International Journal of Software Engineering and Its Applications
Vol.14, No.1 (2020), pp.17-26
http://dx.doi.org/10.21742/ijseia.2020.14.1.03

Machine Learning Techniques in Structural Fire Risk Prediction

Jaesung Chang[1], Jaeyoung Yoon[2], and Gunho Lee[3*]

Dept. Industrial and Information Systems Eng, Soongsil Univ., Sangdoro 369, Dongjakku, Seoul, Republic of Korea
[1]pearlcrum@naver.com, [2]jaeyoung9826@daum.net, [3]ghlee@ssu.ac.kr*

Abstract

Fires often occur, and the damage caused by them is often irreversible. Fire-prone environments can be identified through historical data, and predictive models are recommended to prevent fires in advance. This study uses a variety of machine learning techniques to build fire prediction models and perform a comparative analysis to predict fires. We use data from local fire departments in South Korea to build fire prediction models using decision trees, random forest, XGBoost, extra tree classification, artificial neural networks, and more. Before creating the fire prediction models, we analyze and significant predictive features of a structural fire. We compared the fire prediction models and showed accuracy, F1-score, precision, and recall. The prediction model built with the random forest is the most accurate, but there is a little difference in the accuracy of each model trained with the extra tree classifier, XGBoost, and neural network. For the F1 Score, the model with a neural network shows the best value.

Keywords: *Machine learning methods, Fire prediction, Comparative analysis*

1. Introduction

This study views that the fire can occur in a similar pattern and similar fire can frequently occur. The fire occurs in a variety of ways, in addition to the natural environments that can not be avoided. On the other hand, there are many ways to prevent the fire from experience. In this study, we use data mining techniques to analyze past fire cases to build a fire prediction model. The existing fire prediction models are made by analyzing the relations with natural attributes such as temperature and weather to predict the fire [1]. In addition to the natural attributes, we consider the material of the building frame and environmental factors of the building. We use fire case data from the south Gyeongnam province in South Korea. The fire departments in Gyeongnam province has not enough workforce and equipment, and the area is vulnerable to fire [2]. Gyeongnam has the third highest frequency of fire occurrence. Besides, compared to the population, there were many casualties in five years.

By analyzing the factors affecting the fire, we suggest fire occurring patterns to manage the factors to prevent or reduce the likelihood of fire. Using machine learning techniques such as decision trees, random forest, artificial neural networks, XGBoost, we build fire prediction models and perform a variety of comparative analysis to find the most appropriate model. The models in this study are implemented by the python language.

Article history:
Received (March 13, 2020), Review Result (April 18, 2020), Accepted (May 21, 2020)

Print ISSN: 1738-9984, eISSN: 2208-9802 IJSEIA

2. Related studies

Many cities are densely populated, and there exists the potential for large-scale urban fires. Large fires are becoming an important research area across the world [3]. Researchs on the analysis and assessment of building fire have been conducted. Fire risk of residential building was analyzed based on scenario cluster and its application in fire risk management to define the level of fire risk and to determine whether to take appropriate risk management measures or not [4]. Fires at larger marketplace buildings were analyzed, and a fire risk assessment system was established [5].

Researches on structural fires and how to potentially prevent the fire of structures in such fires are far less performed than other areas of safety science research. This study views that large fire spread is incredibly complex, involving the interaction of temperature, humidity, climate, materials, and structures. Furthermore, it is difficult to model the prediction of a fire mathematically. It has been an interest in the prediction of fire using artificial intelligence. Atlanta Fire Department used artificial attributes rather than natural attributes to build the fire prediction model by using machine learning techniques such as decision trees, SVM, artificial neural networks, and XGboost [6]. They investigated the fire risk factors of the buildings and put the priority level of the risk factors. The model built by SVM shows the best accuracy, 75%. A joint study conducted by Carnegie Mellon University and the New York City Fire Department developed a model using machine learning techniques, logistic regression, ada boost, XGboost, and random forest techniques to find a building that has a high probability of fire [7]. New York City conducted a fire inspection every year, but they were unable to inspect all buildings due to an insufficient workforce. They tried to focus on predicting structures with high fire risk by using artificial attributes such as injury accident data, the number of buildings adjacent to the location of the building, violations of the tax and health and hygiene laws, tax payment history, complaint report phone history, the material of the construction rather than natural attributes.

