Transactions of the Korean Nuclear Society Autumn Meeting
Goyang, Korea, October 24-25, 2019

Classification of Nuclear Fuel Cycle Related Documents by Supervised and Unsupervised Learning Algorithms

Byoungchan Han*, Kibeom Park, Yunpil Jeong, Tongkyu Park
[a]*FNC Technology Co., Ltd., 32F, 13 Heungdeok 1-ro, Giheung-gu, Yongin-si, Gyeonggi-do, Korea*
**Corresponding author: bchan007@fnctech.com*

1. Introduction

Republic of Korea (ROK) made an additional protocol with the IAEA in 2004. The key point of this additional protocol is that ROK should report nuclear or nuclear related activities, particularly fuel cycle related activities to the IAEA. As a part of the efforts to fulfill this additional protocol, a three-year study was launched to develop a collection and analysis system for nuclear fuel cycle related R&D projects and activities at the Korea Institute of Nuclear Nonproliferation and Control (KINAC) in 2018. [1]

This paper introduces one supervised and three unsupervised learning algorithms to classify fuel cycle related documents that can be applied to the collection and analysis system.

2. Methodology

2.1. Long-Short Term Memory

Long-short term memory (LSTM)[2] is one of the modified RNN (Recurrent Neural Network) algorithms[3]. Note that the RNN method is well-known, and thus a detail description of the method is skipped in this section. Researchers have consistently claimed that the RNN model becomes harder to learn relationship between data as their distance in sequence increases, even though it has the advantage of analyzing sequential data. In other words, the distance between data of a target point (present) and its following point (next) needs to be set as close as possible. This phenomenon is called gradient vanishing problem, which is caused by the characteristic of RNN backpropagation and it is well-known drawback of RNN.

On the other hand, LSTM resolves the drawback by employing both multiplication and addition for feedback loops. [2] LSTM unit consists of one cell state and three types of gate. The cell state is a passage for gradient flow from unit to unit. It is a key part to prevent gradient vanishing problem, hence it contains both operations for addition and multiplication. Three types of gate interact with cell state to remember patterns selectively by forgetting and maintaining information at certain point. First of all, forget gate decides whether to discard previous information or not. It calculates sigmoid function output from the past and present information. On the other hand, input gate decides which values to be updated from the present data while output gate decides the amount of the final result to be printed and sent to the following unit cell.

2.2. Self Organizing Map

Self organizing map (SOM) [4] is a method to match each high-dimensional data with two or three dimensional lattice. In other words, it enables visualization of high-dimensional data from dimensionality reduction. Moreover, as these reduced data still preserve topological features, users can separate clusters of dataset with a map.

SOM architecture is a fully-connected neural network. It contains only two layers; input layer and output layer. Data in input layer are represented as vectors on the multidimensional space, and nodes in output layer are represented as lattice points in two-dimensional or three-dimensional form. Each pair of input and output points is connected each other, and this connection is calculated from weight vector.

Whenever a certain input value is given, SOM finds the closest output node from the given input value. Closest output node is called “winning node”, assigns input value in his own inside. Distance between input and output nodes is calculated as Eq. 1, where $\mathrm{X_i(t)}$ is *i*-th input vector at time *t*, and $\mathrm{W_{ij}(t)}$ refers to weight vector between *i*-th input vector and *j*-th output node.

$$\mathrm{d_j} = \sum_{\mathrm{i=0}}^{\mathrm{N-1}} \left(\mathrm{X_i(t)} - \mathrm{W_{ij}(t)}\right)^2 . \qquad (1)$$

Consequently, SOM compensates weight vectors of very winning node and surrounding neighbors such that each pair of input node and conforming winning node gets closer. Winning node is highly modified, as it is the closest node from the input vector, whereas surrounding nodes rate of change diminishes gradually. Eq. 2 presents equation for modifying weighting vector between *i*-th input vector and *j*-th output node, where $\mathrm{\mu_t}$ means to the learning rate and $\lambda_{x_i}^{j,t}$ is the rate of change by distance.

$$\mathrm{W_j(t+1)} = \mathrm{W_j(t)} + \mathrm{\mu_t}\lambda_{x_i}^{j,t}\left(X_i(t) - W_j(t)\right). \qquad (2)$$

2.3. K-Means Algorithm

K-means algorithm [5] is a classic algorithm for clustering. For its simplicity, it requires the number of cluster as much as user wants. Each cluster has the sole centroid, and every data point is allocated to the nearest centroid. Learning process of k-means algorithm can be described as modifying centroids, in order to cluster data accurately.

