있게 적절한 값을 선택해야 하고, 나무마다 사용할 설명변수의 수의 경우 보통 분류문제에서는 전체 입력변수 수의 제곱근 값을 이용한다[17]. 본 연구에서는 하나의 랜덤포레스트가 형성하는 나무의 수를 50만 개로, 나무마다 선택되는 설명변수의 수는 전체 설명변수 19개의 제곱근에 가까운 약 4개로 지정하였다. 사용된 19개의 설명변수에 대해서는 다음절에 소개된다.

3.3 설명변수 구축

전자세금계산서란 전자적 방법으로 발급하는 세금계산서를 말한다. <Figure 3>에서 보이는 자료의 빈칸들이 전자세금계산서를 발행할 때 기입해야 될 값들이다. 그 중 본 연구에 활용한 정보는 ① 공급자, ② 공급받는자, ③ 거래일자, ④ 거래 품목 및 금액에 해당하는 정보들이다.

먼저, 공급자 정보에는 사업자등록번호, 상호, 대표자, 사업장 주소, 업태, 업종 정보가 있고 공급받는자 정보도 동일한 속성들로 이뤄져 있다. 여기서 사업자등록번호는 기업마다 유일하기 때문에 분류 값으로 사용된다. 거래일자 정보에는 전자세금계산서를 작성한 일자가 제공되고 있고, 품목명, 규격, 수량에 해당하는 품목 정보와 단가, 공급가액, 세액, 총 합계금액에 해당하는 금액 정보가 합해져 거래된 품목과 금액에 대한 정보로 제공되고 있다. 그 외에도 발행방향(역/정발행), 청구유형(영수/청구), 세금유형(과세/면세/영세) 등 기타 거래정보들이 전자세금계산서에 존재하나 본 연구에서는 제외하였다.



〈Figure 3〉 Electronic Tax Invoice Form

전자세금계산서는 기업 간 거래가 발생할 때마다 데이터가 생성되는 형태로, 거래처에 따라 여러 건의 세금계산서 데이터가 존재할 수 있다. 본 연구에서는 이러한 거래 데이터로부터 가치사슬 기업들을 분류하는데 사용할 거래 속성들을 추출하였고 이를 알고리즘의 설명변수로 사용하였다. 각 설명변수들의 개념과 계산과정에 대한 본격적인 설명에 앞서, 거래 속성별 도출과정에 사용되는 수식 기호들을 <Table1>에 정의한다.

〈Table 1〉 Definition and Notations

기호	Explanations	Remarks
A	Analysis Target Company	Company ID
j	Supplier of A	Company ID
m	Number of All Months in Analysis Period	Units: Number
p	Purchase Product of A	Name of Items
$t(j)$	Number of Transactions Between Aandj supplier	Units: Number
$t_p(j)$	Number of Transactions with p item Between Aandj supplier	Units: Number
$tm(j)$	Number of Transactions Between Aandj supplier Monthly Sum	Units: Number, $tm \leq m$
$D_i(j)$	The Date that ith Transaction Between Aandj supplier	Date
$S_i(j)$	The Amount that ith Transaction Between Aandj supplier	Units: ₩
$S_{tm(j)}(j)$	Transaction monthly Between Aandj supplier Transaction amount	Units: ₩
$ST_i(j)$	The amount traded in j for the ith analysis month	Units: ₩
$V_{SA(A)}$	All of A's monthly sales vecror, $\{ST_1(A), ST_2(A), \ldots, ST_m(A)\}$	Units: ₩
$V_{S(p)}$	Monthly Purchasing Amount vector with p item, $\{S_1(p), S_2(p), \ldots, S_m(p)\}$	Units: Number, Total m
E_i	Gap of Between $V_{SA(A)}$ and $V_{S(p)}$ Measured on the same month, $\{ST_1(A)-S_1(p), ST_2(A)-S_2(p), \ldots, ST_m(A)-S_m(p)\}$	Units: Number
$La(A), La(j)$	A latitude of business address, j latitude of business address	Location
$Lo(A), Lo(j)$	A longitude of business address, j longitude of business address	Location

가치사슬 기업은 모기업의 제품에 핵심적인 기여를 하는 기업이므로, 모기업과 지속적이고 규칙적인 거래를 하게 될 가능성이 크다. 따라서 이러한 거래 특성은 기본적으로 거래금액이나, 거래기간, 거래횟수 그리고 거래주기 등과 같은 변수들로 설명 될 수 있을 것이다. 우선 거래와 관련되어 생성한 설명변수들을 살펴보면 다음과 같다.

3.3.1 거래빈도(Transaction Frequency)

거래빈도는 두 기업의 관계 형성을 촉진시킨다[2]. 분석기간 동안 많은 거래횟수를 가지고 거래한 기업은 그렇지 않은 기업들보다 자사의 가치사슬 기업일 가능성이 높다. 해당 속성은 분석기간 내 거래 일수 측정하여 계산되며 식 (1)과 같이 정의된다.

$$j\text{거래처의 거래빈도} = t(j) \quad (1)$$

3.3.2 거래금액(Transaction Amount)

거래횟수와 마찬가지로 두 기업 간 거래금액은 자사에 협력적이고 중요 관계에 놓인 거래처를 분류하는데 영향력 있는 속성일 것이다. 설비나 기계 업종의 경우에는 거래횟수는 많지 않으나, 1년에 한 번 정도, 거래규모가 커서 거래금액은 높게 계산된다. 그렇게 되면 거래 횟수는 적지만 금액은 평균 이상의 위치에 속할 수도 있게 된다. 계산과정은 식 (2)의 같이 정의된다.

