J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

공격 메일 식별을 위한 비정형 데이터를 사용한 유전자 알고리즘 기반의 특징선택 알고리즘☆

Feature-selection algorithm based on genetic algorithms using unstructured data for attack mail identification

홍 성 삼[1] 김 동 욱[1] 한 명 묵[1]*
Sung-Sam Hong Dong-Wook Kim Myung-Mook Han

요 약

빅 데이터에서 텍스트 마이닝은 많은 수의 데이터로부터 많은 특징 추출하기 때문에, 클러스터링 및 분류 과정의 계산 복잡도가 높고 분석결과의 신뢰성이 낮아질 수 있다. 특히 텍스트마이닝 과정을 통해 얻는 Term document matrix는 term과 문서간의 특징들을 표현하고 있지만, 희소행렬 형태를 보이게 된다. 본 논문에서는 탐지모델을 위해 텍스트마이닝에서 개선된 GA(Genetic Algorithm)을 이용한 특징 추출 방법을 설계하였다. TF-IDF는 특징 추출에서 문서와 용어간의 관계를 반영하는데 사용된다. 반복과정을 통해 사전에 미리 결정된 만큼의 특징을 선택한다. 또한 탐지모델의 성능 향상을 위해 sparsity score(희소성 점수)를 사용하였다. 스팸메일 세트의 희소성이 높으면 탐지모델의 성능이 낮아져 최적화된 탐지 모델을 찾기가 어렵다. 우리는 fitness function에서 *s(F)*를 사용하여 희소성이 낮고 TF-IDF 점수가 높은 탐지모델을 찾았다. 또한 제안된 알고리즘을 텍스트 분류 실험에 적용하여 성능을 검증하였다. 결과적으로, 제안한 알고리즘은 공격 메일 분류에서 좋은 성능(속도와 정확도)을 보여주었다.

☞ 주제어 : 보안, 비정형 데이터, 지능형 데이터 분석, 특징 선택, 공격 메일

ABSTRACT

Since big-data text mining extracts many features and data, clustering and classification can result in high computational complexity and low reliability of the analysis results. In particular, a term document matrix obtained through text mining represents term-document features, but produces a sparse matrix. We designed an advanced genetic algorithm (GA) to extract features in text mining for detection model. Term frequency inverse document frequency (TF-IDF) is used to reflect the document-term relationships in feature extraction. Through a repetitive process, a predetermined number of features are selected. And, we used the sparsity score to improve the performance of detection model. If a spam mail data set has the high sparsity, detection model have low performance and is difficult to search the optimization detection model. In addition, we find a low sparsity model that have also high TF-IDF score by using *s(F)* where the numerator in fitness function. We also verified its performance by applying the proposed algorithm to text classification. As a result, we have found that our algorithm shows higher performance (speed and accuracy) in attack mail classification.

☞ keyword : Security, Unstructured Data, Intelligent Data Analysis, Feature Selection, Attack Mail

1. Introduction

A Big data refers to a very large amount of data and includes a range of methodologies, such as big data collection, processing, storage, management, and analysis. In particular, text mining of unstructured big data, which has recently been utilized in many industries, is an important unstructured-data analysis technique.

Text mining is likely to extract a larger number of terms (features) as the amount of data increases. Since big-data text mining extracts a large number of features and data, clustering and classification can result in high computational complexity and low reliability of the analysis results. In particular, a term-document matrix (TDM) obtained through text mining represents term-document features, but produces a sparse matrix. In a sparse matrix, useful information cannot

1 Department of Computer Engineering, Gahon University, Gyeonggi-do, 13120, Korea
* Corresponding author (mmhan@gachon.ac.kr)
[Received 7 September 2018, Reviewed 13 September 2018(R2 20 November 2018), Accepted 5 December 2018]
☆ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1C1B3008718)
☆ A preliminary version of this paper was presented at APIC-IST 2017 seminar and was selected as an outstanding paper.

be retrieved and the analysis results cannot be trusted. Therefore, various studies have been conducted on feature selection and data dimensions [1].