Learning process is divided into expectation step and maximization step. Expectation step is the process allocating every data to the nearest centroid using so-called Euclidean distance while maximization step is to modify every centroid's location to cluster's center. K-means algorithm is simply modifying centroids through calculating mean value of conforming data points' coordinates.

2.4. Hierarchical Clustering Analysis

Hierarchical clustering Analysis (HCA) [6] is the algorithm to cluster data using hierarchical tree model. This tree model is called "Dendrogram", which shows the order in which data are combined. It requires every distance or similarity between each data point.

There are two types of HCA algorithm. One is an agglomerative approach and the other is a divisive approach. The agglomerative approach is a bottom-up method. In the method, data points make up their clusters, and formed clusters are bound with each other until sole cluster is left. On the other hand, divisive approach is a top-down method. In the method, cluster is split into smaller group until individual data points remain. Once dendrogram is built, it can be split up at the certain level and presents clusters. Therefore, HCA does not require the number of cluster compared to the k-means algorithm mentioned above

3. Analysis and Results

3.1. Supervised Learning

In our previous study [7], performance of support vector machine (SVM) for classifying documents was presented. The performance of LSTM for classifying documents is described in detail as below. The numerical results in terms of "accuracy" by LSTM and SVM are presented and compared each other in this section. It should be noted that SVM was the most effective algorithm for document classification in the previous study.

To verify the effectiveness of LSTM on classifying documents, three test sets are conducted. Each test set consists of nuclear related documents and nuclear unrelated documents as shown in Table 1. For each test set, 64%, 16%, and 20% of the given data were used as training group, validation group, and test group, respectively. As shown in Table I, results in all three groups are all over 99% in the aspects of accuracy. For those calculations, the dimensionalities of input vector and hidden state vector are set as 100 and 128, respectively. 128 nodes are given as the fully-connected layer after LSTM layer for model's complexity in these problems. Note that the three parameters are user-specified values, and thus a series of sensitivity study needs to be performed to optimize those values[2].

From this observation, it can be concluded that LSTM can be used to effectively classify the nuclear related documents from the given documents set.

As a next step, a series of test (3 independent cases) was done to classify the fuel cycle related documents from nuclear related documents. The performances of LSTM are summarized in Table II. In the first case, the LSTM method can distinguish all the fuel cycle related documents from 600 documents consisting of 300 fuel cycle documents and 300 non-fuel cycle documents.

Table I: Nuclear Related Document Classification

Set number	Nuclear documents	Non-nuclear documents	Accuracy (%)
1	300	300	99.2
2	900	900	100
3	1500	1500	99.7

Table II: Fuel Cycle Related Document Classification

Set number	Fuel cycle documents	Non-fuel cycle documents	Accuracy (%)
1	300	300	100
2	900	900	94.7
3	1500	1500	76.7

Result shown in Table II indicates that performance of LSTM classifying fuel-cycle documents decreases with the number of data increases. However, we suppose that the cause of this phenomenon is due to the mixture of the documents' scope range.

Therefore, we conducted an experiment classifying KNS papers by to justify that LSTM shows sufficient performance on classifying two subjects. KNS papers from summer meeting, 2017 to autumn meeting, 2018 were used for classification data. Each performance between two divisions was calculated one by one with LSTM and SVM. Papers were formerly classified by 11 divisions, since 7th division (Division of Radiation Protection) and 8th division (Division of Radiation Utilization and Instrumentation) were combined until summer meeting, 2017. The amount of values by range is shown as in Table III. Mean value of accuracy of classification using LSTM was 92.35% while SVM was 95.78%. Although SVM shows better performance than LSTM, it can be carefully concluded that LSTM has the ability to classify documents.

Table III: KNS Papers Classification by Division

Accuracy Range (%)	SVM	LSTM
95~100	33	18
90~95	20	24
85~90	0	9
80~85	1	3
75~80	0	1
0~75	1	0

3.2. Unsupervised Learning

In order to confirm that unsupervised learning worked properly, we clustered 600 documents (300 non-nuclear documents, 300 nuclear documents) with known answers. Figures 1 and 2 show the result of clustering using k-means and HCA, respectively. Table IV shows the ratio at which the document was correctly identified. Each unsupervised learning method showed similar results and very good identification ratio for these documents case.

