# Two Classification Algorithms for Nuclear Fuel Cycle Related Documents

Tongkyu Park[1,*], Yunpil Jeong[1], Byoungchan Han[1], Sang Jun Lee[2], Chan Seo Lee[2], and Dong-hoon Shin[2]
[1] FNC Technology Co., Ltd., 32F, 13 Heungdeok 1-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea
[2] Korea Institute of Nuclear Nonproliferation and Control, 1534, Yuseong-daero, Yuseong-gu, Daejeon, Republic of Korea
[*] tongkyu@fnctech.com

## 1. Introduction

The Korea Institute of Nuclear Nonproliferation and Control (KINAC) is currently developing a collection and analysis system for nuclear fuel cycle related R&D projects and activities aiming at fulfilling the additional protocol with the IAEA[1] because information of nuclear fuel cycle related products and activities supported by governmental funding needs to be reported to the IAEA according to the additional protocol. This paper presents two algorithms to be implemented into the collection and analysis system for the automatic classification of fuel cycle related documents.

## 2. Classification algorithms

The natural language processing (NLP) should be done as a first step for classifying collected documents. To achieve this, a well-known TF-IDF (Term Frequency and Inverse Document Frequency) method was used in this research. Consequently, two classification algorithms, i.e., SVM (Support Vector Machine)[2] and XGBoost (Extreme Gradient Boosting)[3], were coupled with TF-IDF.

### *2.1 Support Vector Machine*

SVMs[2] have been proven as one of the most powerful learning algorithm for text categorization. Let's define $N$ is the number of features, and then the goal of this algorithm is to find a hyperplane in an $N$-dimensional space that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. In this research the Radial Basis Function (RBF) kernel SVM was selected. In the kernel, two parameters (gamma and $C$) should be predetermined by user. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning far and high values meaning close while the $C$ parameter trades off correct classification of training examples against maximization of the decision function's margin.

### *2.2 Extreme Gradient Boosting*

The XGBoost algorithm[3] is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. The key concept of this algorithm is so-called decision tree ensembles consisting of a set of classification and regression trees (CART). Several parameters, such as learning_rate, min_child_weight, max_depth, colsample_bytree, and so on, need to be determined by performing sensitivity analyses. The learning_rate is the step size shrinkage used in update to prevent overfitting while the_child_weight is a minimum sum of instance weight needed in a child. The max_depth and colsample_bytree are a maximum depth of a tree and the subsample ratio of columns when constructing each tree, respectively.

## 3. Numerical results

To verify the effectiveness and efficiency of the two proposed classification algorithms coupling with TF-IDF as a NLP tool, a series of sensitivity analyses were performed for two test problems listed in Table 1. Among the 900 documents in each group, 80% (720 documents) were used as training data and thus a total of 360 documents including 180 fuel cycle related documents in each problem were automatically classified with the proposed algorithms.

Table 1. Number of training and test documents

| Problem Number | Fuel cycle documents | Non-fuel cycle documents[a)] | Non-nuclear documents |
|---|---|---|---|
| 1 | 900 | 0 | 900 |
| 2 | 900 | 900 | 0 |

[a)] Documents in nuclear science or engineering

### *3.1 Support Vector Machine*

A total of 48 cases with different values of $C$ and gamma each other, was defined for each problem. Table 2 shows that F1 score for the first problem is 0.995 when $C$ and gamma were set to 10 and 0.1. It means only 2 or 3 documents among 360 documents were

misclassified when the TF-IDF/SVM algorithm was applied. Noted that it is hard to classify the documents in the second problem since all the documents were technical papers in the nuclear science field. It is larger than 0.95 in the aspect of F1 score, and thus it is expected that this algorithm is powerful to classify fuel cycle related documents.

Table 2. Numerical results obtained from SVM

| Problem Number | F1 score | Parameters | |
|---|---|---|---|
| | | C | gamma |
| 1 | 0.995 | 10 | 0.1 |
| 2 | 0.959 | 5 | 1 |

### *3.2 Extreme Gradient Boosting*

A total of 324 cases with different values for the 4 parameters mentioned above were defined for each problem. F1 scores were 0.998 and 0.956, respectively, for the two problems. It can be carefully concluded that this algorithm is also powerful.

## 4. Conclusion

In this study, two classification algorithms were proposed and their effectiveness and efficiency were verified by examining two test problems. Numerical results show that the proposed algorithms were powerful to classify the fuel cycle related documents and can be effectively used as main classification algorithms in the collection and analysis system.

## ACKNOWLEDGEMENT

This work was supported by the Nuclear Safety Research Program though the Korean Foundation Of Nuclear Safety (koFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of Korea. (No. 1803021)

## REFERENCES

[1] S.H. Yoon and D.H. Shin, "A Conceptual Design of the Information Analysis System for Searching Nuclear Fuel Cycle Related R&D Projects", Proc. of the KRS 2018 Autumn Conference, 16(2), October 31-November 2, 2018, Jeju.

[2] C. Cortes and V. Vapnik, "Support Vector Networks", Machine Learning, 20, 273-297 (1995).

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, August 13-17, 2016, San Francisco.