This study focuses on selecting a set of optimized features from the corpus. A genetic algorithm (GA) is used to extract terms (features) as desired, according to the term importance calculated by the equation found. The study revolves around a feature-selection method that lowers the computational complexity and increases the analytical performance. And we have improved FSGA(Feature Selection based on Genetic Algorithm) that is proposed in [1].

We designed an advanced GA to extract features in text mining for detection model. Term frequency inverse document frequency (TF-IDF) [2] is used to reflect the document-term relationships in feature extraction. Through a repetitive process, a predetermined number of features are selected. And, we used the sparsity score to improve the performance of detection model. If a spam mail data set has the high sparsity, detection model have low performance and is difficult to search the optimization detection model. In addition, we find a low sparsity model that have also high TF-IDF score by using $s(F)$ where the numerator in fitness function. We also verified its performance by applying the proposed algorithm to text classification.

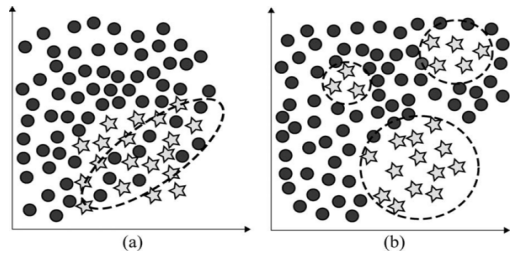
The rest of the paper is organized as follows. In Section 2, related work is introduced. In Section 3, feature-selection techniques using GAs in text mining are described. In Section 4, the experiment and analysis results are presented. Finally, the conclusions are described in Section 5.

2. Related Works

2.1 Intelligent Security Data Analysis

Intelligent security data analysis uses[2] intelligent methods and systems for analyzing [the] security data described above to improve the performance of security technologies, including attack detection, attack analysis, security assessment and vulnerability analysis and to enhance [the] security hardening concerning prediction of new attacks and defense against new attacks.

Intelligent methods or intelligent systems can be described as methods that enable searching, learning, analysis, and



(Figure 1) Type of Imbalance Data
(a) Class Overlapping (b) Small Disjunct

prediction based on knowledge, models, patterns, and features extracted from the collected data and the monitoring data obtained by computers. Intelligent systems can include data mining (classification, clustering, association analysis and ensemble, etc.), information fusion (Bayesian, fuzzy, etc.), soft computing (heuristic search, evolutionary programming, etc.) and artificial intelligence (AI) algorithms. In security systems, many studies have applied these intelligent systems to security technologies, since it is difficult to handle evolving-attack techniques with existing simple filtering-based methods and signature-based analysis methods. In particular, intelligent methods are perfect for research on intrusion detection (insider and outsider) and threat inference (threat assessment or risk assessment).

We targeted intelligent security systems; thus, we studied security systems based on data analysis. This paper introduces the concept of intelligent systems and intelligent-system approaches in intrusion detection and threat inference, on which our studies have focused, to implement intelligent security systems. An improved security system is proposed by applying GAs to intelligent systems. To apply intelligent algorithms for anomaly detection and intrusion reasoning, the following basic concepts and related studies are briefly introduced.

2.1.1 Security Data

It is necessary to define security data and their characteristics prior to the analysis of intelligent security data. Security data broadly means "structured and unstructured data that are collected by various sensors to ensure confidentiality, integrity, and availability of security from security attacks and

threats." From the perspective of attack defense, it can be narrowly defined as "all kinds of data that [are] used to detect security attacks and threats and to analyze their patterns, trends, and signs."

The sensors involved can be various hardware and software, such as host PCs, security software, and security systems, which include intrusion detection systems (IDS), firewalls (FW), and enterprise security management (ESM). These sensors collect and log data, mainly storing monitoring status, real-time conditions, and the analysis results. Each sensor contains many different forms of data.

