지식재산연구 제17권 제2호(2022. 6)
©한국지식재산연구원
The Journal of Intellectual Property
Vol.17 No.2 June 2022
https://doi.org/10.34122/jip.2022.17.2.295
Print ISSN 1975-5945 | Online ISSN 2733-8487

투고일자: 2022년 5월 3일
심사일자: 2022년 5월 19일(심사위원 1), 2022년 5월 10일(심사위원 2), 2022년 5월 18일(심사위원 3)
게재확정일자: 2022년 5월 24일

특허정보와 기계학습을 활용한 산업기술수준 측정 방법 연구

이철주* · 차현진** · 이정우*** · 고병철**** · 한종석*****

Ⅰ. 서 론
Ⅱ. 이론적 배경 및 선행연구 고찰
1. 산업기술수준조사
2. 특허분석을 통한 기술경쟁력 측정
3. 인공신경망, 랜덤포레스트 및 XGboost 알고리즘
Ⅲ. 연구 설계
1. 분석 대상 데이터
2. 변수 정의 및 기초통계량
3. 분석 절차 및 방법
Ⅳ. 분석 결과
1. 변수 간 상관관계 및 선형회귀분석 결과
2. 기술수준 측정 결과
3. 한 · 중 · 일 3개국 분석
Ⅴ. 토론 및 결론

* 한국산업기술평가관리원 수석연구원, 교신저자.
** 한국산업기술평가관리원 책임연구원.
*** 한국산업기술평가관리원 팀장.
**** 한국산업기술평가관리원 단장.
***** 한국산업기술평가관리원 본부장.

본 논문의 완성도 제고를 위해 귀중한 조언을 해 주신 익명의 심사위원님들께 감사드립니다.

초 록

기술정책 또는 기술전략의 수립을 위한 기초 정보로서 기술수준에 대한 정확한 측정이 필요하다. 그러나 통상 실시되고 있는 전문가 설문을 통한 기술수준조사 방식은 객관성이 부족할 수 있고 상당한 시간과 비용이 소요되는 문제가 있어 본 연구는 특허정보를 활용하여 객관적이고 용이하게 기술수준을 측정하는 방법론을 도출하고자 한다. 본 연구는 주요 5개국의 산업기술분야 특허 경쟁력 측정결과와 산업기술수준조사 결과를 연계하여, 특허 경쟁력으로부터 산업기술수준을 측정하였다. 특허지표를 독립변수로 기술수준조사결과를 종속변수로 사용하여 선형회귀 분석을 실시한 결과 각 지표별로 산업기술수준의 결정에 미치는 영향력을 확인하였고, 다음으로 인공신경망, 랜덤포레스트 및 XGboost를 활용하여 기술수준을 측정한 결과 선형회귀 방법 대비 예측 성능이 우수함을 확인하였다. 본 연구는 다양한 특허지표와 기계학습 방법을 도입하여 정확도가 개선된 산업기술수준 측정 방법론을 개발하였다는 데 의의가 있으며, 본 연구 결과가 전문가 설문조사 방식의 기술수준조사를 보완하는 도구로 활용되기를 기대한다.

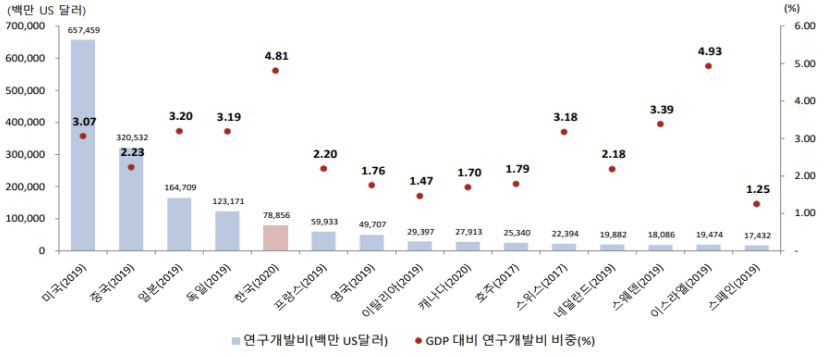
주제어

기술수준, 특허 정보, 특허 지표, 머신러닝, 인공신경망, 랜덤포레스트, XGboost

Ⅰ. 서 론

2020년 기준 우리나라의 GDP 대비 총 연구개발비는 4.81%로 세계 2위 수준이나 우리나라의 총 연구개발비는 78,856백만 달러(약 93조 원)[1] 수준이며(〈그림 1〉), 이는 미국의 12%, 중국의 25%, 일본의 48% 정도에 해당한다. 이와 같이 우리나라의 연구개발비는 전 세계 주요 강대국 대비 절대 규모가 작은 수준으로 이를 효율적으로 사용하기 위해서는 정교한 기술정책 또는 기술전략이 필요하다. 예를 들어 국가 또는 기업에서 기술분야별 투자 우선순위를 결정하거나 국제 공동연구가 필요한 분야를 발굴하기 위해서는 기술분야별로 타 국가 또는 기업 대비 경쟁력, 즉 기술수준에 대한 정확한 측정이 필요하다.

〈그림 1〉 주요 국가별 연구개발비 현황[2]



기술수준 또는 기술 경쟁력을 측정하기 위하여 특허 또는 논문의 양과 인용 강도 등을 측정하는 연구가 다양한 분야에서 행해졌다. 예를 들어 Choi et al.(2019)[3]은 약용 식물에 대하여 10개국의 특허경쟁력을 분석한 바 있으

1) 환율은 1,180.28원/US달러 적용.
2) 김한울 외 2인, "2020년도 연구개발활동조사 결과", 과학기술정보통신부, 2020, 4면.

며, Wu et al.(2019)[4]은 나노기술에 대하여 특허 분석을 통하여 중국과 미국의 경쟁력을 비교하였다. 또한 태양광전지 분야의 경우 구기관 등(2012)[5]과 유소진 · 이재승(2021)[6]은 양적 · 질적 특허지표 분석을 통하여, 하수진 등(2020)[7]은 연구논문 동향을 분석함으로써 주요 국가별 기술수준을 확인하였다.

한편, 상기와 같이 특허, 논문 정보를 활용한 기술수준 측정 방법과 더불어 기술 분야 전문가를 대상으로 한 설문조사 방식의 기술수준조사 또한 활발하게 실시되고 있다. 국내 다수 부처와 산하기관에서는 각 소관 기술분야에 대해 기술수준조사를 실시하고 있는데 부처별 주요 기술수준조사 현황을 살펴보면 다음과 같다. 과학기술정보통신부는 과학기술과 정보통신 분야에 대해, 산업통상자원부(이하 '산업부'라 칭함)는 산업기술 전반에 대해, 보건복지부는 의약과 보건산업 분야에 대하여, 환경부는 환경기술 분야에 대해 기술수준조사를 실시하고 있다(한국과학기술기획평가원, 2022).[8] 기술수준조사는 일반적으로 산업계 임직원, 대학 교수 또는 국공립 연구소 연구원 등 기술 전문가를 대상으로 반복적 설문조사를 실시하여 측정의 신뢰성을 높이는 방식의 델파이 방법(Jo & Jo, 2004[9]; Lee et al., 2014[10])을 활용하고 있는데, 이

3) Choi, Ji Weon et al., "Technology trends and patenting prospects of medicinal plants in Korea", *Korean Journal of Medicinal Crop Science*, Vol.27, No.2(2019), pp.75-85.

4) Wu, L. et al., "Comparing nanotechnology landscapes in the US and China: a patent analysis perspective", *Journal of nanoparticle research*, Vol.21, No.8(2019), pp.1-20.

5) 구기관 외 3인, "국내외 신재생에너지 기술 경쟁력 분석: 태양광 · 연료전지를 중심으로", 『신재생에너지』, 제8권 제3호(2012), 30-37면.

6) 유소진 · 이재승, "스트레처블 태양광전지 분야 특허기반 기술경쟁력 연구", 『한국태양광발전학회지』, 제7권 제1호(2021), 9-16면.

7) 하수진 외 2인, "태양전지 분야 주요 5개국의 연구논문 동향 및 기술수준 조사 · 분석", 『한국기후변화학회지』, 제12권 제1호(2021), 37-47면.

8) 한국과학기술기획평가원, "분야별 기술수준평가", 과학기술정책지원서비스, 〈https://www.k2base.re.kr/techLevelEval/list.do〉, 검색일: 2022년 4월 27일.

9) Jo, Yong-Gon & Jo, Geun-Tae, "Formulating R&D strategy for core technologies in biotechnology using the delphi and the AHP", In Proceedings of the Korean Operations and Management Science Society Conference, The Korean Operations Research and Management Science Society, 2004, pp.185-188.