$$j\text{거래처의 거래금액} = \sum_{i=1}^{t(j)} S_i(j) \quad (2)$$

3.3.3 거래주기(Transaction Cycle)

거래 주기는 거래처별 속한 업종의 특성에 따라 다른 값을 가질 수 있다. 예를 들어, 원자재 및 부품을 납품하는 기업들의 경우 비교적 짧은 주기로 빈번히 거래될 것이고 설비 및 기계 제조 기업들의 경우 전자의 기업들보다는 자주 거래 관계가 발생하지 않아 보다 긴 거래 주기를 가질 것이다. 거래주기는 거래 일자별 차이를 구하고 차이들의 평균을 통해 계산된다. 식 (3)에 이러한 과정이 정의되어 있다.

$$j\text{거래처의 평균거래주기} = \frac{1}{t(j)-1}\sum_{i=1}^{t(j)-1} D_{i+1}(j) - D_i(j) \quad (3)$$

3.3.4 최근 거래월별 거래금액(Amount per Latest Transaction Month: ALTM)

거래가 발생한 최근 열두 달 치의 월별 거래금액을 변수로 사용하였다. 총 12개의 변수가 거래처별로 생성되는데, 거래처별로 거래한 월이 각각 다르기 때문에 “가장 최근 거래 달의 거래금액”에서 “가장 최근 거래 달”은 거래처별로 다를 수 있다. 예를 들어, A매입처는 2016년 5월, 8월, 10월에 거래를 하였고 B매입처는 12월에만 거래를 했다면, 가장 최근 거래 달의 거래금액에는 A매입처의 경우 10월 거래금액이, B매입처의 경우 12월 거래금액이 부여되게 된다. 만일 분석기업과 한 번 거래가 일어난 기업의 경우 총 12개의 변수 중 첫 번째 변수에만 거래금액이 산정되고 나머지 11개의 변수에는 전부 0으로 값이 채워진다. 이 속성은 시계열적인 요인을 고려함과 동시에 단기적인 거래를 하는 거래처들을 거를 수 있는 요인으로 작용될 것이다.

식 (4)의 계산식을 통해 해당 변수의 생성과정이 정의된다. $S_{tm(j)}(j)$은 j거래처가 거래한 가

장 최근 월($tm(j)$)의 거래금액이고, $S_{tm(j)-11}(j)$는 j거래처의 최근 거래 월($tm(j)$)로부터 열한 번째 이전의 거래 월에 해당하는 거래금액을 의미한다. 총 5개월을 거래하여 다섯 달 치 거래금액만 존재하는 거래처의 경우, $S_{tm(j)}(j) \sim S_{tm(j)-4}(j)$까지는 실거래금액이 부여되고 거래가 발생하지 않은 나머지 7개의 거래월별 거래금액, $S_{tm(j)-5}(j) \sim S_{tm(j)-11}(j)$에는 0이 부여된다.

$$j\text{거래처의 최근 12달의 월별 거래금액} = [S_{tm(j)}(j), S_{tm(j)-1}(j), \cdots, S_{tm(j)-11}(j)] \quad (4)$$

3.3.5 거래집중도(Transaction Concentration)

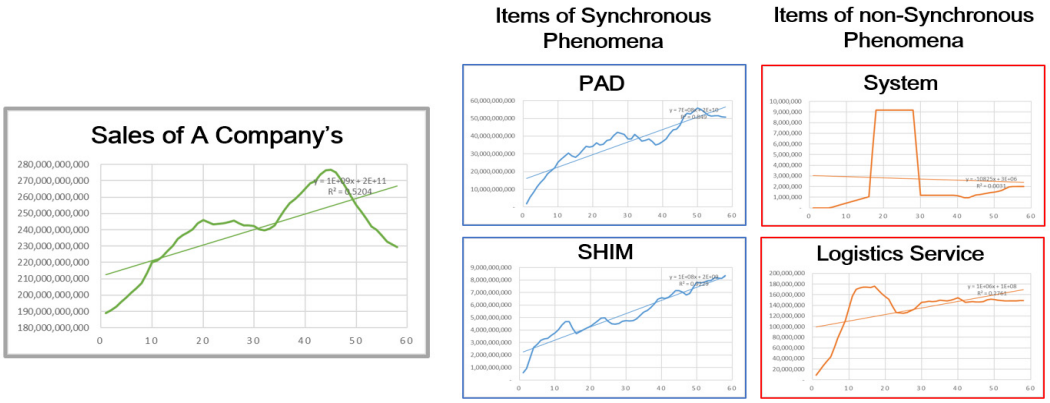
거래 집중도는 특정 거래처에 대한 거래량의 비율을 의미하는데, 집중도가 높을수록 상대방에 대한 의존도가 커지게 되며 거래의 지속성, 두 기업 간 상호작용에도 그 영향을 미치게 된다[30]. 만일 특정 매입처가 분석 기업에 거래집중도가 높은 기업이라면 그 기업은 자사의 핵심 거래처이면서, 가치사슬 기업일 가능성이 높다. 이 속성은 거래처별 전체 매출액 대비 분석 기업에게 올린 매출액의 비율로 계산되며 식 (5)로 정의된다.

$$j\text{거래처의 거래집중도} = \sum_{i=1}^{t(j)} S_i(j) \,/\, \sum_{i=1}^{m} ST_i(j) \quad (5)$$

기본적인 거래 특성 외에도 가치사슬 기업임을 분류하기 위한 기타 특성을 새롭게 개발하여 추가하였는데, 이에 대한 설명은 다음과 같다.