Security data can be classified into the following two categories: data designed to be used for security purposes (e.g., IDS and FW blocking records) and data not created for security purposes but used for security data analysis (e.g., network packets or social network service (SNS) data). Insider attacks and threats have recently increased; thus, SNS, member records, and unstructured data such as CCTV footage, are also included in security data analysis [3]. With these data, studies on attack and threat analysis are being actively being carried out in the field of security data analysis.

2.1.2 Imbalance Problem in Security Data Analysis

These extensive security data sets contain a lot of imbalanced data, which can reduce the performance of learning algorithms and lead to incorrect predictions; therefore, they can degrade the detection rate of security systems and the analysis performance of algorithms. As shown in Figure 1 most data used for solving real-world problems contains imbalanced data, which cannot be used to build a good model when studying algorithms. In particular, security data usually contain considerable normal data with relatively little attack and anomaly data; therefore, security data is characterized imbalance

Data with a lower frequency are not classified into the normal category. In order to process and analyze these data in classification as follows, they must be balanced. That is, to improve poor classification accuracy it is necessary to understand the characteristics of the data frequencies and find effective data categories. It is also necessary to derive applicable and meaningful data through domain-specific data analysis. The necessity of handling these imbalanced data is currently recognized in fields related to data mining. There have been research activities to analyze imbalanced data as follows. The Association for the Advancement of Artificial Intelligence (AAAI) is devoted to applying imbalanced data to all industries involved in artificial intelligence (AI) algorithms that learn from mined data [4]. The International Conference on Machine Learning (ICML) is the leading international conference on machine learning from imbalanced data [4]. The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) focuses on extracting useful information from data and discovering knowledge for computer science [4].

The problem of imbalanced data is commonly considered to be important in the field of data processing for knowledge discovery. There are major challenges in learning from imbalanced data, including the evaluation of learning algorithms and the cost of imbalanced data [5, 6, 7, 8].

There have been many studies on improving conventional algorithms for use with imbalanced data sets. In recent years, there have also been studies on sampling from increasingly large data sets. Several of these have focused on generating a subset of data using a method relating to the selection of the most efficient features.

Security data that are divided into two types, normal and attack data, tend to be imbalanced because they are composed of data requiring dichotomous analysis, such as attack/detection or normal/abnormal (or anomaly) . However, there are not many studies aiming to solve the problem of imbalanced data in data mining; many existing security data analyses are based on data created for research. A significant achievement of solving data imbalance is the performance enhancement of ISDA. Feature selection techniques are used for solving data imbalance in ISDA, and in this paper, we propose a GA-based feature selection for this purpose.

2.2 Feature Selection for Intelligent System Design and Application

Feature selection is a method of selecting a subset of important or relevant features from an entire feature set to improve the analysis time and performance and reduce the dimensionality. Feature selection is mainly divided into wrapper approaches and filter approaches. Wrapper approaches

evaluate feature sets, based on their performance when classifying training data; they use a document classifier to select suitable features. The computational complexity of wrapper approaches is greater than that of filter approaches because they perform the classification of all candidate feature sets. However, the classification accuracy of wrapper approaches is better than that of filter approaches since the classifier properties are considered [9, 10].

Filter approaches evaluate candidate feature sets based on the intrinsic information of the dataset. The distance between data instances representing feature sets or information gain is measured to determine whether each feature should be selected. Filtering approaches are limited because it is impossible to consider the relationships between features because only the weights for independent single features are used for feature selection [9].

Feature selection plays an important role in anomaly detection since it is intended to eliminate unimportant or inappropriate features from all data features [11]. Feature selection usually enhances data generalization to make data more understandable, and it reduces computational complexity, data dimensionality, and redundancy to improve the performance of anomaly detection.

As shown in Figure 1, feature selection is largely made up of three steps: 1) sub-feature set generation, 2) sub-feature set evaluation, and 3) validation. The important part is the sub-feature set evaluation, which has five different approaches: score-based, entropy/mutual information-based, correlation-based, consistency-based, and detection accuracy-based [12].