10) 이동헌 외 2인, "국내외 기술수준조사 및 기술수준평가 연구 동향 분석," *ICROS*, 제20

러한 설문조사 방식의 기술수준조사는 상당한 시간과 비용이 소요되며 전문가의 전문성 또는 견해에 있어 편차(bias)가 있을 수 있어 이를 대체하거나 보완할 수 있는 방법이 필요한 실정이다. 본 연구는 이러한 문제점을 보완하기 위하여 특허정보를 활용하여 객관적이고 신속하게 국내 산업기술의 수준을 측정하는 방법을 제안하고자 한다.

II. 이론적 배경 및 선행연구 고찰

1. 산업기술수준조사

본 연구의 대상인 산업기술수준조사는 산업부 산하의 정부 R&D 투자 전문기관인 한국산업기술평가관리원(이하 'KEIT'라 칭함)에 의해 2년마다 실시되고 있으며, 최근(2021년) 조사 분야는 산업부의 제7차(2019~2023년) 산업기술혁신계획에 명시된 산업부 중점 투자 분야이다. 연구대상을 산업기술수준조사로 한정한 이유는 다음과 같다. 자연현상의 원리를 탐구하고 이론을 정립하는 등 학문적 연구가 중심이 되는 과학기술분야는 논문이 중요한 역할을 하나, 산업기술 분야는 특허의 영향력이 더 큰 분야로서 논문에 대한 정보를 사용하지 않고 특허정보만으로 기술수준의 예측이 가능할 것으로 판단하였기 때문이다. 또한 산업기술수준조사 결과는 다양한 기술분야를 포함하고 있으므로, 정보통신기술이나 환경기술 등과 같이 특정 기술분야가 아닌 다양한 기술 분야를 포괄하는 일반적인 연구 결과를 도출하기 위함이다.

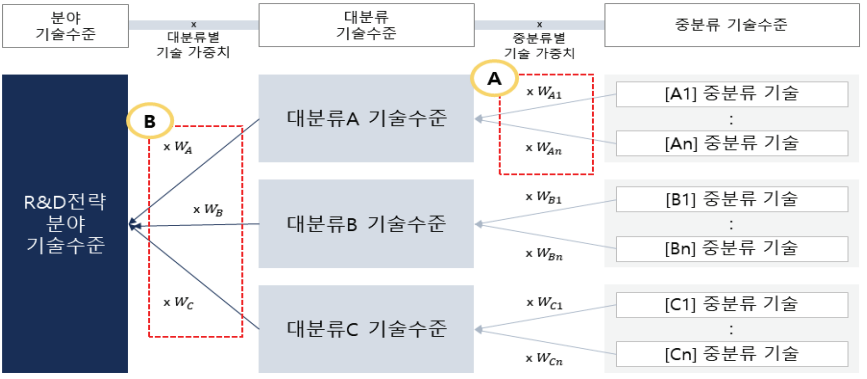
2021년도 산업기술수준조사는 2021. 11. 8.부터 2022. 1. 3.까지 약 8주간에 걸쳐 실시되었고 산 · 학 · 연 등 산업기술 전문가 1,678명이 20개 분야 내 70개 대분류 265개 중분류 기술에 대해 주요 5개국(한국, 미국, 일본, 중국, EU)의 기술수준에 대해 응답하였다.

권 제1호(2014), pp.24-27.

산업기술수준의 산출 방법은 〈그림 2〉와 같은데 이를 살펴보면, 우선 전문가를 통해 가장 하위 단계인 중분류 기술의 기술수준과 각 중분류 기술별 가중치[11]를 조사하여 이로부터 대분류 차원의 기술수준을 결정하게 된다. 다음으로 전문가들을 통해 대분류 기술별 가중치를 조사하여 전략 분야별 기술수준을 최종 결정하게 된다. 예를 들어 〈그림 2〉에서 대분류A에 해당하는 중분류 기술들(A_1~A_n)의 합이 100%가 되도록 가중치를 적용하고(Ⓐ), R&D 전략분야를 구성하고 있는 대분류 기술들(A, B, C)의 합이 100%가 되도록 가중치를 적용하여(Ⓑ) R&D 전략분야의 기술수준이 산출된다.

본 연구에서는 중분류 기술(이하 '세부 기술분야'라 칭함)에 한정하여 분석을 실시하고자 한다. 세부 기술분야를 활용하는 이유는 가중치에 의해 기술수준이 변동되는 문제점이 없으며, 'R&D 전략분야' 또는 '대분류 기술' 단위의 분석과 비교하여 더 많은 데이터의 확보가 가능하여 머신러닝 방법을 활용한 분석에 유리하기 때문이다.

〈그림 2〉 산업기술수준 산출 방법



11) 해당 기술이 상위 기술(분야)이 적절한 기능을 수행하는 데 있어 중요한 정도, 또는 해당기술이 상위 기술(분야)에서 차지하는 비중.

2. 특허분석을 통한 기술경쟁력 측정

특허분석을 통한 국가별 기술경쟁력의 측정에 대한 연구사례는 앞서 서론 부분에서 간략히 살펴보았다. 이들 연구를 포함하여 다수의 연구와 동 분야 분석지침(OECD, 2009[12]; 선동주 등, 2014[13])에서는 국가별로 특정 기술분야에 대한 다양한 특허지표를 측정하여 해당 기술분야에 대한 국가별 기술수준을 측정하고 있다(〈표 1〉).

이러한 특허지표를 활용한 기존 다수 연구들은 수 개의 특허지표 각각에 대하여(구기관 등, 2012[14]; Choi et al., 2019[15]; 유소진 · 이재승, 2021[16]) 또는 특

〈표 1〉 주요 특허 분석 지표

측정지표 명	지표 설명 및 산식
특허활동도 (PAI: Patent Activity Index)	• 특정 기술분야에 대해 특정 국가의 특허 점유율 • PAI = 특정 기술분야에서 특정 국가의 특허 건수 / 특정 기술분야에서 전체 국가의 특허 건수
특허집중도 (PCI: Patent Concentration Index)	• 특정 기술분야에 대해 특정 국가의 출원이 활성화된 정도 • PCI = 특정 기술분야에서 특정 국가의 특허 비중 / 전체 기술분야에서 특정국가의 특허 비중
특허영향력 (PII: Patent Impact Index))	• 특정기술 분야에 대해 특정 국가의 특허가 인용된 정도 • PII = 특정 기술분야에서 특정 국가 특허의 평균 피인용 횟수 / 특정 기술분야에서 전체 국가 특허의 평균 피인용 횟수
특허시장력 (PFS: Patent Family Size)	• 특정기술 분야에 대해 특정 국가의 패밀리 특허 보유 정도 • PFS = 특정 기술분야에서 특정 국가 특허의 평균 패밀리 특허 수 / 특정 기술분야에서 전체 국가 특허의 평균 패밀리 특허 수

12) OECD, "OECD Patent Statistics Manual", OECD, 2009.

13) 선동주 외 6인, "특허성과 지표 활용 가이드라인", 특허청, 2014.

14) 구기관 외 3인, "국내외 신재생에너지 기술 경쟁력 분석: 태양광 · 연료전지를 중심으로", 『신재생에너지』, 제8권 제3호(2012), 30-37면.

15) Choi, Ji Weon et al., "Technology trends and patenting prospects of medicinal plants in Korea", *Korean Journal of Medicinal Crop Science*, Vol.27, No.2(2019), pp.75-85.

16) 유소진 · 이재승, "스트레처블 태양광전지 분야 특허기반 기술경쟁력 연구", 『한국태양광발전학회지』, 제7권 제1호(2021), 9-16면.

허지표를 선형으로 결합하여(서규원, 2011)[17] 특정 기술분야에서 특정 주체(국가 또는 기업 등)의 기술수준을 확인하였다.

반면에 객관적 데이터인 특허지표와 주관적인 설문조사 결과인 기술수준 조사 결과를 연계하는 연구, 즉 특허정보를 분석하고 이로부터 기술수준을 예측하고자 한 선행 연구들은 매우 소수에 불과한데 이를 살펴보면 다음과 같다. 먼저, Cho & Park(2015)[18]은 이동통신기술 분야를 대상으로 특허와 논문의 서지정보를 활용하여 점유율, 집중도, 피인용도 등을 측정하고 각 지표별 가중치를 AHP로 결정하여 기술수준을 측정하는 방법론을 제안하였고 이러한 방법이 전문가 설문조사 방식의 측정 결과와 상관관계가 높음을 확인하였다. 같은 방법론을 사용하여 Han et al.(2018)[19]은 엔지니어링 모델링과 시뮬레이션 기술 분야에서 국가별 경쟁력을 측정한 바 있다. 이어지는 오종학 · 나관식(2018)[20]의 연구는 정보보안기술을 대상으로 특허의 수준이 기술수준에 미치는 영향을 파악하였다. 동 연구는 기술분야별 특허지표(특허활동도, 특허집중도, 특허시장력 및 특허영향력) 측정값을 독립변수, 전문가 설문조사에 의해 측정한 기술수준을 종속변수로 활용하여 다중선형회귀분석을 실시하여 각 독립변수(지표)별 가중치를 결정하였다.