3.3.6 품목 동조화 지수(Product Concordance Index: PCI)

특정 기업의 매출품에 사용되는 원재료 및 부품의 경우에는 이 기업의 매출 흐름에 따라 그 매입액이 달라 질 수 있다. 자사의 제품 수요가 증가하면 이를 만드는데 사용되는 재료들에 대한 매입도 따라서 증가되기 때문이다. 전자세금계산서를 통해 이와 같은 상황이 실제로 일어나고 있는지 살펴보았다. <Figure 4>를 보면, 브레이크 패드를 생산하는 A업체의 매출 흐름이 브레이크 패드 생산을 위해 부품으로 사용되는 패드, SHIM과 같은 품목의 매입흐름과 유사한 것으로 볼 수 있다. 그러나 시스템,



〈Figure 4〉 Examples of Product Concordance and Non-Concordance

물류대행과 같이 원자재 및 부품으로 사용되지 않는 품목의 매입흐름은 상이한 것으로 보였다. 이에 따라, 본 연구에서는 품목 동조현상을 살펴보는 것이 원재료 및 부품을 납품하는 매입처를 찾는데 중요할 것이라 판단하였고, 따라서 품목 동조화 지수(PCI)를 측정할 수 있는 계산식을 세워 값을 도출하였다.

동조 현상을 수치적으로 측정하기 위해 코사인 유사도(Cosine Similarity)와 평균절대치오차(Mean Average Error)를 사용하였다. 품목별로 두 지표 값을 각각 구하고 합산하여 품목별 PCI를 계산하였다. 정의한 계산식은 식 (6)과 같다.

$$PCI_p = Cosine\ Similarity_p + Mean\ Average\ Error_p \quad (6)$$

$$Cosine\ Similarit_p: \frac{V_{ST(A)} \cdot V_{S(p)}}{\|V_{ST(A)}\|\|V_{S(p)}\|},$$

$$Mean\ Average\ Error_p: \frac{\sum_{i=1}^{m}|E_i|}{m}$$

$V_{ST(A)}$은 A의 월별 매출액 벡터로, $\{ST_1(A), ST_2(A), \ldots, ST_m(A)\}$와 같이 표현되며 총 m개의 벡터를 갖는다. $V_{S(p)}$는 p품목으로 매입한 월별 매입액 벡터로 $V_{ST(A)}$와 동일하게 m개의 값을 갖고 $\{S_1(p), S_2(p), \cdots, S_m(p)\}$로 표현된다. E_i은 i번째 월에 해당하는 $V_{ST_i(A)}$, $V_{S_i(p)}$의 차이를 의미하며, $\{ST_1(A)-S_1(p), ST_2(A)-S_2(p), \cdots, ST_m(A)-S_m(p)\}$로 정의된다.

식 (6)을 통해 도출된 품목별 PCI값을 가지고, 거래처별 동조 지수를 계산해야하는데, 이는 거래처가 납품한 거래 품목의 동조화 지수(PCI_p)에 해당 품목으로 납품한 횟수(t_p)를 가중 합산하여 계산된다. 계산 과정은 식 (7)에 정의되어 있다.

$$j\text{거래처의 동조화지수} = \sum_{i=1}^{p} t_i(j) \times PCI_i \quad (7)$$

두 기업 간 거래관계 특성이 아닌, 거래처 기업의 속성을 바탕으로 변수를 생성하였고 이는 다음과 같다.

3.3.7 거래처별 총 매출액(Supplier's Total Sales)

이 속성은 동일 사업군내 기업들이 다른 업종 내 기업들보다 비슷한 매출규모를 가지고 있음을 가정하여 만들어진 변수로 매입처별 분석기간 내 총 매출액의 합이다. 동일 원자재를 납품하는 거래처들은 품목 단가가 유사하기 때문에 총 매출 규모도 일정 금액 이상으로 유사할 것이라 판단하였다. 해당 속성은 식 (8)으로 정의된다.

$$j\text{거래처의 총 매출액} = \sum_{i=1}^{m} ST_i(j) \quad (8)$$

3.3.8 GPS 거리차(Gap of GPS)

분석 대상 기업이 거래처를 선택하는데 고려해야 될 요인이 여러 있을 수 있는데 그 중 근접한 지역에 거래처가 위치했는가도 선택의 한 기준으로 적용될 수 있다. 실제로 동일 업종 내 속하는 기업들은 근교에 뭉쳐져 있는 경우가 많고 이를 산업클러스터라 일반화시켜 부르기도 한다[38]. 특히, 부품 조달의 신속성과 비용문제를 해결하기 위해서는 거리적으로 가까운 업체가 먼 업체보다 주 거래처로써 선택될 가능성이 클 것이다. 따라서 본 연구에서는 두 업체

간 거리 차 변수를 생성하였고 이는 전자세금계산서 데이터에 기입된 공급자와 공급받는자의 주소를 위도와 경도로 변환하여 두 좌표의 차이를 통해 계산된다. 하버사인 공식(Haversine Formula)을 사용해 거리를 측정하였으며 측정 과정은 식 (9)로 정의된다.

