# 특허 패밀리 수를 고려한 머신러닝 기반의 특허 가치 평가 방안

이형진 [1], 유헌창 [2]
[1] 고려대학교 SW/AI 융합대학원 인공지능융합학과 석사과정
[2] 고려대학교 정보대학 컴퓨터학과 교수

hjlee0117@gmail.com, yuhc@korea.ac.kr

# A Study on Machine Learning-Based Method for Patent Valuation Considering the Number of Patent Families

Hyeongjin Lee[1], Heonchang Yu[2]
[1]Dept. of AI Convergence, Graduate School of SW/AI Convergence, Korea University
[2]Dept. of Computer Science & Engineering, Korea university

**요 약**

특허의 가치를 평가하기 위해서는 특허 데이터에 포함된 다양한 지표가 활용될 수 있으며, 최근 다양한 지표를 머신 러닝 기법으로 분석하여 특허의 가치를 평가하는 연구가 증가하고 있다. 특허의 가치를 올바르게 평가하기 위해서는 여러 지표 중에서 어떤 지표가 특허의 가치에 크게 기여하는지 판단할 수 있어야 하며, 이에 따라 지표별로 적절한 가중치를 설정할 수 있어야 한다. 제안된 방법은 회귀 모델 기반으로 다양한 지표에 가중치를 적용하여 특허 피인용수를 예측하였으며, 특허 패밀리 수에 적용되는 가중치를 변경하면서 특허 패밀리 수가 특허의 가치에 미치는 영향을 검증하였고, 특허 가치 평가 과정에서 특허 패밀리 수의 중요성에 대해 확인하였다.

## 1. 서론

특허는 발명의 공개를 대가로 주어지는 독점·배타적인 권리이며, 특허 데이터는 기술에 관한 다양한 정보를 포함하고 있다. 최근에는 대기업 뿐만 아니라 중견, 중소 및 스타트업에 이르기까지 전략적인 특허 포트폴리오를 구축하기 위한 다양한 노력을 수행하고 있다.

특히 기업들은 특허권을 보유함으로써 기술의 독점적인 사용을 법적으로 보장받고 있으며, 경쟁사의 시장 진입을 저해하는 핵심 요소로 특허를 활용하고 있다. 이와 더불어 특허의 가치를 평가하기 위한 다양한 연구들이 진행되고 있으며, 특허권이 공개를 대가로 주어지는 권리라는 점에서 공개된 특허 데이터를 머신러닝 기반으로 분석하여 특허의 가치를 평가하는 연구가 활발하게 이어지고 있다. 기존의 연구들은 '피인용 수'가 특허의 가치를 내포하는 것을 확인하였고 (Dutta & Weiss, 1997) [1], 다양한 특허 지표들에 기초하여 '피인용 수'를 예측하는 방법을 제시하였다 (Nathan Falk et al, 2017) [2]. 다만, 기존의 연구들에서 특허의 가치를 판단하기 위한 지표로서 '특허 패밀리 수'에 대한 중요성을 증명하는 연구는 부족한 실정이다. 이에 따라 본 연구에서는 다양한 특허 지표들 중에서 '특허 패밀리 수'의 중요성에 대해 살펴보고 머신러닝 기법에 의해 그 중요성을 증명하였다.

구체적으로, 회귀 모델을 활용하여 특허 지표 각각을 독립 변수로 설정하였고, 각 독립 변수가 모델의 예측에 기여하는 정도를 수치적으로 조절하기 위해 가중치를 도입하였다. 이때, 특허 패밀리 수를 제외한 나머지 독립 변수에는 동일한 가중치를 부여하였고, 특허 패밀리 수의 가중치를 변화시키면서 피인용수의 예측 정확도에 특허 패밀리 수가 미치는 영향을 성능 평가 지표로 확인하였다.

## 2. 관련 연구

### 2.1 기존의 특허 가치 평가 이론

특허 가치를 평가하기 위한 주요 이론으로는 기술가치 평가에 관한 수익접근법, 비용접근법 및 시장접근법이 활용된다[3]. 여기에서 수익접근법은 평가의 대상이 되는 특허 기술이 특허의 존속기간 동안 발생시킬 경제적 가치를 예측한 방법이고, 비용접근법은 특허 기술을 개발하거나 구입하기 위해 소비되는 비용을 특허의 가치로 산정하는 방법이며, 시장접근법은 평가 대상 특허와 유사한 특허가 관련 분야에서 거래된 이력을 기반으로 특허의 가치를 유추하여 산

정하는 방법이다.