기술수준에 영향을 미치는 특허 지표별 가중치를 결정하는 방식 측면에서 볼 때 오종학 · 나관식(2018)의 연구는 AHP 방식을 사용한 연구(Cho & Park, 2015; Han et al., 2018)보다 객관성과 편의성이 담보된 방법론으로 볼 수 있다. 다만 동 연구는 머신러닝 방법을 활용한 기술수준 예측은 시도하지 않았

17) 서규원, "특허지표를 활용한 기술수준평가 연구방법론의 개발 및 적용", 한국과학기술기획평가원, 2011, 10면.

18) Cho, I. & Park, M., "Technological-level evaluation using patent statistics: model and application in mobile communications", *Cluster Computing*, Vol.18, No.1(2015), pp.259-268.

19) Han, Yuri et al., "Analysis of global competitiveness of engineering modeling and simulation technology for next-manufacturing innovation: Using quantitative analysis of patents and papers." *ICIC Express Letters,* Vol.9, No.4(2018), pp.339-346.

20) 오종학 · 나관식, "특허지표가 선진국과 개발도상국의 기술수준에 미치는 영향 : 정보보안기술을 중심으로", 『한국창업학회지』, 제13권 제3호(2018), 75-93면.

고, 특허지표로 기술수준을 측정한 모델의 수정된 결정계수(adjusted R^2) 값을 0.479로 보고하고 있는데 이를 보다 높이기 위한 후속 연구가 필요하다고 판단된다.

또한 상기 연구들은 기술수준의 측정이 정보통신 분야에 한정되어 있어 본 연구에서는 다양한 기술분야를 포괄하는 보다 설명력과 예측력이 높은 기술수준 예측 모델을 만들고자 한다. 이를 위해 본 연구는 다양한 기술분야가 포함한 데이터 세트를 활용하고 기존 연구에 사용되지 않은 특허지표를 추가하여 예측 모델의 설명력을 높이고자 하며 또한 기술수준 예측 시 기존 연구에서 시도되지 않은 기계학습(machine learning) 방법을 도입하여 측정의 정확성을 높이고자 한다.

3. 인공신경망, 랜덤포레스트 및 XGboost 알고리즘

본 연구에서는 기계학습에서 가장 대표적인 알고리즘 중 하나인 인공신경망(ANN: Artificial Neural Networks), 랜덤포레스트(RF: Random Forest) 및 XGboost를 사용하여 기술수준을 예측하고자 한다.

인공신경망은 뇌의 신경 신호 전달 체계를 모방한 머신러닝 알고리즘으로서 기존 회귀분석 방법론에서 사용되는 설명변수, 종속변수와 함께 은닉층과 은닉노드를 사용하여 종속변수의 결정에 설명변수가 미치는 기여도를 직관적으로 확인하기 어려운 면이 있으나, 다양한 분야의 연구에서 인공신경망이 일반 선형회귀 분석 방법에 비해 예측 성능이 높은 경우가 보고되고 있다(Uysal & Roubi, 1999[21]; Zekic-Susac et al., 2004[22]; Ibrahim & Rusli, 2007[23]).

---

21) Uysal, M. & Roubi, S. E., "Artificial neural networks versus multiple regression in tourism demand analysis", *Journal of Travel Research*, Vol.38(1999), pp.111-118.

22) Zekic-Susac, M. et al., "Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models", In 26th International Conference on Information Technology Interfaces IEEE, 2004, pp.265-270.

23) Ibrahim, Z. & Rusli, D., "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression", In 21st Annual SAS Malaysia Forum, 2007, pp.1-6.

랜덤포레스트와 XGboost의 경우 다수의 분류기를 생성하고 예측값을 결합하는 앙상블(ensemble) 기계학습 방법인데 인공신경망과 비교 사용함으로써 특허지표를 이용한 산업기술수준 측정에 보다 적합한 알고리즘을 확인하고자 한다. 랜덤포레스트는 변수를 무작위로 조합하여 다수의 의사결정 나무(decision tree)를 생성하고 결정트리마다 생성한 예측값 중 가장 높은 빈도를 최종 예측값으로 선정하는 모델이다. 랜덤포레스트 모델은 의사결정 나무를 단독으로 사용하는 경우보다 과적합(over-fitting)을 줄여 성능을 높일 수 있으며, 인공신경망과 같이 예측 성능을 높이기 위해 독립변수를 스케일링할 필요가 없는 장점이 있다(Rodriguez-Galiano et al., 2012[24]; Prajwala, 2015[25]).

랜덤포레스트는 다수의 결정트리가 전체데이터에서 데이터를 무작위로 반복 추출하는 배깅(bagging) 방식인 반면, XGboost는 성능이 낮은 다수의 결정트리가 순차적으로 학습하면서 잘못 예측한 데이터에 가중치를 부여하여 예측의 정확도를 높여 가는 부스팅(boosting) 방식의 알고리즘으로서, 기존의 GradientBoost를 속도 및 과적합 규제 측면 등에서 개선한 알고리즘이다(Chen & Guestrin, 2016[26]). XGboost는 다양한 초모수 값의 적용을 통해 성능향상이 가능한데 본 연구에서 탐색한 초모수들은 다음과 같다. '반복횟수'는 모형의 성능을 개선하기 위한 boosting 시행 횟수이며, '최대 깊이'는 가지 모양의 모형 생성 시 모형의 복잡도를 결정하며, 'gamma'는 가지의 분할 여부를 결정하는 손실함수 크기를 조절하고, 'sub-sample rate'는 과적합 방지를 위해 모형 생성 시 관측치의 일부만을 사용하는 정도이다.

---

24) Rodriguez-Galiano, V. F. et al., "An assessment of the effectiveness of a random forest classifier for land-cover classification", *ISPRS journal of photogrammetry and remote sensing*, Vol.67(2012), pp.93-104.

25) Prajwala, T. R., "A comparative study on decision tree and random forest using R tool", *International journal of advanced research in computer and communication engineering*, Vol.4, No.1(2015), pp.196-199.

26) Chen, Tianqi & Guestrin, Carlos, "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp.785-794.

Ⅲ. 연구 설계

1. 분석 대상 데이터

특허데이터는 16개[27] 산업 R&D 중점 투자 분야(이하 '분야'라 칭함)와 그 하위 기술인 195개 기술(이하 '세부기술 분야'라 칭함)에 관하여 주요 5개국이 미국 특허상표청(USPTO)에 2009년부터 2018년까지 출원한 특허를 대상으로 수집하였다. 미국 특허상표청 데이터를 사용한 이유는 다수 국가 내 기업, 기관과 연구자들이 주요한 특허를 전 세계에서 가장 큰 시장 중 하나인 미국에 출원하고 있고, 미국 특허법상 선행 특허를 엄격히 표시하도록 하고 있어 인용 정보를 활용한 기술수준의 측정이 용이하기 때문이다. 2018년까지로 한정한 사유는 특허제도의 특징상 최근 출원된 특허는 비공개되는 경우가 많아 정확한 건수 파악이 어렵고, 공개 기간이 짧을수록 피인용 횟수가 급격히 감소하기 때문이다. 특허데이터는 2021년 9월 중 특허 전문기관에 의해 추출되고 관련이 없는 건(noise)을 제거하는 작업을 거친 결과 전체 특허 건수는 71,857건이었다.

추출된 특허 데이터 중 분석 대상 데이터는 다음과 같은 방식으로 한정하였다. 우선 전체 195개 세부 기술분야 중 주요 5개국 중 1개 국가라도 세부 기술분야에서 특허출원 건수가 5건 미만인 106개 세부 기술분야를 우선 제외하였다. 특허건수가 0인 경우, 특허지표 중 PAI는 0으로, 그 외의 지표값은 무한대로 측정되며, 특허건수가 매우 적은 경우 특허정보를 이용하여 기술수준을 측정하고자 하는 본 연구의 범위에서 벗어나기 때문이다. 다음으로 특허지표 측정 결과 이상치(outlier)가 임곗값[(Quartile(3)+IQR(1～3)*3)] 이상인 9개 세부 기술분야를 추가로 제외하였다.

---

27) 특허분석 이후 산업기술수준조사가 실시되었는데, 특허분석과 산업기술수준조사에서 사용된 기술분류 체계가 상이한 경우는 분석대상에서 제외(특허분석을 실시했던 20개 분야 중 4개 분야 제외).