3. AFSGA : Advanced FSGA

3.1 Fitness Function Studies and Design

The fitness function in GAs is the fitness equation used to evaluate the superiority of the given solution. In this study, the fitness function for text mining has been designed to evaluate the importance of the given term. The corresponding notation is as follows:

- $F = \{F_1, F_2, \ldots, F_n\}$ ≙ the set of Features(Chromosome)
- $D = \{D_1, D_2, \ldots, D_n\}$ ≙ the set of Documents
- N ≙ the number of documents
- s (F) ≙ sparsity of Document Term Matrix using F
- $x_i = F_i$ ≙ the value of ith Feature in ∈F
- tf_{ik} ≙ term frequence of feature $F_i \in$ F in document $D_k \in$ D
- df_k ≙ document frequence is the number of document included $F_k \in$ F
- idf_k ≙ inverse document frequence of feature $F_k \in$ F in document $D_k \in$ D = $\log((N - df_k)/df_k)$

Fitness function is shown equation (1),

$$\max F = \frac{\sum_{i=1}^{|F|}\sum_{k=1}^{|D|}(tf_{x_i k} \times idf_k)}{e^{s(F)}} \quad (1)$$

where

$$s(F) = 1 - \frac{\sum_{i=1}^{|F|}\sum_{k=1}^{|D|} f(x_{ik})}{N^* n - \sum_{i=1}^{|F|}\sum_{k=1}^{|D|} f(x_{ik})},$$

$$f(x)\begin{cases}1, & x = 0\\ 0, & x \neq 0\end{cases} \quad (2)$$

Regarding the importance of the given term, equation (1) does not simply apply its overall frequency, but uses its relative frequency with respect to each document and term to obtain the fitness value of each solution. Therefore, the relative importance can select the feature. When the frequency of the term is simply used, the term can occur frequently, but its meaning can be used differently in each document. However, this approach is not suitable for low-frequency terms, which may represent important features of the documents. Therefore, the fitness function using TF-IDF was designed by considering the Document-Term relationships.

And, we use the sparsity ratio *s(F)*(equation (2)) to improve the performance of detection model. Sparsity is ratio

of 0 value in data set(or matrix), and if a spam mail data set have the high sparsity, detection model have low performance and is difficult to search the optimization detection model. Therefore, we find a low sparsity model that have also high TF-IDF score by using *s(F)* where the numerator in fitness function.

Furthermore, we use the exponential function in numerator, it is reflected as an exponential function to prevent the value of 0, and the solving capacity for class imbalance is intended to be improved by increasing the importance of sparsity ratio by raising the degree of lowered total fitness as the sparsity ratio becomes higher.

3.2 Time Complexity and Fitness Curve

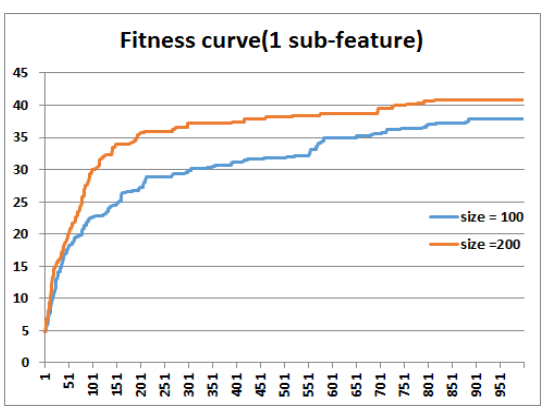
This algorithm has $O(t/l)$ time complexity, where t is total length of selected feature and l is length of sub-selected feature. The time complexity of typical GA is

$$O(n^2 * fitness\ function)$$

[13], where n is length of chromosome. In this algorithm, time complexity of fitness function is $O(d)$, where d is the number of documents. Finally, time complexity of proposal feature selection algorithm is

$$O(n \times d \times t) : (n^2 \times d) \times (t/l) = n \times d \times t$$

Figure 2 is shown Fitness curves of Proposal Algorithm



(Figure 2) Fitness Curve of Proposed Algorithm

in 1 sub-feature selection. To analyze the fitness curve in GA is possible to know level of optimization problem and algorithm performance. We find that optimization problem of spam mail detection model is a very difficult problem by once sub-feature searching because the best optimal detection model is not searched by algorithm while implement a lot of generations. However, we find that our algorithm will find the best optimal model after all, because fitness value is being convergence at about 800 generation and fitness value is more fast converged to increase a population size.