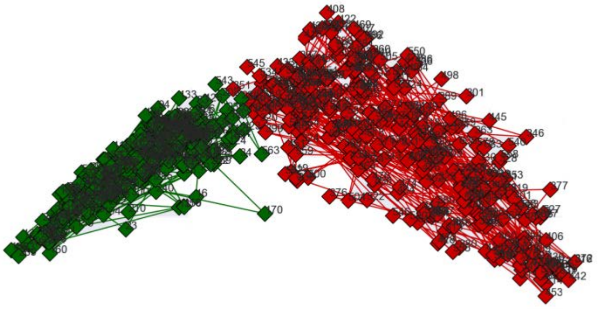

Fig. 1. Documents clustering results using k-means (nuclear vs non-nuclear).

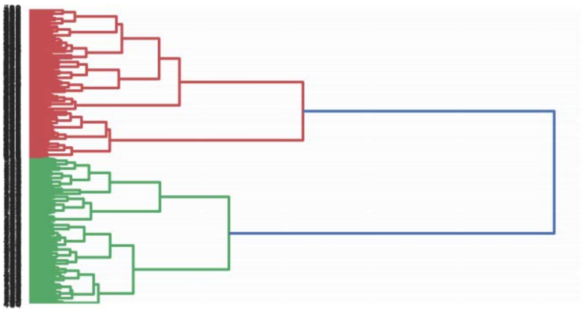

Fig. 2. Documents clustering results using HCA (nuclear vs non-nuclear).

Table IV: Identification Ratio with Clustered Documents(Nuclear vs Non-nuclear)

	Nucl Docs.	Non-Nucl Docs.
Corr./Not Corr.	293/7	298/2
Identification Ratio	97.7%	99.3%

Subsequently, a total of 600 fuel cycle and non-fuel cycle related documents (300 fuel cycle related and 300 non-fuel cycle related documents) were clustered. Figures 3 to 5 and Table V show the results of given 600 documents with known answer. Figures 3 to 5 show the result of clustering using k-means with different number of clusters and HCA.

K-means and HCA showed similar results for 3 clusters case. In case of 2 clusters k-means, the centroid of cluster was not matched well with HCA. To know how many documents were not correctly identified, the identification ratio was checked and showed in Table V.

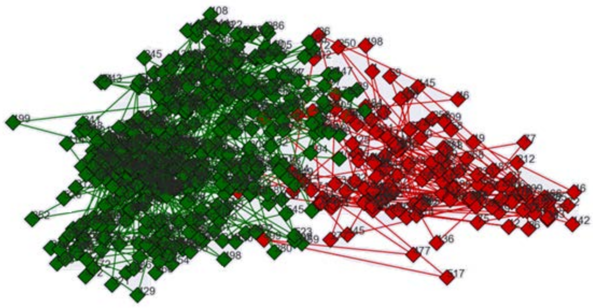

Fig. 3. 600 documents clustering results using k-means - 2 clusters (fuel cycle vs non-fuel cycle).

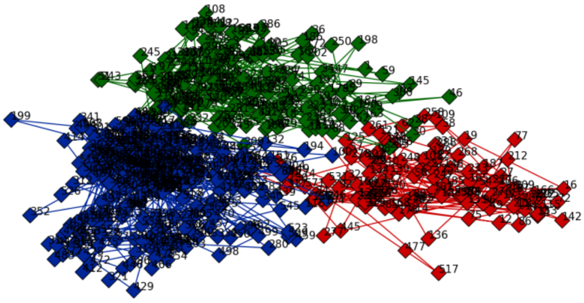

Fig. 4. 600 documents clustering results using k-means - 3 clusters (fuel cycle vs non-fuel cycle).

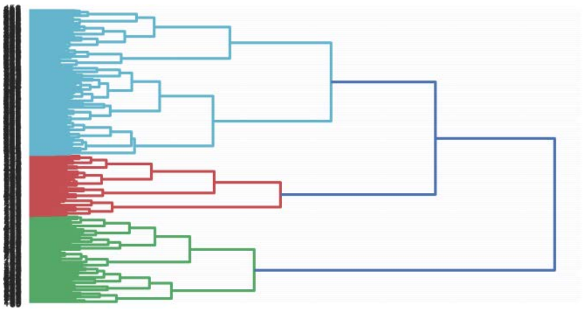

Fig. 5. 600 documents clustering results using HCA (fuel cycle vs non-fuel cycle).

Table 5: Identification Ratio with Clustered Documents (Fuel Cycle vs Non-fuel Cycle)

	Fuel Docs.	Non-Fuel Docs.
Corr./Not Corr.	284/16	128/172
Identification Ratio	94.7%	42.7%

There were a large number of mismatched documents in non-fuel cycle related documents. In this result, it can be seen that unsupervised learning can show correlations or trends between documents, but not accurate classification results.