〈표 2〉 전체 및 분석 데이터

국 가	전체 데이터	분석 데이터
CN	3,655 (5.1%)	2,052 (3.9%)
EU[28]	9,952 (13.8%)	7,271 (13.7%)
JP	11,346 (15.8%)	9,338 (17.6%)
KR	8,259 (11.5%)	4,802 (9.1%)
US	38,645 (53.8%)	29,488 (55.7%)
총합계	71,857 (100%)	52,951 (100%)

정리하면 본 연구는 〈표 2〉, 〈표 3〉과 같이 주요 5개국의 52,951건의 특허 데이터로부터 15개 기술분야 내 80개 세부 기술분야에 대한 특허지표를

〈표 3〉 분석 데이터 현황

번 호	분 야	세부 기술분야 개수	데이터 개수
1	3D프린팅	7	35
2	디지털 엔지니어링	2	10
3	맞춤형 바이오진단/치료	11	55
4	미래형 디스플레이	4	20
5	스마트홈	4	20
6	웨어러블 디바이스	6	30
7	이차전지	4	20
8	자율주행차	4	20
9	전기수소차	7	35
10	지식서비스	1	5
11	차세대반도체	12	60
12	차세대항공	2	10
13	첨단 제조장비	12	60
14	친환경 조선 · 해양플랜트	1	5
15	헬스케어	3	15
합 계		80	400

* 데이터 개수 = 세부 기술분야 개수 × 5개국

28) EU는 독일, 영국, 프랑스 등 EPO(European Patent Office)의 38개 회원국을 합산하여 분석.

측정한 결과를 독립변수로 활용하고, 종속변수의 경우 국가별 세부기술분야에 대한 2021년도 산업기술수준조사 결괏값을 사용하여 연구를 수행하고자 한다(데이터 추출 · 생성 과정은 부록 1, 2 참고).

2. 변수 정의 및 기초통계량

종속변수(산업기술수준)에 영향을 미치는 독립변수들의 경우 기존 선행연구들(Cho & Park, 2015[29]; 오종학 · 나관식, 2018[30])과 동 분야 분석지침(OECD, 2009[31]; 선동주 등, 2014[32])에서 사용한 특허지표를 포함하되 〈표 4〉처럼 일부 지표를 변형 또는 추가하여 사용하였다. PAI의 경우 모든 국가의 평균 특허점유율(5개국인 경우 0.2)로 나눈 값인 PAI_M을 사용하였는데, 이는 PII 또는 PFS 지표와 같이 평균과 동일한 경우 1 값을 갖게 하여 우열을 직관적으로 파악하기 위함이다. PFS의 경우 패밀리 특허수를 계수하는 방법(PFS_1)과 패밀리 국가 수(패밀리 특허가 출원된 국가의 개수)를 계수하는 방법(PFS_2)을 병행하여 사용하였다. 예를 들어 패밀리 특허 10개가 미국에 5건, 일본에 2건, 독일에 3건 출원되었다면 패밀리 국가수는 3인데 이는 3개 국가에 출원되었기 때문이다.

---

29) Cho, I. & Park, M., "Technological-level evaluation using patent statistics: model and application in mobile communications", *Cluster Computing*, Vol.18, No.1(2015), pp.259-268.

30) 오종학 · 나관식, "특허지표가 선진국과 개발도상국의 기술수준에 미치는 영향: 정보보안기술을 중심으로",『한국창업학회지』, 제13권 제3호(2018), 75-93면.

31) OECD, "OECD Patent Statistics Manual", OECD, 2009.

32) 선동주 외 6인, "특허성과 지표 활용 가이드라인", 특허청, 2014.

〈표 4〉 변수 정의

변수 정의	관련 특허지표
PAI_M = 특정 기술분야에서 특정 국가의 특허출원 비중 / 특정 기술분야에서 전체 국가의 평균 특허출원 비중 * 분모는 n개국 대상인 경우 1/n	특허활동도 (PAI: Patent Activity Index)
PCI = 특정 기술분야에서 특정 국가의 특허출원 비중 / 전체 기술분야에서 특정국가의 특허출원 비중	특허집중도 (PCI: Patent Concentration Index)
PII = 특정 기술분야에서 특정 국가 등록 특허의 평균 피인용 횟수 / 특정 기술분야에서 전체 국가 등록 특허의 평균 피인용 횟수	특허영향력 (PII: Patent Impact Index))
PFS_1 = 특정 기술분야에서 특정 국가 등록 특허의 평균 패밀리 특허 수 / 특정 기술분야에서 전체 국가 등록 특허의 평균 패밀리 특허 수 PFS_2 = 특정 기술분야에서 특정 국가 등록 특허의 평균 패밀리 국가 수 / 특정 기술분야에서 전체 국가 등록 특허의 평균 패밀리 국가 수	특허시장력 (PFS: Patent Family Size)
TECH_LEVEL = 특정 기술분야의 2021년도 산업기술수준조사 설문 결과 * 특정 기술분야에서 산업기술수준이 최고인 국가의 산업기술수준을 100으로 보고 환산한 값	–

* 상기에서 기술분야는 세부 기술분야를 지칭

다음으로 상기 변수로 측정한 400개 데이터의 기초통계량은 〈표 5〉와 같다.

〈표 5〉 변수의 기초통계량

	변수명 (variable name)	개수 (count)	평균 (mean)	표준편자 (std.)	최솟값 (min.)	최댓값 (max.)
Dependent Variable	TECH_LEVEL	400	88.48	8.67	69.0	100
Independent Variable	PAI_M	400	1	1.06	0.04	4.41
	PCI	400	0.97	0.68	0.06	3.80
	PII	400	0.75	0.50	0	3.28
	PFS_1	400	0.74	0.48	0	2.56
	PFS_2	400	1.00	0.30	0	2.04

3. 분석 절차 및 방법

상기 종속변수와 독립변수를 활용하여 일반 선형회귀 분석을 실시함으로써 회귀식의 설명력과 기술수준에 영향을 미치는 독립변수(특허지표)별 가중치를 파악하고, 다음으로 인공신경망과 랜덤포레스트 알고리즘을 활용하여 보다 정확하게 기술수준을 예측하고자 한다.

예측의 정확도는 예측값과 실젯값의 차이에 기반한 RMSE(Root Mean Square Error)와 MAE(Mean Average Error)를 활용하고자 하며(수식 1, 2), 추가로 수정된 결정계수(adjusted R^2)를 사용하여 예측 모델의 설명력을 확인하고자 한다.

$$\text{RMSE}= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \text{(수식 1)}$$

$$\text{MAE}= \frac{1}{n}\sum_{i=1}^{n} \mid Y_i - \hat{Y}_i \mid \qquad \text{(수식 2)}$$

참고로 RMSE는 가장 많이 사용되는 지표 중 하나이나 이상치가 있는 경우 MAE보다 값이 크게 측정되는 경향이 있다. 회귀분석은 R을 사용하였고, 인공신경망과 랜덤포레스트 및 XGboost는 Python 프로그램의 머신러닝 라이브러리를 활용하였다.

IV. 분석 결과

1. 변수간 상관관계 및 선형회귀분석 결과

종속변수 및 독립변수 간 피어슨 상관계수를 측정한 결과는 〈표 6〉과 같다.

〈표 6〉 변수 간 상관계수 측정결과

	TECH_LEVEL	PAI_M	PCI	PII	PFS_1	PFS_2
TECH_LEVEL	1					
PAI_M	0.71	1				
PCI	0.41	0.50	1			
PII	0.43	0.43	0.07	1		
PFS_1	0.42	0.49	0.15	0.54	1	
PFS_2	0.17	-0.00	0.16	0.12	0.32	1

설문조사로부터 측정된 세부 기술분야별 산업기술수준(TECH_LEVEL)은 PAI_M과 가장 상관관계가 높고 이어서 PII, PFS_1, PCI, PFS_2 순으로 상관계수가 높게 측정된다. 독립변수 간 상관관계는 전반적으로 높지 않은 수준이나 PAI_M과 PFS_1이 타 지표들과 상관계수가 비교적 높게 측정되었다.

다음으로 선형 회귀분석을 실시하기 전 독립변수들 간의 다중공선성 여부를 확인하였다. 모든 독립변수를 포함하더라도 변수별 VIF(Variation Inflation Factor) 값이 모두 10 이하로 다중공선성에 문제가 없을 것으로 판단되어 모든 독립변수를 사용하여 모형을 만들었고(Model 1), 상관계수가 비교적 높게 측정된 일부 독립변수를 삭제한 모형들(Model 2~4)을 추가로 분석하였다. Model 1~4에 대해 최소자승법에 의한 선형 회귀분석을 실시한 결과는 〈표 7〉과 같다.

먼저 Model 1～4 모두 F 통계량 값을 확인한 결과 회귀식이 통계적으로 유의($p<0.01$)함을 확인하였고, 다음으로 독립변수별 회귀계수를 살펴보면 다음과 같다.