However, no research on the fire prediction model, considering both the natural environment and artificial factors, has been published. In this study, we consider both natural attributes such as temperature, humidity, climate, etc. and artificial attributes such as the number of adjacent buildings, the distance between buildings, building materials, etc. This study aims to provide knowledge that can help prevent fires by discovering patterns or factors that affect the environment susceptible to fire. It is not known that a specific learning algorithm is superior to other algorithms in all industrial problems in the performance like precision, recall, F1-score, etc. This study finds fire prediction models by setting the parameters based on the problem to solve and data to apply and to do a direct comparison analysis.

3. Data and learning algorithms

3.1. Data and attributes

This study uses data on building fires in Gyeongnam province, South Korea, for five years, 01 January 2014 to December 2018. The raw data is collected as a digital form of csv files from the database of LH compass. The table in csv file consists of 59,199 rows and 180 columns. Of the 59,199 data, the number of data that fire has occurred is 7657. This indicates a more proportion of the amount of data that did not fire occur is biased in a ratio of 1:7. In this study, under-sampling is performed to mitigate somewhat the bias of the data for comparative evaluation of the models. The total number of data is 43,741 by randomly

extracting 7657 data from the fire and the remaining non-fire data. Most of the data attributes are prepared based on artificial factors such as distance from tobacco retail stores, distance from the non-smoking area, etc. Many missing values are pre-treated in such a way as to delete the rows or add appropriate value. Additionally, we remove data attributes that are not meaningful, and that can distort the results.

Attributes include discrete type such as steel-concrete structure and lightweight steel structure, and continuous types such as distance, temperature, and humidity. One hot encoding is performed using an algorithm that cannot handle categorical attributes. Data expressed in strings such as building structure and building purpose was arranged into each column by one-hot encoding, and numeric data such as land area, electric energy usage, gas energy usage was used as-is. Additionally, we delete data attributes that are not meaningful, and that can distort the results. The attributes that affect the building fire are confirmed through interviews with firefighters and past studies and are selected according to their importance. The importance of attributes is computed as the normalized total reduction of the criterion brought by that attribute. It is also known as the Gini importance. Only those with a relative importance of over 15% are used. The relative importance of the features is shown in [Figure 1].