Finally it is the result of unsupervised learning analysis of the KNS paper in 5th division. There were totally 87 documents and 3 methods (k-means, HCA, SOM) were used to cluster these documents. Figures 6 to 8 show the clustering results.

The three methods predict that the documents of the 5th division consist of two or three categories (clusters). Note that if the cluster-centroid documents can be investigated and analyzed, it is possible to suggest the possibility of confirming the characteristics of each cluster.

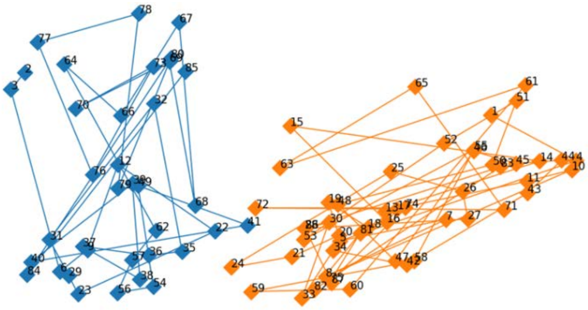

Fig. 6. 5th division clustering results using k-means.

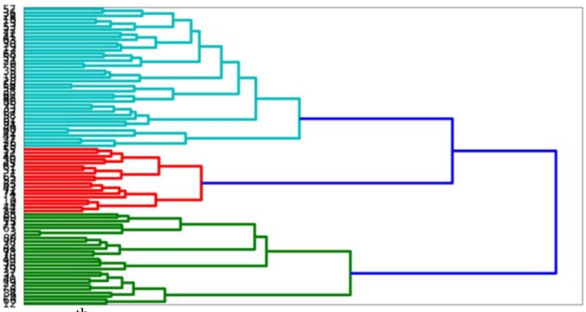

Fig. 7. 5th division clustering results using HCA.

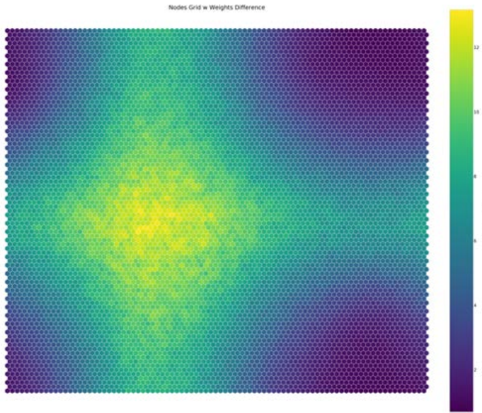

Fig. 8. 5th division clustering results using SOM.

4. Conclusion

This study presented multiple machine-learning algorithms that can pick out fuel cycle related documents from collected raw documents. LSTM and SVM, which are supervised learning algorithms, showed sufficient performance for classifying documents. Moreover, three kinds of unsupervised learning algorithm, SOM, HCA, k-means are suggested for clustering documents. Results indicate that unsupervised learning algorithms could not classify accurately. Regardless of the results, they showed potential in showing correlations or trends which can be applied on semantic web and ontology.

Acknowledgement

This work was supported by the Nuclear Safety Research Program through 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] Sung-ho Yoon and Dong-hoon Shin, “A Conceptual Design of the Information Analysis System for Searching Nuclear Fuel Cycle Related R&D Project”, Proc. of the KRS 2018 Autumn Conference, 16(2), October 31 – November 2, 2018, Jeju, Korea.

[2] Sepp Hochreiter, Jürgen Schmidhuber, "Long short-term memory", Neural Computation, 9(8), pp. 1735-1780, 1997.

[3] Ronald J. Williams, Geoffrey E. Hinton, David E. Rumelhart, "Learning representations by back-propagating errors", Nature, 323(6088), pp. 533-536, 1986.

[4] Teuvo Kohonen, “Self-organized formation of topologically correct feature maps”, Biological Cybernetics, 43, pp. 59-69, 1982.

[5] J. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations”, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 4, pp. 281–297, 1967.

[6] Rokach, Lior, and Oded Maimon, "Clustering methods", Data mining and knowledge discovery handbook, Springer US, pp. 321-352, 2005.

[7] Tongkyu Park, Yunpil Jeong, Byoungchan Han, Sang Jun Lee, Chan Seo Lee, and Dong-hoon Shin, “Two Classification Algorithms for Nuclear Fuel Cycle Related Documents”, Proc. of the KRS 2019 Spring Conference, 17(1), May 8 – May 10, 2019, Busan, Korea.