산업기술수준(TECH_LEVEL)을 결정하는 특허 요인 중 가장 계숫값이 큰 것은 특정 기술분야에서 특정 국가의 점유율로부터 산출된 지표인 PAI_M임을 알 수 있다. PFS의 경우 패밀리 국가 수 정도를 나타내는 PFS_2가 패밀리 특허 수의 정도를 나타내는 PFS_1보다 회귀계수가 더 큰 것이 확인된다. 그런데 PFS_1의 경우 PII 등 타 지표와 상관관계가 높아 PII를 포함하지 않은 Model_4에서만 유의미한 것으로 판단된다. PFS_2가 기술수준에 높은 영향

〈표 7〉 선형 회귀분석 결과

Dependent variable : TECH_LEVEL				
Independent variables	Model_1	Model_2	Model_3	Model_4
PAI_M	5.074***	4.617***	5.011***	4.975***
	(0.395)	(0.385)	(0.363)	(0.376)
PCI	0.877*	1.417***	0.909*	1.150**
	(0.528)	(0.520)	(0.522)	(0.521)
PII	2.527***	2.546***	2.416***	
	(0.722)	(0.735)	(0.668)	
PFS_1	-0.339	0.909		2.018***
	(0.832)	(0.786)		(0.728)
PFS_2	4.405***		4.237***	
	(1.099)		(1.018)	
Constant	76.496***	79.921***	76.530***	80.907***
	(1.112)	(0.726)	(1.109)	(0.677)
Observations	400	400	400	400
Adjusted R2	0.545	0.528	0.546	0.515
F Statistic	96.65***	112.51***	121.02***	142.07***

*p〈0.1; **p〈0.05; ***p〈0.01
Standard Error in parenthesis

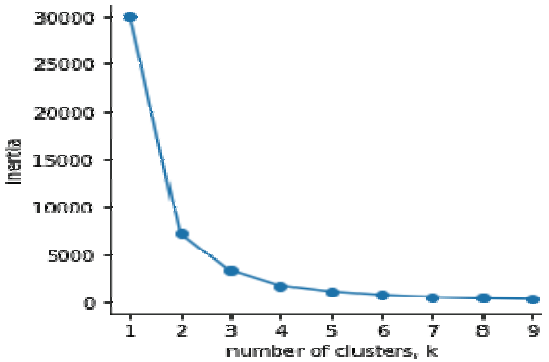
을 미치는 원인을 살펴보면 다수 국가에 패밀리 특허를 출원하는 경우 많은 비용이 소요되므로 기술수준이 높아 해당 국가에서 독점적 지위를 확보 가능한 특허에 한하여 다수 국가에 패밀리 특허를 출원한 것에 기인한다고 판단된다. 한편, 기존 다수 연구에서 기술적 우월성을 확인하는 지표로 가장 많이 사용된 특허영향력(PII)은 PAI_M 또는 PFS_2 대비 회귀식의 계수가 낮게 측정되었다. 마지막으로 집중도를 나타내는 지표인 PCI도 여타 지표보다는 약한 수준으로 기술수준에 유의미한 정(+)의 영향을 확인할 수 있다.

2. 기술수준 측정 결과

이하에서는 〈표 7〉의 Model 1~4를 대상으로 기계학습을 통하여 기술수준을 측정하고자 한다. 인공신경망 알고리즘에서는 기술수준 예측의 성능개선을 위하여 다음과 같이 독립변수 데이터를 정규화하였다. 정규화는 평균과 표준편차를 사용하여 정규화하는 방법 또는 최대-최소 값으로 정규화하는 방법 대신 중앙값(median)과 IQR을 사용하여 정규화[33]함으로써 이상치의 영향을 최소화하는 방식을 사용하였다.[34]

다음으로 데이터를 90:10 비율로 나누어 90%는 훈련 데이터(training data)로 10%는 검증 데이터로 사용하였는데, 데이터 추출 시 400개 산업기술수준 측정 결과를 k-평균 알고리즘(k-means clustering algorithm)을 사용하여(Hartigan & Wong, 1979) 상위(100~92.5점, 150개), 중위(92.2~82.6점, 136개), 하위(82.2~69.0점, 114개) 3개 그룹으로 구분한 후, 층화 무작위 추출 방식을 사용하였다.

〈그림 3〉 클러스터 개수에 따른 군집 내 거리제곱 합



33) 표준화 특허지표 = $(x_i-Q_2(x)) / (Q_3(x)-Q_1(x))$.

34) 최대 또는 최솟값이 그 사이에 있는 값들보다 매우 크거나 작다면 Min-max scaling 결과 치우친 분포를 얻게 될 수 있으며, 본 연구에서도 Min-max scaling을 한 경우는 중앙값과 IQR을 사용한 경우보다 성능이 낮았음.

참고로 산업기술수준을 3개 그룹으로 구분한 이유는 3개의 그룹부터 〈그림 3〉과 같이 군집 내 거리제곱의 합(inertia value)이 급격히 감소하기 때문이다.

상기와 같은 방법으로 추출된 훈련 데이터를 사용하여 i) 최소자승(OLS, Ordinary Least Square)에 의한 선형회귀 방법, ii) 인공신경망(Artificial Neural Networks), iii) 랜덤포레스트(Random Forest), iv) XGboost 알고리즘으로 최적의 매개변수를 확정하고, 검증데이터를 활용하여 예측값의 성능(RMSE, MAE)을 측정한 결과는 〈표 8〉과 같다. 참고로 성능 측정은 10분할 교차검증(10-fold cross-validation)을 실시한 결과이며, 각 방법별로 Model 1~4 중 가장 최고의 성능을 나타낸 모델을 기재하였다.

〈표 8〉 산업기술수준 예측 성능 비교

Method	performance measure			Model
	RMSE	MAE	adj. R^2	
OLS Regression	5.87	4.73	0.47	Model 3
Artificial Neural Networks	5.22	4.03	0.58	
Random Forest	5.39	4.18	0.54	Model 1
XGboost	5.29	4.14	0.56	Model 1

기계학습 방법의 경우 예측 성능을 높이기 위해 하이퍼 파라미터(hyper parameter)를 그리드 서치(grid search)로 탐색하였고, 최고의 성능을 나타낸 하이퍼파라미터 값은 다음과 같다. 인공신경망의 경우 1개의 은닉층에 21개 노드가 사용되었고 hyperbolic tangent를 활성화 함수(activation function)로 사용하였으며 학습 속도(learning rate)는 0.01이었다. 랜덤포레스트의 경우 트리 수 350개, 최대 깊이 7, 최대 특징(feature) 3개에서 가장 높은 성능이 확인되었다. XGboost의 경우 트리 수 240개, 최대 깊이 4, 학습률은 0.03이었고, 서브 샘플(sub sample)은 0.4, gamma는 100이었다. 〈표 8〉을 보면 인공신경망, XGboost, 랜덤포레스트, 선형회귀 순으로 성능이 높은 것이 확인되어 기계학습이 선형 회귀분석보다 예측에 보다 효과적임을 알 수 있다. 이는

기계학습이 예측모형에 다양한 함수 형태를 가정할 수 있고(예를 들어 PII 증가에 따라 기술수준도 증가하나 증가율은 완만히 감소), 인공신경망의 경우 활성화 함수를 사용함으로써 특허지표 값 변화에 따른 기술수준 변화의 비선형성(특허지표가 임곗값 이상인 경우 기술수준의 상승)을 모사할 수 있기 때문으로 판단된다.

3. 한·중·일 3개국 분석

이하에서는 상기와 동일한 방법론을 활용하되 한국, 중국, 일본 3개국 국적 특허에 한정하여 기술수준을 측정해 보고자 한다. 미국과 EU 국적 특허를 제외하고 한중일 3개국만을 대상으로 한 연구는 다음과 같은 장점이 있을 수 있다. 한중일 3국은 동북아시아에 위치하여 지정학적 위치가 유사하며 3개국 모두 비영어권 국가들로 출원 시 언어 장벽이 유사한 조건을 가지게 되어 앞서 미국과 EU를 포함한 5개국 분석 대비 지리적 요인, 언어 요인[35]이 배제되어 보다 산업기술수준 결정 요인의 도출에 있어 객관성이 제고될 것으로 판단된다.

분석용 데이터는 앞서 주요 5개국 분석 시 활용한 데이터 수집 방식을 동일하게 활용하여 264개(88개 분야×3개 국가) 데이터 세트를 구성하였고, 기초통계량은 〈표 9〉와 같다.

35) 미국 특허상표청(USPTO) DB를 사용함으로써 미국 국적 특허들이 점유율이 높고 피인용 횟수 또한 많을 수 있으며, EU 특허의 경우 지리적 위치로 인해 인접국에 패밀리 특허 출원이 많을 수 있다고 판단됨.

〈표 9〉 변수의 기초통계량(한중일 3국 한정)

	변수명 (variable name)	개수 (count)	평균 (mean)	표준편자 (std.)	최솟값 (min.)	최댓값 (max.)
Dependent Variable	TECH_LEVEL	264	93.79	6.76	65.38	100
Independent Variable	PAI_M	264	1.00	0.68	0.03	2.79
	PCI	264	1.04	0.66	0.02	3.94
	PII	264	0.92	0.52	0	3.01
	PFS_1	264	0.94	0.36	0	2.36
	PFS_2	264	0.94	0.23	0	1.64

다음으로 변수 간 피어슨 상관계수를 측정한 결과는 〈표 10〉과 같다.

