https://doi.org/10.33778/kcsa.2023.23.1.117

기계학습 기반 랜섬웨어 공격 탐지를 위한 효과적인 특성 추출기법 비교분석

김 한 석*, 이 수 진**

요 약

점점 더 고도화되고 있는 랜섬웨어 공격을 기계학습 기반 모델로 탐지하기 위해서는, 분류 모델이 고차원의 특성을 가지는 학습데이터를 훈련해야 한다. 그리고 이 경우 '차원의 저주' 현상이 발생하기 쉽다. 따라서 차원의 저주 현상을 회피하면서 학습모델의 정확성을 높이고 실행 속도를 향상하기 위해 특성의 차원 축소가 반드시 선행되어야 한다. 본 논문에서는 특성의 차원이 극단적으로 다른 2종의 데이터세트를 대상으로 3종의 기계학습 모델과 2종의 특성 추출기법을 적용하여 랜섬웨어 분류를 수행하였다. 실험 결과, 이진 분류에서는 특성 차원 축소기법이 성능 향상에 큰 영향을 미치지 않았으며, 다중 분류에서도 데이터세트의 특성 차원이 작을 경우에는 동일하였다. 그러나 학습데이터가 고차원의 특성을 가지는 상황에서 다중 분류를 시도했을 경우 LDA(Linear Discriminant Analysis)가 우수한 성능을 나타냈다.

Comparative Analysis of Dimensionality Reduction Techniques for Advanced Ransomware Detection with Machine Learning

Kim Han Seok*, Lee Soo Jin**

ABSTRACT

To detect advanced ransomware attacks with machine learning-based models, the classification model must train learning data with high-dimensional feature space. And in this case, a 'curse of dimension' phenomenon is likely to occur. Therefore, dimensionality reduction of features must be preceded in order to increase the accuracy of the learning model and improve the execution speed while avoiding the 'curse of dimension' phenomenon. In this paper, we conducted classification of ransomware by applying three machine learning models and two feature extraction techniques to two datasets with extremely different dimensions of feature space. As a result of the experiment, the feature dimensionality reduction techniques did not significantly affect the performance improvement in binary classification, and it was the same even when the dimension of featurespace was small in multi-class clasification. However, when the dataset had high-dimensional feature space, LDA(Linear Discriminant Analysis) showed quite excellent performance.

Keywords : Ransomware, Dimensionality Reduction, Machine Learning, Principal Component Analysis, Linear Discriminant Analysis

접수일(2023년 02월 28일), 수정일(2023년 03월 14일), 게재확정일(2023년 03월 30일)

* 국방대학교 국방과학학과 (주저자)
** 국방대학교 국방과학학과 (교신저자)

1. 서 론

‘몸값(ransom)’과 ‘소프트웨어(ware)’의 합성어인 랜섬웨어는 감염된 피해자의 정보통신기기를 공격하여 금전적 이득을 취하기 위한 목적으로 만들어졌다. 그러나 최근 랜섬웨어의 공격 대상이 국가의 공공기관과 중요 시설로 확장되면서 국가안보까지 위협하고 있다.

뉴질랜드의 사이버보안 기업 Emsisoft가 2023년 1월 발표한 보고서[1]에 의하면, 2022년도에 미국 지방정부를 대상으로 수행된 랜섬웨어 공격은 2021년 대비 약 30% 증가하였다. 국가정보원에서는 2023년 2월 10일 북한의 랜섬웨어에 대응하기 위해 ‘北 랜섬웨어 관련 韓美 합동 사이버보안 권고’를 발표하기도 하였다. 이 권고안에서는 한미 양국이 랜섬웨어 공격을 국가안보의 문제로 상정하였음을 확인할 수 있다.