4. Experiment Result

4.1 Environment

The hardware and operating system environment used in the experiment is as follows.

- CPU: Intel Core i5 650 3.20 GHz
- RAM: 7 GB
- OS: Windows 7 Enterprise K 64bit

Open software R (version 3.02) [14] is a tool that was used for the experiment. In particular, the text mining (TM) package in R [15] was used for text mining. GA algorithms in the R GA package [16] were modified to run in the R environment. The classification algorithm used for the classification experiments is KNN Classifier [17].

The feature-selection experiment was performed using a TF-IDF GA. The clustering results were analyzed and compared with the original corpus. The GA parameters used are as follows:

- Size of population = 200
- Length of chromosome = 55
- Probability of crossover = 0.8
- Probability of mutation = 0.2

4.2 Document Data Set

The document data set used in the experiments includes 300 documents from the LingSpam Data Set [18], which is used for spam mail classification and clustering. The dataset was classified into two clusters: normal mail and spam mail. The clustering performance was measured. These experiments

used 252 normal mails and 48 spam mails.

4.3 Measuring Method of the Text Classification Experiment

In the evaluation step, the score function is applied to evaluate the performance using the feature set derived from the feature-selection method. A heuristic search is used repeatedly to find the desired feature set to meet the given criteria in the learning process until the final feature set is selected. The optimal feature set is chosen using evaluation and searching because a partial feature set is selected from the high-dimensional feature sets.

In this case, the F-Measure is used for the score regarding the document-classification evaluation. The F-measure is widely used to evaluate the results of text classification by applying precision and recall [19]. The precision Pi and the recall Ri are calculated by Definitions (3) and (4), respectively.

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$P_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$F_i = \frac{2 * P_i * R_i}{P_i + R_i} \tag{5}$$

where di indicates the number of documents included in category *i*. N is the number of categories.

4.4 Experiment Results

In class imbalance problem, there are measure methods for performance measure of detection algorithms ETS (Equitable Threat Score), CSI(Critical Success Index), PAG(Post Agreement), ACC(Accuracy)[20, 21, 22, 23, 24].There are usually methods to verify result of dichotomous (yes/no) forecasts on weather, detection, etc. There are according to some indicators in confusion matrix as below Table 1: (using indicators like True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) in F1-Measure).

(Table 2) Confusion Matrix

Forecast \ Label	normal	abnormal
normal	Hit(like TP)	False alarms (like FP)
abnormal	Misses (like FN)	Correct negatives (like TN)

$$ETS = \frac{H - a_r}{H + M + F - a_r}\ , \quad a_r = \frac{(H+M)(H+F)}{H+M+F+C} \tag{6}$$

$$CSI = \frac{H}{H+M+F} \tag{7}$$

$$PAG = \frac{H}{H+F} \tag{8}$$

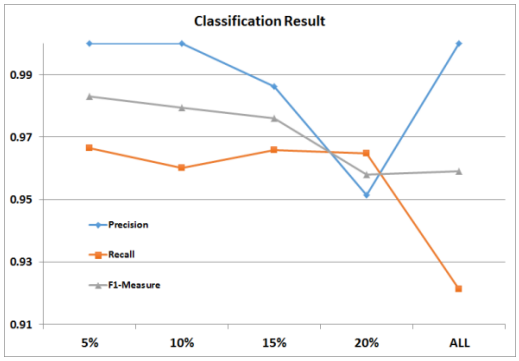
$$ACC = \frac{H+C}{H+M+F+C} \tag{9}$$

The results of the experiment show that, in the case where AFSGA was carried out, the best results were achieved when only 5% was selected for Precision, 50% for Recall, and 50% for F-measure, showing the values of 1, 0.9664, and 0.9829, respectively. When the results are compared with those achieved using all the features, the best performance was achieved when all the features at P = 1, R = 0.9213, F = 0.9590 were used in the cases of F-measure and Precision. On the other hand, a better result was achieved when AFSGA was used in the case of Recall.