〈표 10〉 변수 간 상관계수 측정결과(한중일 3국 한정)

	TECH_ LEVEL	PAI_M	PCI	PII	PFS_1	PFS_2
TECH_LEVEL	1					
PAI_M	0.53	1				
PCI	0.09	0.68	1			
PII	0.17	0.19	0.03	1		
PFS_1	0.22	0.16	0.01	0.28	1	
PFS_2	0.34	0.27	0.13	0.26	0.68	1

독립변수 중 PCI의 경우, 종속변수와 상관계수 측정 결과가 0.09로 0에 가깝고 PAI_M과는 상관계수가 높게(0.64) 측정되는데 이는 3개국만을 대상으로 할 경우 앞서 5개국 측정 결과 대비 국가별 점유율의 차이가 작아 PCI가 PAI_M과는 차별화된 독립변수로서 기능을 하지 못하기 때문으로 판단된다. 이러한 사유로 PCI는 이후 회귀모형을 구성할 때 제외하였다.

PFS_1과 PFS_2 간 상관계수도 0.68에 달하여 동시에 사용할 경우 다중공선성 발생이 확인되어(VIF 측정 값이 10을 초과) 〈표 11〉과 같이 상관관계가 높은 독립변수들을 일부 생략하여 회귀분석을 실시하였다.

〈표 11〉 선형 회귀분석 결과(한중일 3국 한정)

Dependent variable : TECH_LEVEL			
Independent variables	Model_5	Model_6	Model_7
PAI_M	4.959***	5.010***	4.622***
	(0.526)	(0.519)	(0.530)
PII	0.441		0.327
	(0.706)		(0.687)
PFS_1	2.540**	2.706***	
	(1.018)	(0.982)	
PFS_2			6.214***
			(1.628)
Constant	86.038***	85.235***	83.002***
	(1.095)	(1.047)	(1.506)
Observations	264	264	264
Adjusted R2	0.293	0.295	0.314
F Statistic	37.31***	55.90***	41.19**

*p〈0.1; **p〈0.05; ***p〈0.01
Standard Error in parenthesis

〈표 11〉에서 각 독립변수별 측정결과를 살펴보면 다음과 같다. 우선 PAI_M의 회귀 계수 값은 앞서 주요 5개국 결과 유사한 수준임을 확인할 수 있다. 다음으로 피인용 횟수에 기반한 독립변수인 PII의 경우 양의 값을 가지나 산업기술수준에 미치는 영향이 통계적으로 유의미하지 않았다. 이로 인해 한중일 3국의 경우 PII 이외 독립변수들의 값이 통제된 경우 피인용이 많고 적음에 의해 산업기술수준이 좌우되지 않는 경향을 확인할 수 있다. 이는 앞서 미국, 유럽이 포함된 주요 5개국의 경우와 특이하게 다른 양상을 보이는 부분이다. 마지막으로, 독립변수 PFS_1, PFS_2의 경우 둘 다 산업기술수준에 정(+)의 영향력을 미치는데 주요 5개국 분석 결과와 유사한 양상으로 PFS_2가 PFS_1 대비 회귀식의 계수가 높은 값을 가짐이 확인된다.

다음으로 한중일 3국의 산업기술수준을 최소자승 회귀법, 인공신경망, 랜덤포레스트 방법을 사용하여 예측하였고 각 방법의 성능을 비교한 결과는

〈표 12〉와 같다. 분석 방법은 앞서 주요 5개국 분석 방법과 동일(산업기술수준을 상 · 중 · 하위로 구분하여 데이터를 층화 추출하고 10-fold cross-validation)하게 실시하였다.

〈표 12〉 산업기술수준 예측 성능 비교(한중일 3국 한정)

Method	performance measure			Model
	RMSE	MAE	adj. R^2	
OLS Regression	5.59	4.31	0.20	Model 7
Artificial Neural Networks	5.46	4.17	0.23	
Random Forest	5.44	3.92	0.24	
XGboost	5.35	3.91	0.26	

4가지 방법 모두 Model 5~7 중 Model 7에서 가장 높은 성능이 측정되었으며 XGboost 방법이 가장 성능이 우수한 것으로 확인된다. 모든 방법에서 초모수(hyper parameter)는 grid search로 결정하였으며, 최고의 성능을 나타낸 초모수 값은 다음과 같다. 인공신경망의 경우 1개의 은닉층에 110개 노드가 사용되었고 logistic 함수가 활성화 함수(activation function)로 사용되었으며 학습 속도(learning rate)는 0.05였다. 랜덤포레스트의 경우 트리 수 600개, 최대 깊이 12, 최대 특징(feature)은 2개였다. XGboost의 경우 반복횟수 115개, 최대 깊이 2, 학습률 0.07, sub-sample rate은 0.3, gamma는 0이었다.

Ⅴ. 토론 및 결론

본 연구에서는 2개의 데이터 세트(주요 5개국 및 한중일 3개국)를 활용하여 분석을 실시하여 양측 결과를 비교해 보았는데 특허지표 측면에서 이를 종합해 보면 다음과 같다. 특허 점유율(PAI_M)은 2개 데이터 세트 내 다수 모델에서 산업기술수준에 미치는 영향력이 가장 높게 측정되었다. 패밀리 특허수(PFS_1), 패밀리 국가수(PFS_2) 또한 영향력이 높았고 특히 후자가 전자

보다 높은 영향력을 미쳤다. 특이한 연구결과로는 특허영향력(PII)과 특허집중도(PCI)는 5개국 측정 결과에서만 산업기술수준에 통계적으로 유의미한 영향이 있었고 한중일 3국만을 대상으로 하는 경우 그렇지 않았다. 기존 오종학 · 나관식(2018)[36]의 연구에서도 한국과 중국만을 대상으로 한 경우 PII가 기술수준에 미치는 영향이 유의미하지 않음을 확인하였는데 이와 상응하는 결과로 보인다. 그러므로 많은 연구 또는 분석지침에서 기술력 측정 시 피인용 정도에 기반한 PII 지표를 활용하나, 한중일 3국의 경우 피인용 횟수에 기반한 기술적 우위의 비교는 주의를 기울여야 할 것으로 판단된다. 한편, 이러한 현상이 일어나는 원인으로 피인용 특허 건수 중 자체 인용(self citation)건수가 많은 비중을 차지하는 경우를 의심해 볼 수 있는데 이에 대한 분석은 후속 연구를 통해 확인이 필요할 것이다.

이하에서 기존 연구대비 본 연구의 의의, 활용방안 및 개선 사항을 살펴보면 다음과 같다. 먼저 특허정보를 활용하여 기술수준을 예측하는 방법론을 개발하는 기존 연구들이 정보통신 등 일부 분야에 대해 선형 회귀분석 방법을 사용하였으나, 본 연구는 분석데이터로 다양한 산업기술분야를 포함하였고 다양한 특허지표를 사용하였으며 기계학습을 활용함으로써, 산업기술분야 전반에 사용 가능하며 정확도가 개선된 기술수준 예측 방법론을 개발하였다는 데 의의가 있을 것이다.

본 연구의 실질적 활용 방안을 살펴보면 다음과 같다. 우선 다양한 기술분야가 포함된 대규모 기술수준조사를 실시하는 경우 일부 기술분야에서는 기술수준에 대해 전문가별 편차가 큰 경우가 발생하는데 이때 본 연구의 방법론을 사용하여 보다 객관적인 기술수준의 확인이 가능할 것이다. 또한 다양한 산업기술 분야에서 국가R&D 기획 시 소규모의 기술수준조사가 필요하나 기획에 소요되는 예산과 시간이 한정되어 전문가 설문을 위해 고비용 장시간의 투입은 현실적으로 불가능하므로 본 연구의 방법론[37]을 활용하여 기술

36) 오종학 · 나관식, “특허지표가 선진국과 개발도상국의 기술수준에 미치는 영향: 정보보안기술을 중심으로”, 『한국창업학회지』, 제13권 제3호(2018), 75-93면.

37) 실무적으로 기계학습 방법의 기술수준 측정이 어려운 경우도 본 연구에서 도출한 회

수준의 신속한[38] 도출과 비교가 가능할 것이다. 한편 본 예측 모형에 특허데이터 분석기간을 달리하여 투입하는 경우 미래 기술수준의 예측도 가능할 것으로 보이는데,[39] 이 부분은 수년 이후를 예측했던 결과와 실제 측정값을 비교하는 연구가 추가로 필요할 것이다.