$$j\text{거래처와의 거리차} = (6317000 \times w)/1000(km) \quad (9)$$
$$w = 2 \times atan_2(\sqrt{k}, \sqrt{(1-k)})$$
$$k = Sin(rad(La(A) - La(j))/2)^2 + Cos(rad(La(A))) \times Cos(rad(La(j))) \times Sin(rad(Lo(A) - Lo(j))/2)^2$$

전자세금계산서 데이터를 가지고 거래처 별 거래 정보를 위와 같은 속성들로 구축하였다. 이렇게 생성한 설명변수는 거래월별 거래금액의 12개 변수와 나머지 7개 변수를 종합하여 총 19개로 구성되고 개체 수는 분석 기업별 매입 거래관계에 나타난 거래처 수에 따라 형성된다.

4. 사례분석

이번 장에서는 제3장에서 구축한 연구모형을 자동차 부품 산업에 적용한 과정 및 결과를 서술한다. 대상 기업은 자동차 부품 산업 중 브레이크 패드 생산 업체이다. 브레이크 패드 시장은 상위 세 개 기업의 점유율이 95%에 이르는 과점시장이고 그 중 가장 시장 점유율이 높은 A사를 선택하였다. 이 기업의 매입 거래 데이터를 활용하여 학습 데이터 셋을 만들었고, 사용된 데이터의 수집 기간은 2010년 01월부터 2015년 10월까지이다.

4.1 분석 데이터

데이터 수집 기간 동안 A사와 매입 거래관계가 하나라도 있었던 거래처의 수는 총 178곳이었고 이 중 전문가에 의해 가치사슬 관련 기업이라 판명된 기업의 수는 28개였다. 약 15% 정도만이 가치사슬 기업에 해당하는 불균형 데이터이다. 분석기업 A사와 178곳의 매입처간 거래된 총 6,447건의 전자세금계산서 데이터를 토대로, 식 (1)~식 (9)를 활용하여 생성된 총 19가지의 설명변수로 구성된 분석데이터가 <Figure 5>에 제시되어 있다.

4.2 가치사슬 분류 알고리즘

가치사슬 분류 알고리즘에 사용되는 랜덤포

ID	Sup's Total Sales	Amount	concentration	Frequency	Cycle	Latest 1st amount	Latest 2nd amount	...	Latest 12th amount	Gap of GPS	PCI	Target
1	13,591,214,355	12,416,800	0.001	23	93	187,000	336,600	...	523,600	89	42	0
2	1,580,780,000,000	1,170,940	0.000	57	19	6,630	19,900		19,900	74	244	0
3	2,210,780,000,000	385,000	0.000	1	0	385,000	-		-	96	7	0
4	1,815,408,204	1,056,000	0.014	1	0	1,056,000	-		-	93	-4	0
5	61,756,363,692	183,947,500	0.018	7	30	167,200,000	2,791,250		-	93	26	0
6	202,243,000,000	18,617,289	0.001	2	65	5,660,466	12,956,823		-	96	-1	0
7	109,947,000,000	4,653,139,790	0.042	68	31	41,825,190	24,201,430		29,943,760	71	1010	1
8	130,213,000,000	26,997,505	0.000	55	37	183,744	934,164		344,256	89	1340	0
9	1,235,120,000,000	968,000	0.000	2	90	440,000	528,000		-	89	21	1
10	40,950,029,779	67,000,000	0.006	2	363	33,000,000	34,000,000		-	82	-5	0

〈Figure 5〉 Example of Analysis Data

레스트를 구성하는 나무의 수는 50만 개이고, 나무마다 가질 수 있는 입력변수의 수는 4개이다. 각 의사결정나무에서는 모델 구축에 사용된 학습용 데이터와 그렇지 않은 검증용 데이터가 70% 대 30%의 비율로 분할된다. 본 연구에서는 OOB의 예측 값을 통해 각 거래처별 가치사슬 확률을 계산하였고 임계값 설정을 통해 가치사슬 기업과 아닌 기업으로 분류했다.

4.3 알고리즘 성능 평가

분류 모델의 성능을 평가하기 위해 정확도(Accuracy), 오분류율(Misclassification Rate), F1-척도(F1-Score) 그리고 AUC(Area Under the Curve)를 사용하였다. 정확도는 실제 1인 것을 1로, 실제 0인 것을 0으로 예측한 비율이고 오분류율은 실제 1인 것을 0으로, 실제 0인 것을 1로 잘못 예측한 비율이다. F1-척도는 재현율(Recall)과 정밀도(Precision)의 조화평균으로 계산되는데 여기서 재현율은 실제 1인 것 중 예측도 1로 한 비율이고 정밀도는 예측된 1중 실제 1인 것의 비율을 의미한다. 마지막으로 AUC는 ROC(Receiver Operation Characteristic)곡선 그래프의 면적 값으로 그래프의 Y축은 재현율이고 X축은 거짓 긍정률(Fall-Out)인데 이는 실제 0인 것을 1로 잘못 예측한 비율이다. 1이라고 예측한 확률 값이 높은 순으로 개체를 나열하고 그룹을 나눠 그룹별 X축과 Y축 값을 구해 그래프를 그리는데 상위 그룹에 실제 1이 많이 포함되어 재현율의 값이 커지면 AUC 면적도 커지게 된다. 이상의 지표들은 모델의 성능 평가를 위해 빈번히 사용되며 <Figure 6>에 등식이 표현되어 있다.