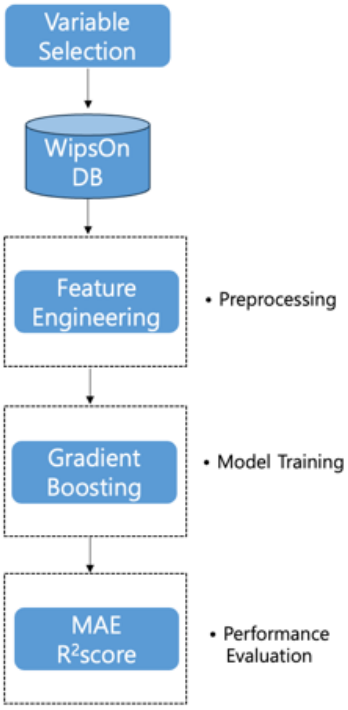
2.2 특허 가치 평가에 관한 선행 연구

관련 선행 연구는 특허의 가치를 판단하기 위한 평가 지표에 관한 연구와, 평가 지표에 기초하여 특허의 가치를 판단하기 위한 모델에 관한 연구로 나눌 수 있다. 먼저, 평가 지표와 관련하여 특허의 가치는 다양한 평가 지표에 의해 평가될 수 있으며, 구체적으로 평가 지표는 발명자 수, 자기인용, 청구항 수 및 패밀리 특허 수를 포함할 수 있다(Lee et al, 2006) [4]. 또한, 선행 연구에서는 평가 지표에 가중치를 설정하여 새로운 품질 지표를 생성하거나(Wu et al, 2016) [5], 특허 문헌의 첫 페이지에 포함된 정보를 평가 지표로 정하여 특허의 가치를 평가하기도 하였다(Lin et al, 2007)[6]. 다음으로, 특허의 가치를 판단하기 위한 모델과 관련하여 선행 연구에서는 새로운 특허 품질 지표를 SVM 과 같은 머신러닝 모델에 입력하여 특허 가치를 평가하거나(Wu et al, 2016)[4], MLP 알고리즘에 기초하여 특허의 품질을 예측하는 연구도 수행되었다 (Erdogan et al, 2022)[7].

## 3. 머신러닝 기반 특허 가치 평가 기법 설계

3.1 특허 가치 평가 모델링

본 연구는 아래 Fig. 3.1 과 같이 변수를 선택하고, 데이터베이스에서 특허 데이터를 수집하여 전처리한 후 변수의 가중치를 달리하면서 모델을 학습시켜 모델의 성능을 판단하였다.




(그림 1) 머신러닝 파이프라인

독립변수와 종속변수를 선택하고, WipsOn 특허 데이터베이스(www.wipson.co.kr)에서 엑셀 형태로 특허 데이터를 수집하였다. 이후 수집된 특허 데이터에서 독립변수의 기여도를 다르게 설정하는 feature engineering 을 수행하고, 그래디언트 부스팅 알고리즘을 활용하여 XGboost 로 모델을 학습시켰다. XGboost 는 그래디언트 부스팅 학습 과정을 병렬로 처리할 수 있어서 처리 속도가 빠르고, DNN(Deep Neural Network) 보다 수치형 특성이나 범주형 특성이 명확하게 정의된 정형 데이터에서 매우 효과적이므로 특허 데이터에 적합하여 선택하였다. 모델의 성능 평가는 회귀 분석에서 모델의 성능을 평가하는 주요 지표인 MAE(Mean Absolute Error)와 $R^2$score 를 활용하였다.

3.2 변수 선정

3.2.1 독립변수

특허 데이터에는 다양한 범주형 데이터와 수치형 데이터가 포함되며, 독립변수로는 수치형 데이터 중 기존 연구에서 중요한 가치를 내포하고 있다고 확인된 변수를 위주로 선정하였다. 예를 들어 Lanjouw & Schankerman(2004)은 독립 변수로 청구항 수, 피인용 수, 인용 문헌 수 및 특허 패밀리 수에 가중치를 부여하고, 특허의 품질을 예측하였다.[8]

그러나, 기존의 연구들에서는 '특허 패밀리 수'가 특허의 가치를 나타내는 가장 중요한 변수임을 설명하지 못하였다. 여기에서 '특허 패밀리 수'는 발명을 국제적으로 보호하기 위하여 동일한 발명에 대해 여러 국가에서 출원된 일련의 관련 특허의 수를 의미한다.

