Transactions of the Korean Nuclear Society Autumn Meeting
December 16-18, 2020

Evaluation of Proper Hyperparameters in Machine Learning Algorithms for Fuel Cycle Related Documents Classification

2020. 12. 16

Byoungchan Han, Byeounghyeok Ha, and Tongkyu Park

bchan007@fnctech.com

㈜미래와도전
FNC Technology Co., Ltd.

www.fnctech.com

CONTENTS

1 Introduction

2 Methodology

3 Analysis

4 Results

5 Conclusion

1. Introduction

■ The Additional Protocol (IAEA)

▶ Every nuclear activities should be reported

- Particularly fuel cycle related activities

▶ Growth of unreported nuclear activities

- Depreciation of international credibility
- Ex) Different academic studies, still contain fuel cycle related activities

▶ Collection and analysis system for nuclear fuel cycle activities

- Crawling domestic research paper
- Classify product through machine learning algorithms
- Create metadata database through analysis of documents

2. Methodology

■ Naïve Bayesian

▶ Conditional probability model

- Let $P(C_k|x_1, \dots, x_n)$ as the probability of being vector $\boldsymbol{X}$ is contained in category C_k where $\boldsymbol{X} = (x_1, \dots, x_n)$.
- According to Bayes' Theorem,
  $P(C_k|x_1, \dots, x_n) = \frac{P(C_k)P(x_1,\dots,x_n|C_k)}{P(x_1,\dots,x_n)} \propto P(C_k)\prod_i P(x_i|C_k)$.
- Naïve Bayesian can classify documents by calculating word frequencies.

■ SVM model

▶ Algorithm for finding the optimal boundary to classify data

- Convert documents to vectors. (TF-IDF preprocessing)
- Calculate the hyperplane maximizing the margin

2. Methodology

■ XGBoost

▶ Ensemble of decision trees

- Each decision tree has nodes to classify features of data.
- Decision tree divides data into smaller groups based on their features.
- Once the data points continue to be classified and reach the deepest level of the tree, they are finally classified to their fitted groups as the leaves are paired with their own categories.

■ Experimental Setting (1/3)

▶ Case Design

- Generated a group of documents and divided it into five pieces.
- Four of them were used as training group, and the other as test group in machine-learning models.
- Five independent experiments were performed to obtain the mean and the standard deviation of the F1 scores. (5-fold method)

▶ Lists of Documents

- Fuel cycle documents
- Non-fuel cycle documents, but in the field of nuclear energy
- Documents unrelated to nuclear energy

■ Experimental Setting (2/3)

▶ Group Configuration

- 9 groups are built of varying composition of documents

Case	Fuel Cycle Docs.	Non-Fuel Cycle Docs.	Non-Nuclear Docs.	Total
I	300	0	300	600
II	300	300	0	600
III	300	150	150	600
IV	900	0	900	1,800
V	900	900	0	1,800
VI	900	450	450	1,800
VII	1,500	0	1,500	3,000
VIII	1,500	1,500	0	3,000
IX	1,500	750	750	3,000

■ Experimental Setting (3/3)

▶Group Configuration

- The purpose of cases I, IV, and VII was to classify fuel cycle documents from documents unrelated to nuclear energy.
- The purpose of cases II, V, VIII was to classify fuel cycle documents from documents related to nuclear energy.
- The purpose of cases III, VI, IX was to classify fuel cycle documents from documents on various topics.

3. Analysis

■ Experimental Method (1/2)

▶ Grid Search

- Grid search approach is to obtain the best hyperparameters for optimization problem from a list of parameter options provided.
- For the SVM model, parameter **C** and $\boldsymbol{\gamma}$ were adjusted for optimization.
- For the XGBoost model, "**learning rate**", "**minimum child weight**", and "**tree depth**" were tuned for optimization.
- In the case of Naïve Bayesian model, optimization is unnecessary as there is no parameter to modify.
- The range of each parameter to be used in grid search optimization was determined by erasing meaningless values through multiple pre-testing.

3. Analysis

■ Experimental Method (2/2)

▶ Range of SVM Hyperparameters for Optimization

- C : in the range of $10^{-3} \sim 10^{5}$
- γ : in the range of $10^{-3} \sim 10^{3}$

▶ Range of XGBoost Hyperparameters for Optimization

- Learning Rate : in the range of $10^{-2} \sim 1$
- Minimum Child Weight : in the range of 0.5 ~ 1
- Tree Depth : in the range of 4 ~ 8
- These three parameters were chosen for being most influential on the performance of the XGBoost model.