		Predict	
		$\hat{Y}=1$	$\hat{Y}=0$
Real	$Y=1$	***TP*** Number of Objects $y=1\ \&\ \hat{y}=1$	***FN*** Number of Objects $y=1\ \&\ \hat{y}=0$
	$Y=0$	***FP*** Number of Objects $y=0\ \&\ \hat{y}=1$	***TN*** Number of Objects $y=0\ \&\ \hat{y}=0$

(a) Confusion Matrix

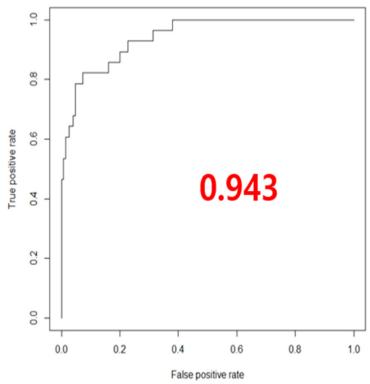
Performance Index	Equation
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$
Misclassification Rate	$\frac{FP+FN}{TP+FN+FP+TN}$
Recall	$Recall=\frac{TP}{TP+FN}$
Precision	$Precison=\frac{TP}{TP+FP}$
F1-Score	$2\cdot\frac{precision\cdot recall}{precision+recall}$
Fall-Out	$\frac{FP}{TN+FP}$

(b) Equation by Measurement

〈Figure 6〉 The Equation of the Performance Evaluation Index

분류 성능 지표를 계산하기 위해서는 예측치가 확률이 아닌 이항분류 값으로 도출되어야 하고 이를 위해서는 특정 값 이상은 Y=1로 분류할 수 있는 임계값이 정해져야한다. 본 연구에서는 임계값을 F1-척도가 가장 높을 때 사용된 임계값으로 선택하였고, 그 값은 약 30%였다. 임계값과는 무관히 성능이 계산될 수 있는 AUC를 제외하고 나머지 지표에 해당 임계값을 적용시켜 측정지표별 성능을 측정하였다. 그 결과는 <Figure 7>에 제시되어 있다.

오분류율을 제외하고는 모두 값이 1에 가까울수록 모델의 성능이 우수하다 평가된다. 구축된 모델은 정확도 92%, 오분류율 8%로 높은 성능을 보였다. 또한 모든 개체의 예측력을 평가하는 정확도와는 다르게, 가치사슬 기업들만 고려하여 예측력을 평가하는 F1-척도를 살펴



Cut-off : 30%	
Performance	**A Company**
Accuracy	0.92
Misclassification Rate	0.08
F1-Score	0.76
AUC	0.94

〈Figure 7〉 Classification Performance of Random Forest

보면 76%로 비교적 높은 편이었다. 즉, 28개의 가치사슬 기업을 모델이 비교적 잘 분류하고 있는 것이다.

이상의 지표들은 적절한 임계값이 존재해야만 계산될 수 있다. 즉, 임계값의 변화에 따라 성능이 달라진다는 것이다. 때문에 임계값에 영향을 받지 않는 성능 지표로 모델을 평가할 필요가 있는데, 이는 AUC를 통해 평가될 수 있다. 본 연구에서는 AUC 값이 94%로 랜덤하게 예측했을 때의 성능 50%와 비교했을 때 상당한 차이를 가짐을 확인할 수 있었다.

4.4 사례분석 결과

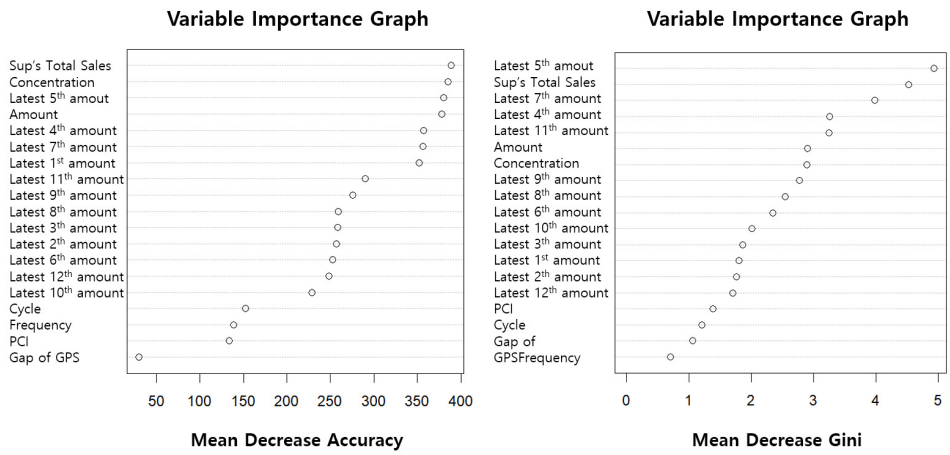
기계학습모형은 두 가지 즉, 예측 성능과 예측 결과에 대한 설명력에 의해 유용성이 평가된다. 제4.3절에서 예측 성능을 살펴보았고 이 절에서는 예측 결과에 대한 설명력을 기술하고자 한다. 모형의 설명력이란 목표변수에 영향을 미치는 설명변수들의 상대적 중요도를 파악해내는 것이다.