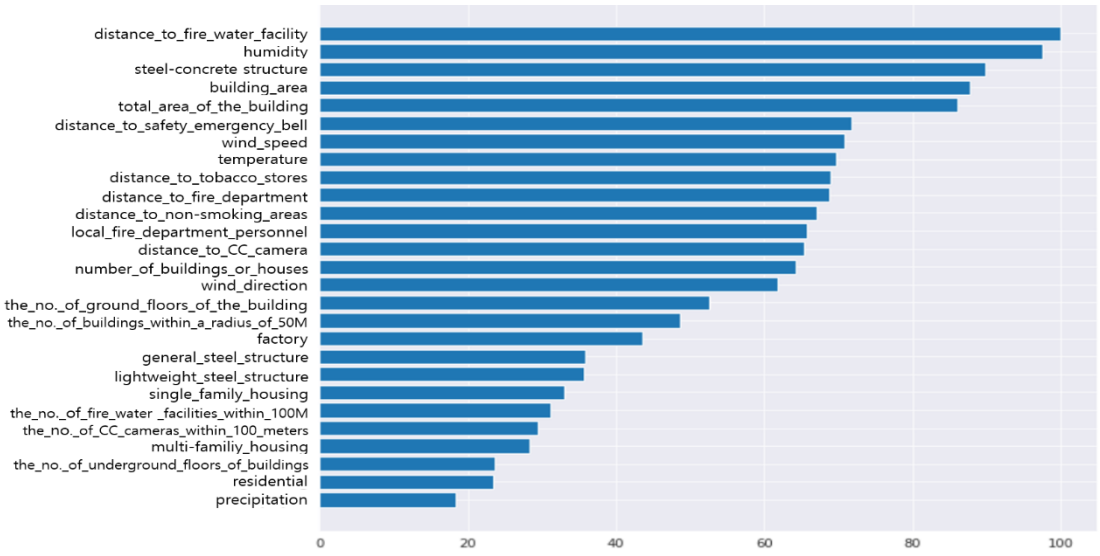

Figure 1. Attribute importance

Most of the attributes prepared is based on artificial factors such as distance from tobacco retail stores, distance from the non-smoking area, etc.

Humidity: In general, fires occur more frequently and significantly in dry winters with low humidity. The higher the humidity in the summer, the less likely it is to cause a fire. As a result, humidity can also be seen to have a significant effect on the fire.

Temperature: Temperature is natural environmental conditions that affect the fire, just like humidity. In the summer, when the temperature rises, fires such as vehicle fires and factory fires due to overheating are frequent. However, it is not possible to fire more frequently in other seasons. This is because other factors, such as humidity and wind speed, are affected

Material_of_architectural_frame: The frame of the building includes a steel-concrete structure and wooden structure and brick structure. It is commonly known that wooden

structures are vulnerable to fire, but the reinforced structure is the lowest in fire resistance. As a result, as well as accessible to fire, it can drop rapidly in durable once it occurs [8].

Distance_to_non-smoking_area: The closer the distance from the non-smoking area, the less the occurrence caused by cigarette butts or cigarette fires that cause the fire because the surrounding area is a non-smoking area. Experts say that fires caused by cigarette fires are one of the most frequent factor of fire. Hence, it can be seen that the distance to the non-smoking area has a significant association with the fire.

Distance_to_tobacco_stores: The distance from the tobacco retail store is also inferred in the same sense as a non-smoking area. The closer you are to the non-smoking area, the fewer cigarette butts or cigarettes will cause less fire. At this point, you can deduce that the distance from the retail store also affects the same context [9].

Distance_to_CC_camera: There may be cases where a person is set on fire with bad intentions, or a fire may occur due to a cigarette fire. In these cases, if the CC camera is installed at close range, it will be able to be a preventive effect of the fire to some extent [10].

Distance_to_fire_water_facility: The water facility supplies water to fire trucks. If there are enough facilities to provide water to the fire truck, it will be possible to supply water to the fire truck in the fire efficiently. The fire inside the building can also be quickly extinguished early in the fire. It can be seen that it has a high level of preventive facilities that are well equipped with a fire extinguisher