3.2.2 종속변수

종속변수는 시간의 흐름에 따라 누적될 수 있는 수치형 데이터 중에서 기존 연구에서 특허의 가치를 내포하고 있는 것으로 확인된 피인용 수로 선정하였다. 구체적으로, Lin et al.(2007)은 바이오 분야에서 특허 명세서의 첫 페이지에 기재된 정보에 기초하여 피인용 수를 예측하였고[6], Nathan et al.(2017)은 통계 모델을 구축하고 통계 모델에 기초하여 예측 인용수를 도출하였다[2].

여기에서 피인용 수는 특허 심사 과정에서 심사관이 다른 특허의 심사에 활용하기 위해 인용한 횟수를 의미하며, 특정 기술 분야에서 피인용 수가 높다는 것은 해당 기술 분야의 제반 기술 또는 핵심 기술로 판단될 가능성이 높으므로, 피인용 수를 종속변수로 선정하였다.

## 4. 특허 데이터를 이용한 모델 학습

본 연구에서는 그림 2 에서와 같이 특허 패밀리 수 이외의 독립변수에 대한 가중치를 고정하고, 특허 패밀리 수의 가중치를 0.1 에서 0.9 까지 높여가며 종속변수를 예측하였다. 이를 위해 X_train 데이터를 스케일링하기 위한 스케일 펙터를 계산하고, 스케일 펙터를 X_train 데이터에 적용하여 X_train_scaled 로 학습을 진행하였다.

```
user_defined_importances = {
    '청구항 수': 0.7,
    '독립항 수': 0.7,
    '출원인 수': 0.7,
    '발명자 수': 0.7,
    '인용 문헌 수(B1)': 0.7,
    '개별도면 수': 0.7,
    '비 특허 참고문헌 수(B1)': 0.7,
    'WIPS패밀리 문헌 수(출원기준)': 0.1,
    'EPO패밀리 문헌 수(출원기준)': 0.1
}
```

(그림 2) 독립변수 가중치

여기에서 독립변수에 대해 설명하면, 청구항 수는 특허의 권리범위를 결정하는 청구항의 총 수를 의미하고, 독립항 수는 다른 청구항을 인용하는 형식을 취하지 않는 청구항의 수를 의미하고, 출원인 수는 명세서에 기재된 출원인의 인원수를 의미하고, 발명자 수는 발명에 기여한 인원수를 의미한다. 또한, 인용 문헌 수(B1)는 본 특허에서 다른 문헌을 인용한 개수를 의미하고, 개별도면 수는 본 특허에서 작성된 도면의 개수를 의미하고, 비 특허 참고문헌 수(B1)는 본 특허에서 참고한 문헌의 개수를 의미하고, WIPS 패밀리 문헌 수(출원기준)는 특허 출원 당시 기준으로 우선권 기초가 일부 중복되는 특허 패밀리 수를 의미하고, EPO 패밀리 문헌 수(출원기준)는 특허 출원 당시 기준으로 우선권 기초가 전부 중복되는 특허 패밀리 수를 의미한다.

가중치는 각 피처(독립변수)가 모델의 예측에 기여하는 정도를 수치적으로 나타낸 것을 의미하며, 0 에 가까울수록 가중치가 작고, 1 에 가까울수록 가중치가 큰 것을 의미한다. 특허 패밀리 수의 가중치를 변경하는 과정에서 나머지 독립변수에 대한 가중치를 0.7 로 고정하였는데, 여기에서의 0.7 은 모델의 $R^2$score 가 0 을 초과하기 위한 기준값이며, 모델이 데이터의 분산을 설명할 수 있는 최소값으로서 실험적으로 결정되었다.

