Classification Assistance for Korean R&D Project Using Machine Learning Techniques

Jiseong Son[1], Yuchul Jung[2], Kwang-Nam Choi[1], Chul-Su Lim[1*]

[1]{jsson, choi, cslim}@kisti.re.kr, [2]jyc@kumoh.ac.kr

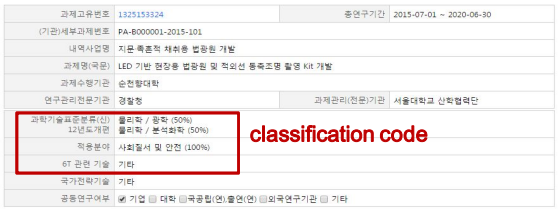
[1] NTIS Center, Korea Institute of Science and Technology

[2] Department of Computer Engineering, Kumoh National Institute of Technology

Introduction

- Currently, classifications codes should be manually input by researcher in Korean R&D project information registration system.
- However, an accurate classification of R&D project has a few issues
  - it has many kinds of codes in *Korean national science and technology standard classification system* that is consist of 3 levels.
  - a convergence research area doesn't have a well-defined boundary.
- *Therefore, researchers spend a lot of time for finding classification code that is suited for their research area and cause some errors such as wrong input code.*
  - it needs an expert of each research area to modify wrong input code and requires more time and cost for analyzing R&D project information.

Korean National S&T Standard Classification

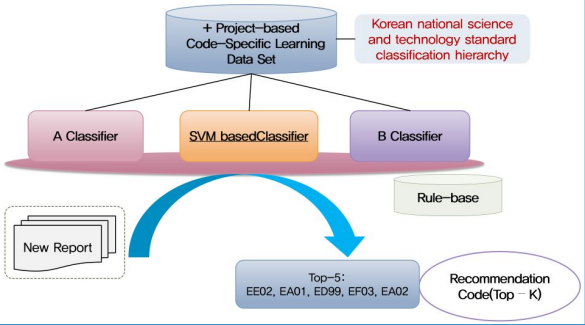


CLASS_CD	CLASS_CD_NM_ENG
EE	Information/Communication
EE01	Information Processing Theory
EE0101	Computer theory
EE0102	Algorithm
EE0103	Compiler
EE0104	Programming language/natural language processing
EE0105	Database
EE0106	Software engineering
EE0107	Operating system
EE0108	Artificial intelligence
EE0109	HCI
EE0110	Cloud computing/Grid computing
EE0111	Real time system
EE0112	Information searching/information browsing
EE0199	Other information process theory
EE02	Software
EE0201	Embedded software
EE0202	Software solution
EE0203	System Integration
EE0204	Internet software
EE0299	Other software

Proposed Method

we implement classification assistance using SVM algorithm and experiment with machine learning techniques using R&D project information

- R&D Project information are input and generated Data Set.
- We process morpheme analysis, stopword, and etc.
- Features are extracted by TF-IDF.
- Finally, it recommend suitable Top-K class code for classification of research area with R&D project based on Korean national science and technology standard classification hierarchy.



Implementation

Environment and Data Set

- **Environment:** Python 2.7, Jupyter Notebook, NumPy, SciPy, Scikit-learn
- **Data Set:** we tested for middle classification(ie. EE01) level using Korean R&D projects information from 2014 to 2015.

1st,2nd,4th Test	EE. Information/Communication Dataset: Overall 8,552 reports
3rd Test	E. Artifact Dataset: Overall 42,610 reports

PJT_ID	KOR_PJT_NM	KOR_KYWD	ENG_KYWD	RSCH_GOAL_ABSTRACT	RSCH_ABSTRACT	EXP_EFCT_ABSTRACT
1425088266	스마트 안말기 기	모바일Web/App	Mobile Web/APp,P	기업 모바일 Wep/App 연계 - 기업용 ERP 연동 개	0	경제적 효과- 서험 서비스
1711014031	국방IT·SW융합지	아이디어오디션	Idea Audition, Defe	1.국방산업별 수요기업과 IT 네트워크 구축, 국방IT 융	다수의 군사혁신 전문가 풀	
1711013230	5G 광대역 통신	밀리터파 통신	MillimeterWaveCom	기존의 크고 고비용인 밀리미	1. 밀리미터파 CMOS 용	1. 밀리미터파 CMOS 회로
1425089252	근거리 통신기술	블루투스 저전력	BLE Bluetooth Low	o PIPS(Patient Indoor Positi	o 입원 환자 위치추적 내	o 의료인력 부족 보완 - 무
1425087925	모바일 통합 Volu	모바일내장코드	MBID M-Generator	모바일 개발 기술의 향상과 (이번 개발에 개발될 최첨기술적측면1)기술 향상- 도		

RSCH_AREA_CLS1_CD	RSCH_AREA_CLS2_CD	RSCH_AREA_CLS3_CD	RSCH_AREA1_WEIGHT	RSCH_AREA2_WEIGHT	RSCH_AREA3_WEIGHT	STAN_YR
EE0202			100			2014
EE1499	EE0299		50	50	0	2014
EE1101	EE1105	ED0607	50	30	20	2014
EE0202			100			2014
EE0202			100			2014

Evaluation

Test Result

- Train/Test ratio = 9 : 1
  - By making several trials with different train/test combinations, we can use the average accuracy

```
EE05 Acc.: 0.842105263158 Correct: 16.0 Tested: 19.0 #Train 241.0
EE13 Acc.: 0.85 Correct: 17.0 Tested: 20.0 #Train 184.0
EE99 Acc.: 0.868421052632 Correct: 66.0 Tested: 76.0 #Train 708.0
EE07 Acc.: 0.85 Correct: 17.0 Tested: 20.0 #Train 227.0
EE14 Acc.: 0.285714285714 Correct: 4.0 Tested: 14.0 #Train 115.0
EE11 Acc.: 0.91935483871 Correct: 57.0 Tested: 62.0 #Train 578.0
EE08 Acc.: 0.791666666667 Correct: 19.0 Tested: 24.0 #Train 208.0
EE09 Acc.: 0.927272727273 Correct: 51.0 Tested: 55.0 #Train 506.0
EE06 Acc.: 0.947368421053 Correct: 72.0 Tested: 76.0 #Train 650.0
EE10 Acc.: 0.895833333333 Correct: 43.0 Tested: 48.0 #Train 439.0
EE04 Acc.: 0.88 Correct: 22.0 Tested: 25.0 #Train 242.0
EE12 Acc.: 0.875 Correct: 28.0 Tested: 32.0 #Train 177.0
EE02 Acc.: 1.0 Correct: 290.0 Tested: 290.0 #Train 2580.0
EE03 Acc.: 0.898734177215 Correct: 71.0 Tested: 79.0 #Train 671.0
EE01 Acc.: 0.978873239437 Correct: 139.0 Tested: 142.0 #Train 1306.0
============================================================
Total Test Examples 982
Micro Acc.(item level) 0.928716904277
```

No.	Top-N	Data Set	Accuracy(%)
1	Top-1	EE. Information/Communication	50%
2	Top-10	EE. Information/Communication	90%
3	Top-10	E. Artifact	82%
4	Top-5	EE. Information/Communication	93% ~ 95%
		E. Artifact	85% ~ 87%

Conclusion and Future Works

- The classification of R&D project is important for management and distribution of information, efficient management for human resource, measurement of statistical data and efficient planning projects in research area.
- Using the result of measurement,
  - improve accuracy and usability of classification for R&D area.
  - reduce time for verification of input codes.

- As future work,
  - plan to apply various domain data.
  - attempt to test all class code of Korean national science and technology standard classification.