다음으로 본 연구의 주된 한계는 첫째, 본 방법론은 해당 기술분야 내 특허 건수, 인용건수, 패밀리 특허 수 등이 충분하지 않으면 특허정보를 활용한 기술수준 예측이 불가능하다는 점이다. 본 연구와 같이 세부 기술분야의 기술수준을 예측하는 경우 특허정보(특허 건수, 피인용 건수 등)의 부족이 더욱 심하게 된다. 반면, 상위 기술을 분석하면 특허정보를 충분히 확보할 수 있으나, 세부기술이 상위 기술에서 차지하는 가중치에 따라 기술수준이 변동되는 단점이 존재한다. 본 연구는 많은 데이터를 확보하는 동시에 가중치를 고려하지 않아도 되는 세부 기술분야를 연구대상으로 삼았으나, 특허수가 많지 않은 다수의 세부 기술분야는 분석대상에서 제외할 수밖에 없어 최초에 예상했던 충분한 수준의 데이터량을 확보하지 못하게 되는 문제가 있었다.

둘째, 산업기술수준조사를 포함한 다수의 기술수준조사가 해당 기술분야 내 다수 전문가의 응답을 평균하여 측정하는데 이때 평가자의 수가 적거나 평가자 간 의견의 일치 정도인 동의도[40]가 낮은 경우는 종속변수로 활용한 기술수준조사 측정값 자체의 신뢰성이 부족할 수 있을 것이다. 한편 동의도가 낮거나 응답 수가 적은 기술분야의 데이터는 제외하고 모형을 구성하는 경우 예측의 정확도가 높아지는지를 확인하는 것도 향후 후속 연구로서 의

귀식에 기술분야별 특허지표 측정결과를 투입하면 신속하고 대략적인 기술수준의 측정이 가능.

38) 설문조사에는 통상 6개월(특허분석 및 설문 준비 3개월, 전문가 설문실시 2개월, 설문조사 결과 정리 1개월) 이상이 소요되는데, 본 연구방법을 사용하면 설문실시 및 정리 기간인 3개월 내외의 단축이 예상됨.

39) 본 연구의 모형은 2009~2018년 특허데이터를 투입하여 2021년도 기술수준을 측정하므로, 2023년 기술수준을 예측하고자 할 경우 본 모형에 2011~2020년 특허데이터를 투입하면 된다.

40) 동의도 = 1 - (관찰된 분산 / 무작위 분산).

미가 있을 것이다.

셋째, 특허데이터 추출기간을 기존 선행연구와 같이 최근 10년으로 한정하였으나 이러한 구간을 달리하는 경우(예를 들어 최근 7년, 최근 13년) 예측의 설명력이 어떻게 변화하는가를 탐색하는 연구도 필요할 것이다.

마지막으로 기술수준 예측 모델의 설명력과 성능을 보다 개선하기 위하여 본 연구에서 사용한 지표 이외 추가적인 특허지표를 사용하거나 특허 외에 논문 정보 등을 추가로 활용할 경우 설명력과 예측 성능의 향상이 기대되며, 기계학습 방법 또한 인공신경망 또는 렌덤포레스트, XGboost 이외의 다양한 알고리즘과 XAI(eXplainable AI) 방법론을 사용한 후속 연구가 가능할 것이다.

마무리하면, 본 연구는 객관적 데이터 기반의 기술수준 측정 방법론 도출에 기여하고자 수행되었으며, 본 연구결과가 다양한 기술분야에서 기존 전문가 설문조사 방식의 기술수준조사를 보완 또는 대체하는 측면에서 활용될 수 있을 것이다.

참고문헌

〈학술지(국내 및 동양)〉

구기관 외 3인, "국내외 신재생에너지 기술 경쟁력 분석: 태양광 · 연료전지를 중심으로", 『신재생에너지』, 제8권 제3호(2012).

오종학 · 나관식, "특허지표가 선진국과 개발도상국의 기술수준에 미치는 영향: 정보보안기술을 중심으로", 『한국창업학회지』, 제13권 제3호(2018).

유소진 · 이재승. "스트레처블 태양광전지 분야 특허기반 기술경쟁력 연구", 『한국태양광발전학회지』, 제7권 제1호(2021).

하수진 외 2인, "태양전지 분야 주요 5개국의 연구논문 동향 및 기술수준 조사 · 분석", 『한국기후변화학회지』, 제12권 제1호(2021).

Choi, Ji Weon et al., "Technology trends and patenting prospects of medicinal plants in Korea", *Korean Journal of Medicinal Crop Science*, Vol.27, No.2 (2019).

이동헌 외 2인, "국내외 기술수준조사 및 기술수준평가 연구 동향 분석", *ICROS*, 제20권 제1호(2014).

〈학술지(서양)〉

Cho, I. & Park, M., "Technological-level evaluation using patent statistics: model and application in mobile communications", *Cluster Computing*, Vol.18, No.1(2015).

Han, Yuri et al., "Analysis of global competitiveness of engineering modeling and simulation technology for next-manufacturing innovation: Using quantitative analysis of patents and papers", *ICIC Express Letters*, Vol.9, No.4(2018).

Hartigan, J. A. & Wong, M. A., "Algorithm AS 136: A K-Means Clustering Algorithm", *Journal of the Royal Statistical Society, Series C*, Vol.28, No.1(1979).

Prajwala, T. R., "A comparative study on decision tree and random forest using R tool", *International journal of advanced research in computer and communication engineering*, Vol.4, No.1(2015).

Rodriguez-Galiano, V. F et al., "An assessment of the effectiveness of a random forest classifier for land-cover classification", *ISPRS journal of photogrammetry and remote sensing*, Vol.67(2012).

Uysal, M. & Roubi, S. E., "Artificial neural networks versus multiple regression in

tourism demand analysis", *Journal of Travel Research*, Vol.38(1999).

Wu, L. et al., "Comparing nanotechnology landscapes in the US and China: a patent analysis perspective", *Journal of nanoparticle research*, Vol.21, No.8(2019).

〈연구보고서〉

김한울 외 2인, "2020년도 연구개발활동조사 결과", 과학기술정보통신부, 2020.

서규원, "특허지표를 활용한 기술수준평가 연구방법론의 개발 및 적용", 한국과학기술기획평가원, 2011.

선동주 외 6인, "특허성과 지표 활용 가이드라인", 특허청, 2014.

OECD, "OECD Patent Statistics Manual", OECD, 2009.

〈인터넷 자료〉

한국과학기술기획평가원, "분야별 기술수준평가", 과학기술정책지원서비스, 〈https://www.k2base.re.kr/techLevelEval/list.do〉, 검색일: 2022년 4월 27일.

Banerjee, P., "A Guide on XGBoost hyperparameters tuning", Kaggle, 〈https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook〉, 검색일: 2022년 5월 27일.

〈기타 자료〉

Chen, Tianqi & Guestrin, Carlos, "Xgboost: A scalable tree boosting system", Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.

Jo, Yong-Gon & Jo, Geun-Tae, "Formulating R&D strategy for core technologies in biotechnology using the delphi and the AHP", In Proceedings of the Korean Operations and Management Science Society Conference, The Korean Operations Research and Management Science Society, 2004.

Ibrahim, Z., & Rusli, D., "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression", In 21st Annual SAS Malaysia Forum, 2007.

Zekic-Susac, M. et al., "Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models", In 26th International Conference on Information Technology Interfaces IEEE, 2004.

〈부록 1〉 산업기술수준조사 설문 항목 및 내용

□ ('21년도) 산업기술수준조사 설문 항목 및 측정 방법

구 분	조사항목	문항정의	평가척도	측정방법	
기술적 중요도	중요도	상위기술(대분류)이 적절한 기능을 수행하게 하는 데 있어 해당기술의 중요한 정도	비율척도	(100점 만점 기준) 해당기술 중요도 점수	
	시급성	적정기술 구현이 필요한 시기	명목척도	① 3년 이내 ② 5년 이내 ③ 5년 이후	
	파급효과	해당 기술이 타 요소기술에 미치는 영향력 정도(응용범위)	등간척도	① 영향 없음 …… ⑤ 영향 미침	
기술수준 변화양상	기술변화 속도	해당 기술력의 과거 대비 변화 양상	등간척도	① 아주 느림 …… ⑤ 아주 빠름	
기술수준 측정지표	최고 기술국	해당 기술에 대한 최고 기술국 파악	명목척도	① 한국 ② 미국 ③ 중국 ④ 일본 ⑤ 유럽	
	기술적 상대수준	최고 기술국 대비 주요국의 상대적인 기술수준	비율척도	선도수준(90~99%)…추격수준(80~89%)…후발수준(70~79%)…낙후수준(70% 미만)	
	기술수준 격차	현재 최고기술 보유국 수준에 도달하는데 소요되는 시간격차	비율척도	( )개월	
	기술수준 비교	최고 기술국 대비 한국의 기술수준단계	명목척도	① 개념 단계 ② 기술 정립 단계 ③ 기술 개발 단계 ④ 초기 상용화 단계	⑤ 기술진화 및 시장 성장 단계 ⑥ 기술완성 및 시장 성숙 단계
	기술격차 발생원인	해당분야(대분류) 기술격차가 발생한 원인	명목척도	① R&D투자 부족 ② R&D인프라 부족 ③ 인력부족	④ 국내 공동연구 부족 ⑤ 국제 공동연구 부족 ⑥ 법/제도/규제개선 미흡 ⑦ 시장부족
	기술격차 해소방안	해당분야(대분류) 기술격차 해소를 위한 방안	명목척도	① R&D 투자 확대 ② R&D 인프라 확충 ③ 인력양성 강화	④ 국내 공동연구 강화 ⑤ 국제 공동연구 강화 ⑥ 법/제도/규제개선 ⑦ 시장 활성화
기타	응답 확신도	평가 내용에 대한 본인 확신도	등간척도	① 아주 낮음 …… ⑤ 아주 높음	