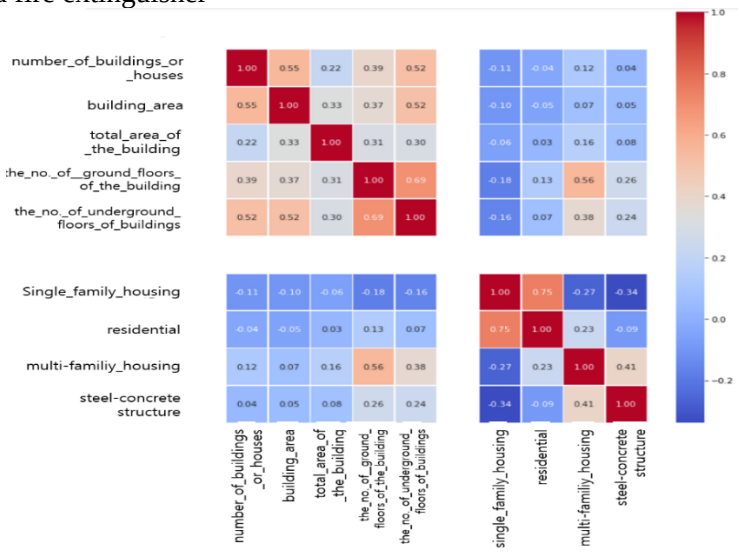

Figure 2. Correlation between features

A correlation is a number between -1 and +1 that measures the degree of association between two attributes (call them X and Y). In this case, large values of X tend to be associated with large values of Y, and small amounts of X tend to be associated with low values of Y. The correlation between single-family housing and residential is about 0.75 since the single-family house is commonly used as residential buildings in Figure 2. A positive value for the correlation implies a positive association. The number of ground floors of the building and the number of underground floors of the building is also shown to be 0.69. The number of underground floors of the building, the number of buildings or_houses, and the building_area are also correlated with a correlation value of 0.55.

We find that the larger the number of buildings, the greater the building area because it is a large building. In the case of the use of steel-concrete structure and multi-family housing,

there is a high correlation. Therefore, based on the above analysis, we simplified the use of attributes.

3.2. Algorithms

The prediction models can be built by algorithms such as SVMs, neural nets, logistic regression, naive-Bayes, memory-based learning, forests random, decision trees bagged trees, boosted trees, and boosted stumps, show good results in accuracy, reproducibility, and precision [11]. In this study, fire prediction models are built using machine learning algorithms. The data prepared in this study is applied to a decision tree, an extremely randomized tree classifier, random forest, neural network, and XGboost.

Decision tree: The decision tree is suitable for creating classification models. It is easy to organize decision rules into groups of interest and classify into several subgroups from the decision tree. Since the analysis process is expressed by the tree structure, users could easily understand the process. In general, it is sensitive to the sample size since the sample becomes smaller as the subnode progresses [12]. This study should handle the various attributes of the data. Of the data mining techniques, the decision tree is considered to be appropriate to process the multiple attributes of the data. The decision tree is very suitable for creating classification models.

Random forest: Random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. It has known that Random Forests significantly outperform Tree Bagging and other random tree ensemble methods in terms of accuracy [13].

Extremely randomized tree classifier: Extremely randomized tree classifier is a further development from the random forest, which performs similarly to the random forest [14]. However, there are performance differences that, namely decision trees show high variance, random forests show medium variance, and extra trees show low variance.

XGboost is a method that is often used in the creation of predictive models in recent years as a gradient boosting algorithm that puts forward parallel processing and optimization as an advantage [15]. It is used only to produce optimal results with a classification model that focuses on performance and speed. It also has the advantage of being able to add additional learning data to a model that has already been trained once.

Neural network: The neural networks are simple models of the way the nervous system operates. The basic units are neurons, which are typically organized into layers. It works by simulating a large number of interconnected processing units that resemble abstract versions of neurons. The neural network is composed of nodes and layers similar to human neurological neurons, and the node refers to the essential elements of the neural network model [16]. This technique is used to build models that are necessary to recognize the unique patterns or structures in the data. It is convenient to create a model for not knowing exactly whether the data has a linear relationship and a non-linear relationship or having two characteristics. Also, it is more suitable than other techniques for dealing with noisy, incomplete, and inconsistent data. To create a useful neural network model, it needs training data that includes a wide range of attributes. The model requires a precise setting based on the optimized conditions because of the technique that the result value can vary greatly depending on how the number of hidden layers and neurons is set. Neural networks in this study consist of two hidden layers, where the first hidden layer includes 11 nodes, and the

second does two nodes. It uses Adam optimizer to optimize weight and also the Relu function as an activation function.

4. Analysis of models

4.1. The patterns according to the fire risk level

The probability of a fire occurring can be obtained by using a multi-layer perceptron that extracts the probability of a fire in the multi-layer classifier model [17]. For a further understanding of this analysis, it was classified as one group with a high probability of a fire, more than 0.9, and the other group with a low probability, less than 0.1, that is less likely to cause fires. [Figure 2] shows a simple example of the data distribution according to the multi-family housing, steel-concrete frame, the fire risk level as a tree.