An analysis of the overall experiment result shows that the higher the feature selection ratio is, the more the values of Precision, F-measure, and Recall improve in general, with the exception of several cases shown in Table 2, Figure 3.

(Table 3) Classification Result

Precision					
	5%	10%	15%	20%	ALL
KNN	1	1	0.9861	0.9513	1
Recall					
	5%	10%	15%	20%	ALL
KNN	0.9664	0.96	0.9659	0.9647	0.9213
F1-Measure					
	5%	10%	15%	20%	ALL
KNN	0.9829	0.9795	0.9759	0.958	0.9590



(Figure 3) Result of F1-measure(graph)

In the classification case, as learning is carried out through already-classified criteria, it can be seen that the more data is used for learning, the more the classification result improves.

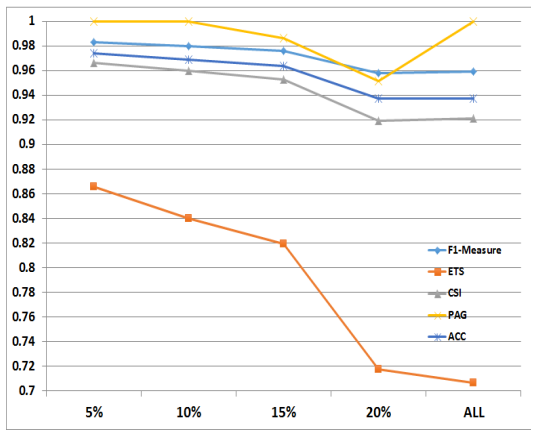
However, as the classification performance did not deteriorate very much, even when AFSGA was used, and a superior result was achieved depending on the selection ratio, AFSGA's superior performance and utilization potential could also be verified in document classification. However, to show better classification performance, AFSGA's performance should be enhanced.

The ETS/CSI/PAG/ACC results of the experiment show in Table 3 and Figure 4 that, in the case where FSGA was carried out, the best results were achieved when only 5% was selected showing the values of ETS = 0.8657, CSI = 0.9664, PAG = 1, and ACC = 0.9739, respectively. There is better performance than classification using all of features. ETS and CSI score is reflected as that algorithm has a performance to solve the imbalance class problem and to search the optimal forecast model. Therefore, the proposal algorithm has high-performance on spam mail detection.

Table 4 shows the experiment results of the classification time for each feature-selection ratio. Figure 5 shows a graph of the result. It can be seen that the higher the feature-selection ratio becomes, the more the performance time increases. Accordingly, when we see the results of the classification performance experiment and the results of the performance time, document classification using FSGA can be valuably used, depending on the purpose and the environment.

(Table 4) Result of ETS/CSI/PAG/ACC - Table

	5%	10%	15%	20%	All
ETS	0.8657	0.84	0.8193	0.7176	0.7063
CSI	0.9664	0.96	0.953	0.9194	0.9213
PAG	1	1	0.9861	0.9513	1
ACC	0.9739	0.9687	0.9635	0.9375	0.9375



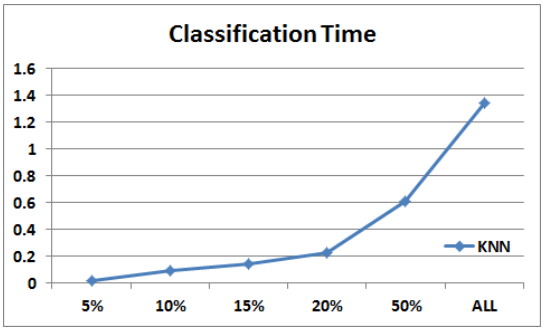
(Figure 4) Result of ETS/CSI/PAG/ACC(graph)

As the classification time can be greatly reduced while maintaining the document-classification performance to some extent within a usable level, it can be valuably utilized in a big-data environment, for real-time document classification, or where computing resources are insufficient. If the purpose is to accurately carry out classification, irrespective of the document-classification time, a better result may be achieved by using all of the features.