랜덤포레스트에서 설명변수의 중요도를 측정하기 위해서는 크게 두 가지 중요도 지표를 사용한다[1]. 먼저, MDG(Mean Decrease Gini)라는 지표로 이 값은 랜덤포레스트를 이루는 각 나무가 가지를 뻗어나갈 때마다 선택되는 변수들의 불순도 감소량을 측정하여 전체 나무로부터의 그 평균치 값을 사용한다. 때문에 특정 변수의 MDG값이 높다는 것은 그 만큼 그 변수를 가지고 개체들을 분류하게 되면 불순도를 감소시키는 방향으로 즉, 동일 범주끼리 묶이도록 하는데 일조한다는 것을 의미한다. 또한 정확도의 개념으로도 변수의 중요성을 판단할 수 있는데 이는 MDA(Mean Decrease Accuracy)라 정의된다. MDA는 구축된 나무의 정확도가 특정 변수 제거 후 재구축했을 때 감소되는 정확도의 차이를 변수별로 평균화한 것이다. 분류 정확도를 높이는데 큰 영향을 준 변수일수록 그 변수를 제거했을 때의 정확도 감소량은 커지게 된다. 따라서 랜덤포레스트의 변수 중요도를 측정하는 두 지표는 모두 값이 커질수록 변수의 중요도가 높아지게 된다.

앞서 설명한 두 지표에 의해 본 연구 모형의 변수들을 평가한 그래프가 <Figure 8>에 제시되어 있다. 그래프 상단에 위치할수록 변수의 중요도가 큰 것으로 거래금액, 거래집중도, 거

래처별 총 매출액이 주로 위치하였다. 거래집중도 변수와 거래금액 변수는 거래처와 분석 대상 기업, 각각의 입장에서 상대 기업에 대한 의존도를 나타내는 속성이다. 먼저, 거래 집중도가 높은 거래처란 분석 기업이 거래처의 주 고객사라는 것을 의미한다. 반대로 분석 기업 입장에서 보다 많은 금액을 매입하는 거래처의 경우, 해당 거래처는 분석 기업 입장에서 주 공급처를 의미한다. 두 변수가 가치사슬 기업을 분류하는데 중요하게 작용한 결과를 보면, 거래 관계의 상호의존성이 가치사슬 기업을 분류하는 주요 속성인 것으로 해석된다. 또한 거래처별 총 매출액도 모델에서 중요도가 높은 변수로 선택되었는데, 이는 가치사슬 기업들을 일정 매출 규모로 구분할 수 있다는 것을 의미한다.

변수 중요도 하위에 위치한 거래 속성들 중 본 연구에서 새롭게 정의한 품목 동조화 지수가 포함되고 있다. 그 원인에 대해 몇 가지 고찰해보면 먼저, 본 연구에서 제시한 동조 측정 지표인 코사인과 평균절대치오차가 충분한 동조 현상 측정 지표로써 사용되지 못했을 가능성이다. 시계열 유사도의 측정방법은 고정적 접근(Lock-Step), 탐색적 접근(Elastic), 임계값 기반 접근(Threshold-based), 패턴기반 접근(Pattern-based)으로 분류되고 각각의 접근법에 해당하는 유클리디안 거리(Euclidean Distance), 동적정합(Dynamic Time Warping), 최장 공통 부분열(Longest Common Sub Sequence) 등의 측정 지표들이 존재한다[6]. 이상의 지표들 중 품목 동조 현상을 측정할 수 있는 가장 적절한 지표를 선택하는 과정이 본 연구에서는 충분히 다뤄지지 못했다. 또한, 재고가 발생하는 상황을 고려하지 않고 두 시계열 그래프의 동일 시점에 해당하는 매출 상승, 하락으로만 동조를 측정했다는 데에 한계를 갖는다. 즉, 두 시계열 그래프에 존재할 수 있는 시차(Lag)값을 고려하지 못한 채 품목별 동조 지수를 산출했다는 것이다. 이상의 이유들로 품목 동조화 지수가 가치사슬 기업을 분류하는데 유용하게 작용하지 못했음을 파악하였고 추후 추가적인 연구 수행을 통해 해당 지수를 개선해 나가고자 한다.



〈Figure 8〉 Variable Importance Graph

5. 결 론

본 연구에서는 기존의 산업 및 기업중심의 가치사슬 분류 문제를 기업-품목중심의 가치사슬 분류 문제로 처음 접근하였으며 또한 이러한 과정을 자동적으로 수행할 수 있는 분류 모형을 제시하였다. 모형에 사용된 알고리즘은 학습의 편의성과 높은 예측력을 가진 랜덤포레스트이다. 실시간 전자세금계산서 데이터를 재구성하여 19개의 속성들로 이뤄진 학습 데이터를 구축하였으며, 여기에는 거래주기, 거래횟수, 거래금액 등 일반적인 거래 속성 변수들과 매출액, 거리 차와 같은 거래처별 특성 변수, 마지막으로 거래 품목들의 동조현상을 코사인 유사도와 MAE를 활용해 새롭게 정의한 품목 동조화지수 등이 포함된다. 이러한 속성들은 각 거래처별 설명변수로 사용되고 이를 통해 예측할 값은 해당 거래처가 가치사슬 거래처가 될 가능성이다.