□ 산업기술수준조사 설문지(기술수준 질의 분야 발췌)

○ 최고기술국의 기술수준을 100%로 가정할 때, 나머지 국가들의 상대적인 기술수준을 평가해 주세요.
(아래 상대수준 평가기준을 참고하시어, 1%~99% 범위 내로 작성해 주세요)

상대수준 평가기준
선도수준(90%~99%) : 기술분야를 선도하는 수준
추격수준(80%~89%) : 선진기술의 모방/개량이 가능한 수준
후발수준(70%~79%) : 선진기술의 도입/적용이 가능한 수준
낙후수준(70% 미만) : 연구개발 능력이 취약한 수준

[최고기술 보유국 100%로 자동입력]

대분류	중분류	최고기술국 대비 상대적 기술수준				
		① 한국	② 미국	③ 중국	④ 일본	⑤ 유럽
자율주행 핵심부품	주행환경 인지기술 (19년 한국 : XX.X%)	____%	____%	____%	____%	____%
자율주행 핵심부품	자율주행 통합제어 (19년 한국 : XX.X%)	____%	____%	____%	____%	____%
커넥티비티 및 서비스	운전자 모니터링 및 제어권 전환 (19년 한국 : XX.X%)	____%	____%	____%	____%	____%

〈부록 2〉 데이터 추출 및 생성 과정

〈특허정보(Raw Data)〉

분야	대분류	중분류	출원번호	출원년도	등록여부	대표출원인 국적	출원인	피인용 회수	패밀리 특허수	패밀리 국가수
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-221805	2018	N	JP	Toyota Jidosha Kabushiki Kaisha	0	3	3
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-215113	2018	N	KR	HYUNDAI MOTOR COMPANY | KIA MOTORS CORPORATION | AJOU UNIVERSITY INDUSTRY-ACADEMIC COOPERATION FOUNDATION	0	3	3
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-213827	2018	N	US	Ouster, Inc.	0	6	2
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-212066	2018	Y	EU	VOLKSWAGEN AKTIENGESELLSCHAFT	0	6	5
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-204329	2018	N	KR	Hyundai Motor Company | Kia Motors Corporation	0	4	4
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-197276	2018	N	EU	THALES	0	5	5
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-185980	2018	N	EU	Continental Automotive Systems, Inc.	0	2	2
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-184866	2018	N	US	Peloton Technology, Inc.	0	1	1
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-183248	2018	Y	JP	DENSO CORPORATION	1	13	3
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-177734	2018	Y	US	MAGNA ELECTRONICS INC.	3	32	5
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-176529	2018	Y	US	Luminar Technologies, Inc.	3	17	2
자율주행차	자율주행 핵심부품	주행환경 인지기술	16-176607	2018	Y	US	Luminar Technologies, Inc.	0	17	2
:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:

〈설명변수(특허지표) 추출(파이썬 코딩)〉

분야	대분류	중분류	NATION	출원건수	PAI	PAI_M	PCI	PII	PFS_1	PFS_2
자율주행차	자율주행핵심부품	주행환경 인지기술	CN	40	0.03	0.14	0.54	0.29	0.35	0.94
자율주행차	자율주행핵심부품	주행환경 인지기술	EU	264	0.18	0.91	1.32	0.78	0.54	1.30
자율주행차	자율주행핵심부품	주행환경 인지기술	JP	363	0.25	1.26	1.59	0.58	0.46	1.04
자율주행차	자율주행핵심부품	주행환경 인지기술	KR	105	0.07	0.36	0.63	0.56	0.39	0.93
자율주행차	자율주행핵심부품	주행환경 인지기술	US	671	0.47	2.33	0.86	1.44	1.62	0.87
자율주행차	자율주행핵심부품	자율주행 통합제어	CN	69	0.02	0.12	0.49	0.24	0.30	1.15
자율주행차	자율주행핵심부품	자율주행 통합제어	EU	347	0.13	0.63	0.90	0.57	0.34	1.15
자율주행차	자율주행핵심부품	자율주행 통합제어	JP	501	0.18	0.90	1.15	0.69	0.32	1.14
자율주행차	자율주행핵심부품	자율주행 통합제어	KR	272	0.10	0.49	0.85	0.42	0.26	1.06
자율주행차	자율주행핵심부품	자율주행 통합제어	US	1582	0.57	2.85	1.06	1.29	1.47	0.91
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	CN	11	0.02	0.09	0.35	0.48	2.47	1.18
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	EU	77	0.12	0.62	0.90	1.10	0.68	1.22
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	JP	227	0.37	1.83	2.32	0.72	1.23	1.24
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	KR	60	0.10	0.48	0.84	0.59	0.69	1.07
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	US	246	0.40	1.98	0.74	1.30	0.93	0.74
:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:

〈목적변수(전문가 기술수준 설문조사 결과) 추가〉

분야	대분류	중분류	NATION	출원건수	설명변수						목적변수
					PAI	PAI_M	PCI	PII	PFS_1	PFS_2	TECH_LEVEL
자율주행차	자율주행핵심부품	주행환경 인지기술	CN	40	0.03	0.14	0.54	0.29	0.35	0.94	77.8
자율주행차	자율주행핵심부품	주행환경 인지기술	EU	264	0.18	0.91	1.32	0.78	0.54	1.30	88.2
자율주행차	자율주행핵심부품	주행환경 인지기술	JP	363	0.25	1.26	1.59	0.58	0.46	1.04	82.9
자율주행차	자율주행핵심부품	주행환경 인지기술	KR	105	0.07	0.36	0.63	0.56	0.39	0.93	81.5
자율주행차	자율주행핵심부품	주행환경 인지기술	US	671	0.47	2.33	0.86	1.44	1.62	0.87	100.0
자율주행차	자율주행핵심부품	자율주행 통합제어	CN	69	0.02	0.12	0.49	0.24	0.30	1.15	75.8
자율주행차	자율주행핵심부품	자율주행 통합제어	EU	347	0.13	0.63	0.90	0.57	0.34	1.15	88.4
자율주행차	자율주행핵심부품	자율주행 통합제어	JP	501	0.18	0.90	1.15	0.69	0.32	1.14	83.4
자율주행차	자율주행핵심부품	자율주행 통합제어	KR	272	0.10	0.49	0.85	0.42	0.26	1.06	80.4
자율주행차	자율주행핵심부품	자율주행 통합제어	US	1582	0.57	2.85	1.06	1.29	1.47	0.91	100.0
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	CN	11	0.02	0.09	0.35	0.48	2.47	1.18	74.9
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	EU	77	0.12	0.62	0.90	1.10	0.68	1.22	91.5
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	JP	227	0.37	1.83	2.32	0.72	1.23	1.24	86.6
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	KR	60	0.10	0.48	0.84	0.59	0.69	1.07	82.6
자율주행차	자율주행핵심부품	운전자 모니터링 및 제어권 전환	US	246	0.40	1.98	0.74	1.30	0.93	0.74	100.0
:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:

Research on Industrial Technology Level Evaluation Method using Patent Information and Machine Learning

Lee, Cheolju; Cha, Hyunjin; Lee, Jungwoo; Ko, Byungchul & Han, Jongseok

Accurate measurement of technology level is required as basic information for establishing technology policy or strategy. However, technology level survey from experts may lack objectivity and consume considerable time and costs. Therefore, this study was conducted to derive a methodology to measure the level of technology objectively and easily by using patent information. We attempted to measure the industrial technology level from patent information by linking survey results with the patent competitiveness data of 5 major countries in 80 industrial technology fields. Using various patent indices as independent variables and the technology level survey result as a dependent variable, linear regression analysis was performed, identifying the influence of each index on the determination of technology level. Next, the technology level was predicted using artificial neural networks, random forest, and XGboost, confirming the better performance of machine learning than linear regression method. This study is meaningful in that it developed a technology level evaluation methodology with improved accuracy by using various patent indices and machine learning. And we expect our research would be used as a tool to supplement the expert survey method.

Keyword

technology level, patent information, patent index, machine learning, artificial neural networks, random forest, XGboost