랜섬웨어에 의한 위협이 증가한 만큼 랜섬웨어 탐지의 효율과 속도를 높이는 것이 중요한 이슈로 부각되고 있으며, 최근에는 기계학습을 기반으로 랜섬웨어 공격을 탐지하는 연구가 활발하게 진행되고 있다. 그러나 랜섬웨어는 더 많은 감염을 유도하기 위해 다양한 전략, 방법론 및 플랫폼을 채택하며 점점 더 고도화되고 있어 모델이 학습해야 하는 특성의 수 또한 점점 더 늘어나고 있다.[2]

기계학습 기반 탐지모델의 경우 학습해야 할 특성의 수가 늘어날수록 특성 공간의 차원이 함께 증가한다. 이는 함수의 최적화가 어려워지는 ‘차원의 저주(Curse of Dimensionality)’[3] 현상으로 이어질 수 있다. 따라서 모델의 탐지 성능을 향상하기 위해서는 학습 특성에 차원 축소기법을 적용하는 것이 필수적이다.

이에 본 논문에서는 차원 축소기법 중 특성 추출기법을 연구 범위로 하여 대중적인 특성 추출기법인 주성분 분석(Principal Component Analysis, PCA)[4]과 분류에 뛰어난 성능을 보이는 선형 판별 분석(Linear Discriminant Analysis, LDA)[5]을 비교하여 어떤 특성 추출기법이 랜섬웨어 탐지에 더욱 적합한 방법인지 비교 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 선행연구를 고찰하고 3장에서 특성 추출기법인 PCA와 LDA의 개념을 살펴본다. 4장에서는 데이터세트의 특성 크기에 따른 특성 추출기법의 성능을 비교하기 위한 실험을 제안하고 결과를 분석한다. 마지막 5장에서 결론을 맺는다.

2. 관련 연구

랜섬웨어 공격을 탐지하기 위한 연구는 꾸준히 이어지고 있다. [6]에서는 Locky 랜섬웨어 탐지를 위해 실시간 네트워크 정보를 학습 특성으로 하는 기계학습 기반 모델을 제안하였다. [7]에서는 랜섬웨어를 쿠쿠 샌드박스(Cuckoo Sandbox) 환경에서 동적으로 분석하여 랜섬웨어 탐지에 필요한 다양한 시그니처를 획득한 후, 특성 선택 기법을 통해 학습 차원을 축소하고 다수의 기계학습 알고리즘을 적용해 랜섬웨어 탐지를 수행하였다. [8]에서는 랜섬웨어의 시그니처를 N-Gram 시퀀스를 통해 유사도를 계산하는 방법으로 랜섬웨어를 탐지하는 방법을 제안했다. [9]에서는 동적분석을 통해 API Call 정보를 수집하고 이를 학습한 LightGBM 기반의 랜섬웨어 분류 모델을 제안하였다.

특성 추출기법을 활용하여 기계학습의 학습효율을 끌어올리기 위한 연구도 진행되었다. [10]에서는 IoT 기기에 대한 악성코드 공격을 탐지하기 위해 PCA 기법을 적용하였다. 실험 결과 학습 특성을 원본의 9% 수준으로 낮추더라도 탐지 성능에는 차이가 발생하지 않았다. [11]에서는 안드로이드 악성코드를 탐지하기 위해 PCA와 LDA를 서로 다른 기계학습 모델과 함께 사용했을 때 어떤 효과가 있는지 분석하였다. 실험 결과 LDA를 활용한 방식이 PCA보다 좋은 성능을 보였다. 그러나 이진 분류에 관한 연구만 진행했다는 한계가 있었다. [12]에서는 악성코드 분석을 위해 PCA와 LDA 기법을 사용하였고, 각 기법에 장단점이 있음을 확인하였다. 그러나 실험에 사용된 데이터의 특성이 41개로, 최근 기계학습에 사용되는 데이터의 학습 특성보다 훨씬 적은 숫자이기 때문에 최근 실태를 반영하지 못한다는 한계가 있었다.