사례 데이터를 사용하여 가치사슬 분류 알고리즘을 적용한 결과, 모델의 분류 성능은 정확도 92%, 오분류율 8%, F1-척도 76% 그리고 AUC 94%로 비교적 높은 성능을 보였다. 이러한 모델의 성능을 높이는데 영향을 준 설명변수들을 살펴보기 위해, 랜덤포레스트의 변수 중요도 측정 지표인 MDG와 MDA를 사용하여 비교하였고 그 결과 거래집중도, 거래금액, 거래처별 총 매출액, 최근 거래월별 거래금액 등이 가치사슬 기업을 분류하는데 보다 중요하게 사용되고 있음을 알 수 있었다.

본 연구는 특정 기업의 거래처 기업들이 그 기업의 가치사슬에 속하는지 여부에 따라 거래처 기업들을 분류하는 데에 있어 기계학습 알고리즘을 통해 자동적으로 분류가 가능한지에만 초점을 두고 있다. 때문에 가치사슬 분류 알고리즘을 타 산업에 적용하여 알고리즘의 일반화 가능성을 판단하는 것은 추후 연구과제가 될 것이다. 또한 본 연구의 분석 범위는 기업-품목중심의 후방 1차 거래처들의 가치사슬 분류였으나 향후 연구에는 분석 범위를 넓혀가는 것도 고려할 가치가 있는 문제일 것이다.

References

[1] Archer, K. J. and Kimes, R. V., “Empirical characterization of random forest variable importance measures,” Computational Statistics & Data Analysis, Vol. 52, No. 4, pp. 2249-2260, 2008.

[2] Barney, J. B. and Ouchi, W. G., Organizational economics, San Francisco: Jossey-Bass, 1986.

[3] Breiman, L., “Random Forests,” Machine learning, Vol. 45, No. 1, pp. 5-32, 2001.

[4] Brown, I. and Mues, C., “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” Expert Systems with Applications, Vol. 39, No. 3, pp. 3446-3453, 2012.

[5] Choi, S. H. and Choi, J. I., “GVC Case Analysis of the Motor Industry: Focusing on Hyundai Motor,” Journal of Digital Convergence, Vol. 14, No. 12, pp. 73-84, 2016.

[6] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E., “Querying

and mining of time series data: experimental comparison of representations and distance measures," Proceedings of the VLDB Endowment, Vol. 1, No. 2, pp. 1542-1552, 2008.

[7] Gang, S. H., Kim, C. H., and Chung, H. Y., "Machining process improvement of automobile hub assembly parts," Proceedings of the Korea Academy Industrial Cooperation Society, pp. 242-244, 2015.

[8] Go, S. W., Hong, D. P., Gang, S. H., Do, J. H., Lee, G. H., and Yu, S., S., Structural changes and countermeasures in each sector of the economy in the digital economy 3, Korea Information Strategy Development Institute, pp. 1-208, 2005.

[9] Han, E. J., Screening Test Data Analysis for Cataract Happening Prediction Model using Random forest, Yonsei University Graduate School of Medical Statistics, Master's Thesis, 2004.

[10] Hong, J. S., Park, K. H., and Park, J. R., "Hybrid Classifiers of Classification Techniques for Mixed Data," Journal of the Korean Institute of Industrial Engineers, Vol. 43, No. 5, pp. 341-349, 2017.

[11] Kim, C. S., Jo, H. J., and Jeong, J. H., "Modular Production and Hyundai Production System: The Case of Hyundai Mobis," Economy and Society, Vol. 92, pp. 351-385, 2011.

[12] Kim, J. H., "Development of Fostering Strategies for MICE Industry through the Value Chain Analysis," Northeast Asia Tourism Research, Vol. 7, No. 4, pp. 131-150, 2011.

[13] Kim, K. S., "The Characteristics of Corporate Growth and Innovation in the Materials, Components, and Equipments Sectors of Korean Display Industrial Value Chain," Journal of Korea Technology Innovation Society, Vol. 20, No. 1 pp. 205-238, 2017.

[14] Kim, S. J. and An, H. C., "Random Forest's Assessment Model for Corporate Bond Ratings," Korea Intelligent Information Systems Society Spring Conference, pp. 371-376, 2014.

[15] Kim, T. J., Lee, J. H., and Hong, J. S., "Supply Network Analysis of Second and Third Outsourcing Firms with E-Invoice at Automobile Parts Industry: Focused to Brake Manufacturing Firms," The Jounal of Society for e-Business Studies, Vol. 21, No. 3, pp. 79-99, 2016.

[16] Kotsiantis, S. B., Zaharakis, I., and Pintelas, P., "Supervised machine learning: A review of classification techniques," Informatica, Vol. 31, pp. 249-268, 2007.

[17] Kwon, A. N., Variable Selection Using Random Forest, Inha University Graduate School of Statistics, Master's Thesis, 2013.

[18] Lee, H. J., Park, J. S., and Kim, M. T., "Transformation of Value Chain and Business Models in the 3G Mobile Service Industry," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 1833-1836, 2005.

[19] Lee, H. S., Lim, D. H., and Mun, Y. S., “Value chain analysis system using company data,” Korean Institute Of Industrial Engineers Fall Conference, pp. 1974-1985, 2016.

[20] Lee, R. E., Kim, K. T., Lee, S. J., Jeong, G. J., Lee, S. J., Lee, H. S., Mun, Y. S., and Lim, D. H., “Data-based Value Chain Construction Algorithm Development and Smart Device Application,” Korean Operations Research and Management Society Spring Conference, pp. 109-128, 2016.