When the overall results of the document-classification experiment are considered, the proposed FSGA algorithm can be said to be efficiently usable for document classification.

(Table 5) A Result of Classification Time - Table

Classification Time(s)						
	5%	**10%**	**15%**	**20%**	**50%**	**ALL**
KNN	0.02	0.09	0.14	0.23	0.61	1.34



(Figure 5) Result of Classification Time

5. Conclusion

In this paper, a feature-selection (term selection) method was proposed to enhance the effectiveness of analysis in text mining. A new GA was designed for text mining, due to its optimal search performance. In addition, to maintain genetic diversity, the algorithm was modified to select the final feature set using partial feature sets.

And, we use the sparsity ratio s(F) to improve the performance of detection model. Sparsity is ratio of 0 value in data set(or matrix), and if a spam mail data set have the high sparsity, detection model have low performance and is difficult to search the optimization detection model. Therefore, we find a low sparsity model that have also high TF-IDF score by using S(F) where the numerator in fitness function.

In the document-classification experiment, the classification performance was shown to be better when all features were used than when AFSGA was used, with the exception of the Recall result. ETS and CSI score is reflected as that algorithm has a performance to solve the imbalance class problem and to search the optimal forecast model.

참고문헌(Reference)

[1] Sung-Sam Hong, Wanhee Lee, and Myung-Mook Han, "The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification," International Journal of Advance Soft Computing Application, Vol. 5, No. 3, 2013.

[2] Sung-Sam Hong, Dong-Wook Kim and Myung-Mook Han, "Feature-Selection Algorithm based on Genetic Algorithms for Intelligent Security Data Analysis of Unstructured Data," KSII The 12th Asia Pacific International Conference on Information Science and Technology(APIC-IST), Chiangmai, Thailand, 2017

[3] Daniel L. Costa, Matthew L. Collins, Samuel J. Perl, Michael J. Albrethsen, George Silowash, and Derrick Spooner, "An Ontology for Insider Threat Indicators," Proceedings of the 9th Conference on Semantic Technology for Intelligence, Defense, and Security, Fairfax VA, pp.48-53, 2014.

[4] He, Haibo and Edwardo Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, Vol.21, No.9, pp.1263-1284, 2009.
https://doi.org/10.1109/tkde.2008.239

[5] Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz, "Editorial: special issue on learning from imbalanced data sets," ACM SIGKDD Explorations Newsletter, Vol.6, No.1, pp.1-6, 2004.

[5] Eun-Jin Kim, Uk Heo, Byoung-Chul Kim, Il-Kyu Eom, and Young-In Kim, "More Realistic Data Generation for the Imbalanced Class Problem," Journal of Korean Institute Of Information Technology, Vol.9, No.11, pp.143-150, 2011.

[6] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol.42, No.4, pp.463-484, 2012.
https://doi.org/10.1109/tsmcc.2011.2161285

[7] Zhongbin Sun, Qinbao Song, Xiaoyan Zhu, Heli Sun, Baowen Xu, and Yuming Zhou, "A novel ensemble method for classifying imbalanced data," The Journal of the Pattern Recognition Society, Vol.48, No.5, pp.1623-1637, 2015.
https://doi.org/10.1016/j.patcog.2014.11.014

[8] Robertson, Stephen. "Understanding inverse document frequency: On theoretical arguments for IDF," Journal of Documentation, Vol.60, No.5, pp.503-520, 2004.