4. Results

■ Performance of Naïve Bayesian

Case	F1 Score			
	Test Set		Training Set	
	Mean	SD	Mean	SD
I	0.989	0.011	0.997	0.001
II	0.915	0.031	0.944	0.006
III	0.895	0.024	0.965	0.002
IV	0.982	0.006	0.993	0.001
V	0.952	0.006	0.978	0.002
VI	0.960	0.011	0.980	0.002
VII	0.978	0.004	0.988	0.001
VIII	0.779	0.013	0.837	0.004
IX	0.838	0.009	0.865	0.002

4. Results

■ Performance of Naïve Bayesian

- ▶ Naïve Bayesian was found to be effective in separating fuel cycle documents from non-nuclear documents as F1 scores of test set in Case I, IV and VII are above 0.97.
- ▶ Classification of fuel cycle documents and non-fuel cycle documents was found to be more difficult than classification of fuel cycle documents and non-nuclear documents.
- ▶ F1 scores of Case II, V, VIII showed insignificant difference with the Case III, VI, and IX.

4. Results

■ Performance of SVM (the Best and the Second Best)

Case	F1 Score		Hyperparameter	
	Mean	SD	C	γ
I	0.993	0.003	**5**	**0.1**
	0.993	0.003	100	0.01
II	0.901	0.041	**5**	**0.1**
	0.893	0.034	10	0.1
III	0.956	0.006	**10**	**0.1**
	0.955	0.012	5	0.1
IV	0.995	0.003	**10**	**0.1**
	0.995	0.003	100	0.01
V	0.959	0.005	**5**	**1**
	0.959	0.005	10	1
VI	0.971	0.006	**100**	**0.01**
	0.970	0.006	10	0.1
VII	0.997	0.002	**10**	**0.1**
	0.997	0.002	100	0.01
VIII	0.783	0.011	**10**	**0.1**
	0.782	0.016	10	1
IX	0.863	0.015	**10**	**0.1**
	0.863	0.015	10	0.1

4. Results

■ Performance of SVM

- ▶SVM showed optimal performance when the hyperparameter C is in the range of 5 and 100, with γ in the range of 0.01 to 0.1.
- ▶Optimized SVM model showed superior performances than Naïve Bayesian in the eight of the nine cases.
- ▶SVM showed extremely accurate classification performance close to 100% in separating fuel cycle documents and non-nuclear documents.
- ▶As the document set in Case III, VI, and IX contains non-nuclear documents, whereas that in Case II, V, VIII is not, F1 Scores of the Case III, VI, IX showed better results than the Case II, V, VIII.

4. Results

■ Performance of XGBoost (the Best and the Second Best)

Case	F1 Score	Hyperparameter		
	Mean	Learning Rate	Min Child Weight	Tree Depth
I	0.997	**0.2**	**1**	**4**
	0.997	0.2	1	6
II	0.956	**0.05**	**0.75**	**4**
	0.954	0.1	0.5	8
III	0.997	**0.05**	**0.5**	**4**
	0.997	0.05	0.75	4
IV	0.998	**0.1**	**1**	**4**
	0.998	0.1	1	6
V	0.956	**0. 05**	**0.75**	**4**
	0.954	0.1	0.5	8
VI	0.967	**0.2**	**1**	**4**
	0.967	0.2	1	8
VII	0.999	**0.1**	**1**	**4**
	0.998	0.1	0.5	4
VIII	0.818	**0.1**	**0.5**	**8**
	0.815	0.1	1	8
IX	0.922	**0.2**	**1**	**8**
	0.922	0.2	0.5	6

4. Results

■ Performance of XGBoost (the Worst)

Case	F1 Score	Hyperparameter		
	Mean	Learning Rate	Min Child Weight	Tree Depth
I	0.993	0.1	0.75	6
II	0.925	0.05	0.5	8
III	0.993	0.1	0.75	4
IV	0.994	0.05	0.5	4
V	0.925	0.05	0.5	8
VI	0.953	0.05	1	8
VII	0.992	0.05	1	4
VIII	0.788	0.2	1	8
IX	0.907	0.2	1	8

4. Results

■ Performance of XGBoost

▶ Learning Rate in the range of 0.05 to 0.2, Minimum Child Weight around 1, and Tree Depth in the range of 4 to 8 showed sufficient result in document classification.

▶ Optimized XGBoost model showed even more superior performances than SVM model in every test case.

▶ However, tendency of hyperparameters to optimize the XGBoost model was hard to find.

▶ The difference between the best F1 scores and the worst was less than 0.03 in all cases. It implies that XGBoost model is still effective to classify documents without tuning hyperparameters.

5. Conclusion

■ Classification performances of optimized models

- ▶ Naïve Bayesian, SVM and XGBoost all showed effective on classifying documents.
- ▶ The performance of the three models was followed by XGBoost > SVM > Naïve Bayesian .

■ Optimal hyperparameters

- ▶ The SVM model was optimized at 5~100 for C and 0.01~0.1 for γ.
- ▶ The XGBoost model was optimized at 0.05~0.2 for learning rate, 4~8 for tree depth, and near 1 for minimum child weight.
- ▶ Optimization of the XGBoost model did not significantly change its performance.

THANK YOU