[21] Liaw, A., and Wiener, M., “Classification and Regression by Random Forest,” R News, Vol. 2/3, pp. 18-22, 2002.

[22] Linden, G., Kraemer, K. L., and Dedrick, J., “Who captures value in a global innovation network?: the case of Apple’s iPod,” Communications of the ACM, Vol. 52, No. 3, pp. 140-144, 2009.

[23] Li, R. H. and Belford, G. G., “Instability of decision tree classification algorithms,” In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 570-575, 2002.

[24] Liu, M., Wang, M., Wang, J., and Li, D., “Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar,” Sensors and Actuators B: Chemical, Vol. 177, pp. 970-980, 2013.

[25] Macher, J. T., Mowery, D. C., and Simcoe, T. S., “e-Business and disintegration of the semiconductor industry value chain,” Industry and Innovation, Vol. 9, No. 3, 155-181, 2002.

[26] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., and de Mendonça, A., “Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests,” BMC research notes, Vol. 4, No. 1, p. 299, 2011.

[27] Murthy, S. K., “Automatic construction of decision trees from data: A multi-disciplinary survey,” Data mining and knowledge discovery, Vol. 2, No. 4, pp. 345-389, 1998.

[28] Pal, M., “Random forest classifier for remote sensing classification,” International Journal of Remote Sensing, Vol. 26, No. 1, pp. 217-222, 2005.

[29] Park, C. D., Chae, Y. J., and Park, J. G., “An Analysis on the Value Chain of Korean Bioenergy Industry,” Journal of Energy Engineering, Vol. 23, No. 2, pp. 102-113, 2014.

[30] Park, J. M., “An Empirical Study on the Impact of Relationship between Parent and Collaboration Companies on Business Performance,” Journal of Industrial Economics and Business, Vol. 15, No. 1, pp.

303-319, 2002.

[31] Park, K. S. and Lee, C. W., “Value Chain System and Management of Cultural Contents Industry in Daegu,” Journal of The Korean Association of Regional Geographers, Vol. 13, No. 2, pp. 171-186, 2007.

[32] Porter, M. E. and Advantage, C., Creating and Sustaining Superior Performance, 1985.

[33] Rahani, A. R. and Al-Ashraf, M., “Production flow analysis through value stream mapping: a lean manufacturing process case study,” Procedia Engineering, Vol. 41, pp. 1727-1734, 2012.

[34] Ryu, J. H., Choi, T. G., and Park, J. G., “An Analysis on the Value Chain and the Value System of the Korean Wind Power Industry,” Journal of Energy Engineering, Vol. 23, No. 1, pp. 46-57, 2014.

[35] Statnikov, A., Wang, L., and Aliferis, C. F., “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” BMC bioinformatics, Vol. 9, No. 1, p. 319, 2008.

[36] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P., “Random forest: a classification and regression tool for compound classification and QSAR modeling,” Journal of chemical information and computer sciences, Vol. 43, No. 6, pp. 1947-1958, 2003.

[37 Yeo, I. G., Lee, U. H., Gang, S. R., Im, B. H., Ji, Y. G., Min, N. G., Jo, N. Y., Lee, J. M., Jang, Y. S., and Kim, Y. M., Planning Report of the Technology Roadmap for the Industrial Convergence-Transportation System(Automobile), pp. 1-197, Korea Industrial Technology Development Agency, 2010.

[38] Yun, M. S., “Industry Cluster,” INCHAM Business News, pp. 15-23, 2003.

[39] Zhang, G., Patuwo, B. E., and Hu, M. Y., “Forecasting with artificial neural networks:: The state of the art,” International Journal of forecasting, Vol. 14, No. 1, pp. 35-62, 1998.

저 자 소 개

김태진 (E-mail: taejin@scp.re.kr)
1993년 서울대학교 자원공학과 (학사)
1996년 서울대학교 자원공학과 (석사)
2011년 서울과학기술대학교 산업정보시스템공학 (박사)
2014년~현재 (주)에스씨플랫폼 대표
관심분야 산업별 가치사슬망 분석, 데이터마이닝, 빅데이터

홍정식 (E-mail: hong@seoultech.ac.kr)
1982년 서울대학교 산업공학과 (학사)
1985년 서울대학교 산업공학과 (석사)
1988년 서울대학교 산업공학과 (박사)
1989년~현재 서울과학기술대학교 글로벌융합산업공학과 교수
관심분야 신제품 확산 및 수요예측, 데이터마이닝

전윤수 (E-mail: yunsu3761@seoultech.ac.kr)
2016년 서울과학기술대학교 글로벌융합산업공학과 (학사)
2016년~현재 서울과학기술대학교 일반대학원 데이터사이언스학과 (석사)
관심분야 데이터마이닝, 산업공학

박종률 (E-mail: jrpark924@seoultech.ac.kr)
2016년 서울과학기술대학교 글로벌융합산업공학과 (학사)
2016년~현재 서울과학기술대학교 일반대학원 데이터사이언스학과 (석사)
관심분야 데이터마이닝, 산업공학

안태욱 (E-mail: taeukahn@hanmail.net)
2001년 국립금오공과대학교 컴퓨터공학과 (학사)
2003년~현재 ㈜비즈니스온커뮤니케이션 이사
관심분야 전자세금계산서, 산업별 가치사슬망 분석, 지능형 서비스