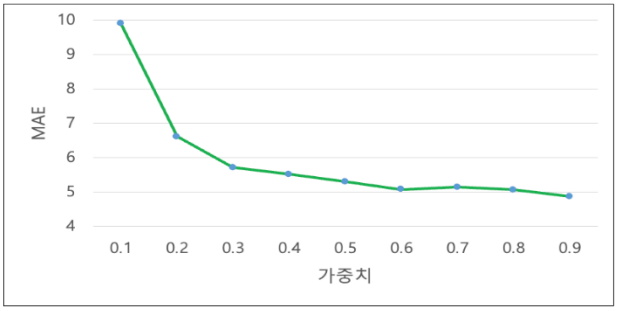
## 5. 머신러닝과 특허데이터를 이용한 모델 검증

본 연구에서는 분야를 한정하여 머신러닝으로 배터리를 진단하는 기술 분야에 대해 (그림 3) 의 특허 검색식으로 검색된 55,722 건을 학습데이터로 활용하였다.

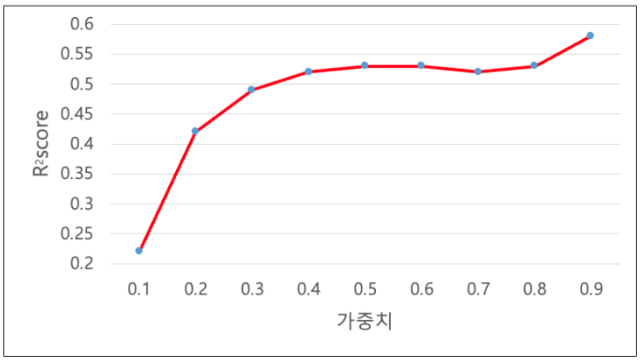
(머신러닝 or 인공지능 or ai or "machine learning") and (배터리 or battery) and (진단 or diagnosis)

(그림 3) 특허 검색식

연구 방법에서 제시한 독립변수들에 가중치를 부여하고 XGboost 에 의해 회귀분석을 수행한 결과, 아래의 (그림 4) 및 (그림 5)와 같이 분석되었다.




(그림 4) MAE 결과




(그림 5) $R^2$score 결과

즉, 특허 패밀리 수 이외의 독립변수에 대한 가중치를 고정하고, 특허 패밀리 수의 가중치를 0.1 에서 0.9 까지 높여가며 종속변수를 예측한 결과, 특허 패밀리 수의 가중치를 높일수록 MAE 는 낮아지고, $R^2$score 는 높아지는 경향을 보였다.

<표 1> MAE 및 $R^2$score 결과

| 패밀리 수 가중치 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| MAE | 9.9 | 6.61 | 5.7 | 5.51 | 5.29 | 5.07 | 5.13 | 5.06 | 4.86 |
| $R^2$score | 0.22 | 0.42 | 0.49 | 0.52 | 0.53 | 0.53 | 0.52 | 0.53 | 0.58 |

표 1 을 참조하면, 특허 패밀리 수 이외의 독립변수의 가중치를 0.7 로 고정하고, 특허 패밀리 수의 가중치를 0.1 로 설정한 결과, MAE 값은 9.9 로 높게 나타났고, $R^2$score 값은 0.22 로 낮게 나타났다. 반면, 특허 패밀리 수 이외의 독립변수의 가중치를 0.7 로 고정하고, 특허 패밀리 수의 가중치를 0.9 로 설정한 결과, MAE 값은 4.86 으로 상대적으로 낮게 나타났고, $R^2$score 값은 0.58 로 상대적으로 높게 나타났다.

여기에서 MAE 값은 모델의 예측이 실제 값에서 얼마나 떨어져 있는지를 나타내는 지표로, 예측 값과 실제 값 사이의 절대 오차의 평균으로 계산되어 상대적으로 낮은 값을 가지는 모델의 성능이 더 우수하다. 또한, $R^2$score 값은 모델이 데이터의 변동성을 얼마나 잘 설명하고 있는지를 측정하며, 0 에서 1 사이의 값 중에서 1 에 가까울수록 모델이 데이터를 더 잘 설명하고 있다고 해석할 수 있다.