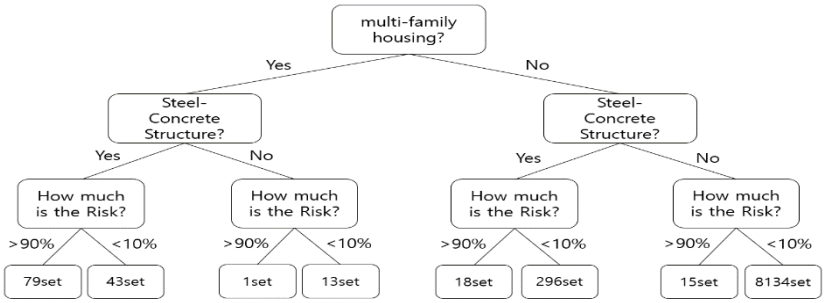

Figure 2. The distribution of data according to the fire risk level

Of the 113 buildings, according to a high probability of a fire, nearly 70.8% of them is multi-family housing, and 74.6% is a steel-concrete frame of the building. Of the 8430 cases for the low-risk structures, multi-family housing is more than 99%, and a steel-concrete frame is more than 96%. We find that the risk of fire is high if it is a steel-concrete frame building and multi-family housing.

4.2. Performance of fire prediction models

K fold cross-validation is performed to build a predictive model in this study. It has the advantage of making a more accurate estimate of out-of-sample accuracy since every observation is used for both training and testing. We use cross-validation and build K different models, so we are able to make predictions on all of our data.

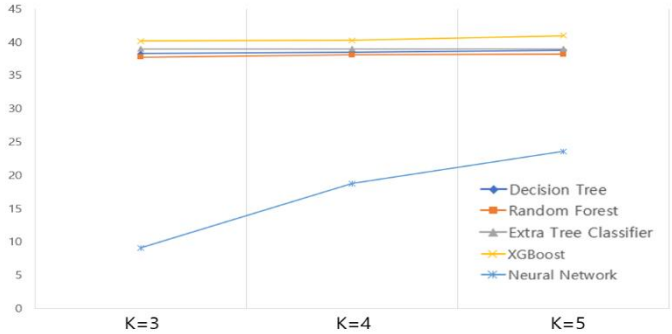

Figure 3. F1-score comparison

In **[**Figure 3**]**, F1 score is a measure of a test's accuracy in a statistical analysis of binary classification. The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples [18].

Table 1. Comparison of the model performance (%)

Algorithm	Decision tree	Random forest	Extra tree classifier	XGboost	Neural network
Accuracy	83.8	89.3	88.8	89.2	89.1
Precision	63.9	91.2	87.1	88.1	85.6
Recall	65.1	55.1	58.3	57.0	59.3
F1 Score	62.6	68.8	68.9	69.8	70.1

[Table 1] shows the accuracy and the F1-score of predictive models with five machine learning algorithms. The prediction model built with the random forest is the most accurate, but there is a little difference in the accuracy of each model trained with the extra tree classifier, XGboost, and neural network. For the F1 Score, the model with a neural network shows the best value, 0.701.

5. Conclusion

In this study, we built fire prediction models by using machine learning algorithms and compared the performance and results of the models. Key attributes affecting the fire were identified.

In the future study, we would like to use this study as follows: The first is to develop applications for firefighters that utilize natural conditions such as humidity, temperature, wind_direction, wind_speed, and precipitation. All of the natural conditions used here are based on the attributes that cause the fire. First, the real-time data containing the natural conditions received from the Meteorological Agency Open API are stored in the database in the application, and the database is linked to the fire prediction module. This will reduce careless behavior as much as possible on fire-prone days. Besides, firefighters can thoroughly prepare equipment for the outing on high-risk days or replenish tribal personnel. This will be able to prevent fires and minimize the damage in the event of an outbreak.

