# 머신러닝 기반 유효특허 분류기 개발에 관한 연구

장예훈 (가천대학교), 최성철 (가천대학교), 서원철 (부경대학교)

## Ⅰ. 연구배경 및 목적

### 연구 배경

- 기술 연구 개발 단계에서 특허권 확보는 매우 중요하므로 해당 기술의 유효특허 검색은 필수적인 작업임
- 유효특허 분류 업무 특성상 반드시 사람이 직접 실제 유효특허 수의 10~10,000배의 문서 검토 작업 후 유효특허 여부의 판별이 가능하지만 인력, 시간 등의 비용소모가 심각하며 수작업으로 인한 오차와 중복이 발생함

### 연구 목적

- 특허 데이터의 서지적 정보와 네트워크 정보, text 정보를 활용하여 머신러닝 기반의 유효특허 분류기를 개발하여 유효특허 분류 작업을 자동화 함
- 유효특허 분류기를 통해 1차적으로 분류 작업을 거치면 기존의 검토해야 하는 문서의 절대적인 수치가 줄어들게 되며 분류 과정에서 발생하는 비용과 오차 또한 줄일 수 있음

## Ⅱ. 연구방법 및 결과

### 1. 데이터 수집

- 특허나라에서 제공하는 '관리형 해상최종처분장 조성기술 개발'의 영문 검색식을 이용하여 WIPS DB에 유효특허 검색 조건에 맞게 검색한 모든 US 특허 데이터를 사용함

<표 1> 실험 데이터 set의 valid와 garbage 개수와 비율

| | 개수 (백분율) |
|---|---|
| **Valid** | 29 (0.0152) |
| **Garbage** | 1876 (0.9848) |
| **Total** | 1905 (1) |

- 전체 데이터 중에서 극소수의 유효특허를 찾아내고 Garbage 특허를 제거하는 모델에서는 Recall점수가 중요한 모델 평가지표가 됨

$$recall = \frac{TP}{TP+FN} = \frac{\text{유효특허 중 튜효특허라고 예측한 개수}}{\text{유효특허의 개수}}$$

### 2. 연구 방법

#### 1) 1차 데이터 전처리 및 실험

- WIPS DB에서 기본 제공되는 IPC, US Class, 날짜정보(출원, 공개, 등록), 국제공개등록번호, 우선권번호, 문헌종류코드 등 특허의 서지적 정보를 각 데이터의 특성에 맞게 가공함
- 전체 데이터를 7:3의 비율로 training, test set 분리
- 머신러닝 모델 학습 결과

<표 2> 1차실험결과-Decision Tree

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 0.99 | 1.00 | 0.99 |
| 1 | 0.67 | **0.29** | 0.40 |
| accuracy | | 0.99 | |

<표 3> 1차실험결과-Random Forest Classifier

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 0.99 | 1.00 | 1.00 |
| 1 | 1.00 | **0.57** | 0.73 |
| accuracy | | 0.99 | |

#### 2) 2차 데이터 전처리 및 실험

- 특허 데이터 set 중 등록번호를 이용하여 구글에서 제공하는 USPTO Bulk data에서 assignee, application citation, grant citation 정보를 추출하여 pivot table 형태의 네트워크 데이터 생성 및 전체 학습 데이터에 추가함
- 분류 모델로 학습 해야 하는 valid 특허 개수와 garbage 특허의 개수 비율이 1:9로 클래스 불균형 문제가 발생함. 본 연구에서는 분류 개수가 적은 valid 특허 클래스의 데이터를 SMOTE 기법을 활용하여 생성 후 각각의 데이터 set에 추가하여 valid 특허와 garbage 특허의 비율을 1:1로 조정하여 이 문제를 해결함
- 머신러닝 모델 학습 결과

<표 4> 2차실험결과-Decision Tree

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 0.99 | 0.99 | 0.99 |
| 1 | 0.36 | **0.57** | 0.44 |
| accuracy | | 0.98 | |

<표 5> 2차실험결과-Random Forest Classifier

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 |
| 1 | 1.00 | **0.71** | 0.83 |
| accuracy | | 1.00 | |

#### 3) 3차 데이터 전처리 및 실험

- 특허 데이터에서 abstract, title, claim 등과 같은 text 데이터의 TF-IDF 값을 구한 뒤 전체 학습 데이터에 추가함
- 머신러닝 모델 학습 결과

<표 6> 3차실험결과-Decision Tree

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 1.00 | 0.96 | 0.98 |
| 1 | 0.19 | **0.71** | 0.30 |
| accuracy | | 0.95 | |

<표 7> 3차실험결과-Random Forest Classifier

| | precision | **recall** | f1-score |
|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 |
| 1 | 1.00 | **0.85** | 0.92 |
| accuracy | | 0.99 | |

### 3. 결과

| data | Model | accuracy | f1-score | recall |
|---|---|---|---|---|
| only preprocessed data | Logistic Regression | 0.98 | 0.00 | 0.00 |
| | SVM | 0.98 | 0.00 | 0.00 |
| | Decision Tree | 0.99 | 0.67 | 0.29 |
| | Random Forest | 0.99 | 1.00 | **0.57** |
| preprocessed data with network data + SMOTE | Logistic Regression | 0.87 | 0.02 | 0.14 |
| | SVM | 0.70 | 0.02 | 0.43 |
| | Decision Tree | 0.98 | 0.36 | 0.57 |
| | Random Forest | 1.00 | 0.36 | **0.71** |
| preprocessed data with network data and text data + SMOTE | Decision Tree | 0.95 | 0.19 | 0.71 |
| | Random Forest | 0.99 | 1.00 | **0.86** |

## Ⅲ. 결론 및 개선방향

- 특허의 서지적 데이터 뿐만 아니라 네트워크 데이터, text 데이터를 학습 데이터에 포함하면 예측 모델의 recall 점수가 향상됨
- 머신러닝 모델 중 tree 계열의 모델 성능이 우수함
- 학습 속도 단축을 위해 데이터 set의 size를 줄이는 작업이 필요함
- 머신러닝 모델 뿐만 아니라 딥러닝 모델로 확장 가능함
- 기계번역 모델을 활용하면 여러 국가의 유효특허 분류기로 확장 가능함

이 성과는 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. NRF-2015R1C1A1A01056185).