https://doi.org/10.1108/00220410410560582

[9] Joo-ho In, Jung-ho Kim, and Soo-hoan Cahe, “Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification,” Journal of Korean Society for Internet Information, Vol.14, No.5, pp 49-57, 2013.
https://doi.org/10.7472/jksii.2013.14.5.49

[10] John, G. Kohavi, and R. Pfleger, K., “Irrelevant Feature and the Subset Selection Problem”, In Proceedings of 11th International Conference on Machine Learning, New Brunswick, NJ, pp.121-129, 1994.
https://doi.org/10.1016/b978-1-55860-335-6.50023-4

[11] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, ″Network Anomaly Detection: Methods, Systems and Tools,″ IEEE COMMUNICATIONS SURVEYS & TUTORIALS, Vol.16, No.1, pp.303-336, 2014.
https://doi.org/10.1109/surv.2013.052213.00046

[12] J. Van Rijsbergen, 1979, Information Retrieval, second ed., Buttersworth, London

[13] http://www.r-project.org/

[14] http://cran.r-project.org/web/packages/tm/index.html

[15] http://cran.r-project.org/web/packages/GA/index.html

[16] htps://cran.r-project.org/package=e1071

[17] Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos, “An Evaluation of Naive Bayesian Anti-Spam Filtering″, 11th European Conference on Machine Learning (ECML 2000), Warsaw, Poland, pp. 9-17, 2000.

[18] Bratko, Andrej; et al. ″Spam filtering using statistical data compression models,″ The Journal of Machine Learning Research, No.7, pp. 2673-2698, 2006

[19] THOMAS M. HAMILL, and JOSIP JURAS, “Measuring forecast skill: is it real skill or is it the varying climatology?,” Quarterly Journal of the Royal Meteorological Society, Vol.132, No.621c, pp.2905-2923, 2006.
https://doi.org/10.1256/qj.06.25

[20] Roberts, N. M., and H. W. Lean, “Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events,” Monthly Weather Review, Vol.136, No.1, pp. 78-97, 2008.
https://doi.org/10.1175/2007mwr2123.1

[21] http://www.cawcr.gov.au/projects/verification/

[22] Nigro, M.A., J.J. Cassano and M.W. Seefeldt, “A weather-pattern-based approach to evaluate the Antarctic Mesoscale Prediction System (AMPS) forecasts : Comparison to automatic weather station observations,” Weather Forecasting, Vol.26, No.2, pp.184-198, 2011.
https://doi.org/10.1175/2010waf2222444.1

[23] Wilks, D.S., Statistical Methods in the Atmospheric Sciences. 3rd Edition. Elsevier, p. 676, 2011.
https://doi.org/10.1016/c2010-0-65519-2

◐ 저 자 소 개 ◑

홍 성 삼(Sung-Sam Hong)
2009년 가천대학교 전자거래학과 공학사
2011년 가천대학교 일반대학원 전자계산학과 공학석사
2016년 가천대학교 일반대학원 전자계산학과 공학박사
2016년~현재 : 가천대학교 컴퓨터 공학과 연구교수
관심분야 : Data Mining, Artificial Intelligence, Bigdata, Security, Feature Engineering
E-mail : sungsamhong0@gachon.ac.kr

김 동 욱(Dong-Wook Kim)
2015년 : 가천대학교 컴퓨터공학 공학사
2017년 : 가천대학교 일반대학원 컴퓨터공학 공학석사
2017년~현재 : 가천대학교 컴퓨터공학 박사과정
관심분야 : Data Mining, Artificial Intelligence, Anomaly Detection
E-mail : kog7306@naver.com

한 명 묵(Myung-Mook Han)
1980년 : 연세대학교 공과대학 공학사
1987년 : 뉴욕공과대학교 대학원 컴퓨터공학과 공학석사
1997년 : 오사카시립대학교 대학원 정보공학부 이학박사
1998년~현재 : 가천대학교 컴퓨터소프트웨어학과 교수
관심분야 : Information Security, Algorithm, Data Mining, Soft Computing, Bigdata
E-mail : mmhan@gachon.ac.kr