일반적으로 특허 패밀리 출원은 국내에 특허 출원을 먼저 진행하고, 국내 특허 출원을 우선권 기초 출원으로 하여 해외에 패밀리 출원을 진행한다. 이 과정에서 패밀리 출원을 진행하기 위해서는 해외 특허청에 지불해야 하는 관납료와 해외 특허대리인에 지불해야 하는 수수료가 국내 출원과 비교하여 적게는 수배에서 많게는 수십배가 요구된다. 따라서, 해외에 패밀리 출원을 진행하고자 하는 기업이나 개인은 신중하게 특허의 가치를 판단할 수 밖에 없다. 이에 따라, 특허 실무상 해외 출원을 진행하고자 하는 출원인과 국내 대리인은 해외 패밀리 출원 심의 절차와 같은 엄격한 과정을 거쳐 특허의 가치를 면밀히 판단하고 있다. 결국 해외 출원의 비용과 시간의 부담에도 불구하고 특허의 가치를 인정받아 해외 출원을 진행했

다는 점에서 해외 패밀리 출원의 수는 그 자체로 특허의 가치를 대변하고 있음을 확인할 수 있다.

이에 따라, 특허 패밀리 수의 가중치가 0.1로 설정된 경우보다, 0.9 로 설정된 경우가 모델의 예측 성능이 더 우수한 것을 확인할 수 있으며, 특허 패밀리 수의 가중치가 다른 독립변수의 가중치(0.7)보다 높아진 경우에도 성능이 개선되었으므로, 연구 방법에서 제안한 것과 같이 특허 패밀리 수가 특허의 가치 산정에 미치는 영향이 크다는 것을 확인하였다.

## 6. 결론 및 향후 과제

본 연구에서는 국내 특허 데이터베이스 업체인 WipsOn 에서 제공하는 특허 데이터를 활용하여 특정 분야에서 특허의 가치를 평가하기 위해 중요한 변수가 무엇인지 확인하였다. 특허 데이터에 포함되는 여러 변수 중 특허 패밀리 수는 기업과 전문가의 특허 가치 판단의 산물이고, 특허 업계에서도 청구항 수나 독립항 수와 같은 지표보다 특허의 가치를 더욱 잘 내포하고 있는 것으로 판단되고 있다. 따라서 본 연구는 선행 연구에서 패밀리 수의 중요성에 대한 논의가 부족하다는 점에 착안하여 연구가 진행되었고 데이터로 중요성을 확인했다는 점에 의의가 있다.

다만, 본 연구의 한계점은 분야를 한정하는 과정에서 분석된 데이터의 n 수가 적다는 것과, 전문가 집단의 설문조사와 같이 패밀리 수의 중요성에 대한 근거를 추가하지 못했다는 점과, 여러 모델에 의한 분석을 수행하지 못했다는 점에 한계가 있다. 따라서, 분야의 한정을 완화하여 데이터의 n 수를 증가시키고, 변리사와 같은 전문가 집단의 설문조사를 수행하고, 복수의 모델에 의한 분석을 추가한다면 더욱 설득력 있고 개선된 성능의 연구가 수행될 것으로 기대한다.

## 참고문헌

[1] Dutta, S., & Weiss, A. M. The Relationship Between a Firm's Level of Technological Innovativeness and its Pattern of Partnership Agreements. Management Science, 43(3), 1997, 343-356.

[2] Nathan, F., Kenneth, T. Patent Valuation with Forecasts of Forward Citations. Journal of Business Valuation and Economic Loss Analysis, 12(1), 2016, 101-121.

[3] 박현우,이종택, “초기단계 기술의 가치 평가 방법론 적용 프레임워크”, 「기술혁신학회지」, 2012, 15(2): 242-261.

[4] Lee, Y. G., Lee, J. D., Song, Y. I. An analysis of citation counts of ETRI-Invented US patents. ETRI Journal, 28(4), 2006, 541-544.

[5] Wu, J. L., Chang, P. C., Tsao, C. C., Fan, C. Y. A patent quality analysis and classification system using self-organizing maps with support vector machine. Applied Soft Computing. 41, 2016, 305-316.

[6] Lin, B. W., Chen, C. J., Wu, H. L. Predicting citations to biotechnology patents based on the information from the patent documents. International Journal of Technology Management, 40(1-3), 2007, 87-100.

[7] Erdogan, Z., Altuntas, S., Dereli, T. Predicting Patent Quality Based on Machine Learning Approach. IEEE Transactions on Engineering Management, 2022, 1-14.

[8] Lanjouw, J., Schankerman, M. Patent quality and research productivity: Measuring innovation with multiple indicators. The Economic Journal, 114(495), 2004, 441-465.