Second, this study can provide necessary information on building a fire prevention system through building risk classification. If you focus on building at high fire risk, you will be able to achieve a more significant effect.

Acknowledgments

This study is supported by Soongsil University.

References

[1] M. S. Won, K. S. Koo, and M. B. Lee, “An analysis of forest fire occurrence hazards by changing temperature and humidity of ten-day intervals for 30 years in spring,” Korean Journal of Agricultural and Forest Meteorology, vol.8, no.4, pp.250-259, **(2006)**

[2] http://www.nfa.go.kr/nfa/releaseinformation/statisticalinformation/main/=view&cntId=20&category=&pageIdx=&searchCondition=&searchKeyword, March 5, **(2020)**
[3] S. L. Manzello, R. Blanchi, M. J. Gollner, D. Gorham, S. McAllister, E. Pastor, E. Planas, P. Reszka, and S. Suzuki, "Summary of workshop large outdoor fires and the built environment," Fire Safety Journal, vol.100, pp.76-92, **(2018)**
[4] J. Xin, and C. Huang, "Fire risk analysis of residential buildings based on scenario clusters and its application in fire risk management," Fire Safety Journal, vol.62, Part A, pp.72-78, **(2013)**
[5] J. Yang and Y. Chen, "Research and application of fire risk assessment system for marketplace buildings," Procedia Engineering, vol.71, pp.476-480, **(2014)**
[6] M. Madaio, O. L. Haimson, W. Zhang, X. Cheng, M. Hinds-Aldrich, B. Dilkina, and D.H.P. Chau, "Identifying and prioritizing fire inspections: a case study of predicting fire risk in Atlanta," Bloomberg Data for Good Exchange, New York, NY, USA. **(2015)**
[7] M. A. Madaio, "Predictive modeling of building fire risk-designing and evaluating predictive models of fire risk to prioritize property fire inspections," A metro 21 Research Project, Carnegie Mellon University,13-27, **(2018)**
[8] O. Jungbluth, U.S. Patent No. 4,196,558. Washington, DC: U.S. Patent and Trademark Office, **(1980)**
[9] M. Ahrens, "Smoking and fire," American Journal of Public Health, vol.94, no.7, pp.1076-1077, **(2004)**
[10] S. Lu, G. Li, P. Mei, and H. Zhang, "Suppressive effects of fire prevention campaign in China: A time series analysis," Safety Science, vol.86, pp.69-77, **(2016)**
[11] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," In Proceedings of the 23rd international conference on Machine learning, June: 25-29; Pittsburgh, USA, **(2006)**
[12] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on Systems, Man, and Cybernetics, vol.21, no.3, pp.660-674, **(1991)**
[13] L. Breiman, "Random forests," Machine Learning, vol.45, no.1, pp.5-32, **(2001)**
[14] https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers 8507ac21d54b March 7, **(2020)**
[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, San Francisco, USA, August: 22-27, **(2016)**
[16] L. Yang, C. W. Dawson, M. R. Brown, and M. Gell, "Neural network and GA approaches for dwelling fire occurrence prediction," Knowledge-Based Systems, vol.19, no.4, pp.213-219, **(2006)**
[17] T. Windeatt, "Ensemble MLP classifier design," In Computational Intelligence Paradigms, Springer, **(2008)**
[18] https://en.wikipedia.org/wiki/F1_score, March 6, **(2020)**

Authors

Jaesung Chang
He is a student at Soongsil University, Department of Industrial Engineering. Areas of interest are machine learning, big data, and block chain.

Jaeyoung Yoon

He is a student at Soongsil University, Department of Industrial Engineering. Areas of interest are machine learning, Robotic Process Automation, and data analysis.

Gun Ho Lee

He received a Ph.D. degree from the Department of Industrial Engineering at the University of Iowa in 1996. From 1997 to present, he is a professor at Soongsil University in Korea. Areas of interest are machine learning, data mining, and artificial intelligence systems.

This page is empty by intention.