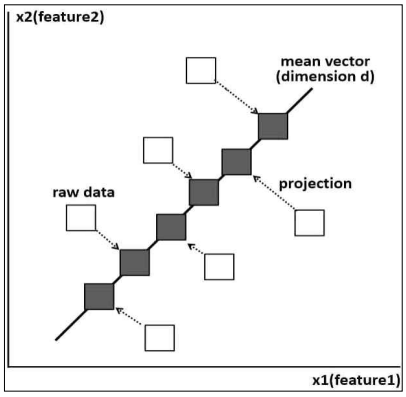
3. 특성 추출기법

3.1 주성분 분석 (PCA)

주성분 분석은 고차원의 데이터를 저차원의 데이터로 환원하는 기법이다. 고차원 공간에서 연관관계에 있는 표본들을 저차원 공간의 표본들로 변환하기 위하여 선형 및 직교 변환을 사용한다.

고차원 데이터의 차원을 감소시키기 위해 공분산 행렬을 이용하여 원래의 데이터 구조를 가장 잘 유지할 수 있는 벡터를 찾고, 해당 벡터에 표본들을 정사영(projection)시키는 방식으로 이루어진다. 즉 <그림 1>에서 보는 바와 같이, 기존 데이터의 분산을 최대한 보존하는 새로운 축을 찾고 그 축에 데이터를 정사영시키는 것이다. 주어진 데이터의 무수히 많은 표본 중 전체적인 분산을 가장 잘 설명해주는 성분을 주성분이라고 부른다.

PCA는 기존 데이터의 분산을 최대한 보존하는 새로운 특성 벡터를 찾는 기법이므로 새로운 특성의 신뢰도는 '특성이 기존 데이터의 분산을 얼마나 잘 설명할 수 있는지'로 이해할 수 있다. 이것을 수치로 나타낸 것을 '분산도'라고 한다.

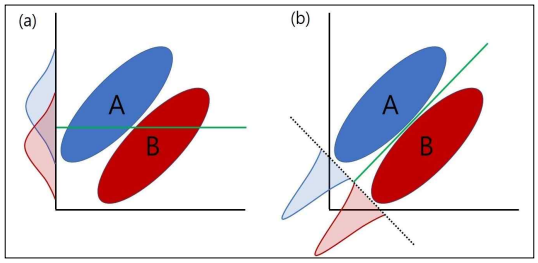


(그림 1) 주성분 분석의 차원 축소 원리

3.2 선형 판별 분석 (LDA)

LDA 역시 고차원의 데이터를 저차원의 데이터로 환원하는 기법이지만, PCA와 달리 LDA는 클래스를 기반으로 특성을 추출한다. LDA는 고차원 데이터의 차원을 감소시키기 위해 기존의 클래스를 기반으로 각 클래스 집단 사이의 평균 차이는 크고 집단 내부의 분산은 작게 구분할 수 있는 벡터를 찾고, 해당 벡터에 표본들을 정사영시키는 방식으로 특성을 추출한다.

<그림 2>의 (a)와 (b)를 비교하였을 때 (a)와 같이 축을 설정하면 각 클래스 내부의 분산이 높아지며 클래스 집단 사이의 평균 차이도 작아지지만, (b)와 같이 축을 설정하면 클래스 내부의 분산은 낮고 클래스 사이의 평균 차이는 높아진다. 새로운 차원의 축을 설정할 때 클래스가 사전에 구분되어 있어야 하므로 LDA는 지도학습으로 분류된다.



(그림 2) 선형 판별 분석의 차원 축소 원리

LDA도 분산도를 통해 새로운 특성의 신뢰성을 확인할 수 있다. 그러나 PCA와 달리 LDA는 클래스를 기반으로 특성을 추출하기 때문에 특성의 수를 클래스의 수보다 하나 적게 설정하면 분산도를 100%에 가깝게 확보할 수 있는 특징이 있다.

4. 실험 및 결과

4.1 데이터세트

본 연구에서는 <표 1>에서처럼 학습데이터의 특성 차원이 작은 경우와 큰 경우에서 이진 분류 및 다중 분류를 수행하여 PCA와 LDA가 차원의 크기에 따라 어떤 성능을 나타내는지 비교하였다.

'Feature Dynamic API' 데이터세트[10]는 8종의 랜섬웨어를 쿠쿠 샌드박스(Cuckoo Sandbox) 환경에서 동적 분석하여 랜섬웨어의 API Call을 수집하였다. 각 특성의 값은 API가 호출된 횟수이다.

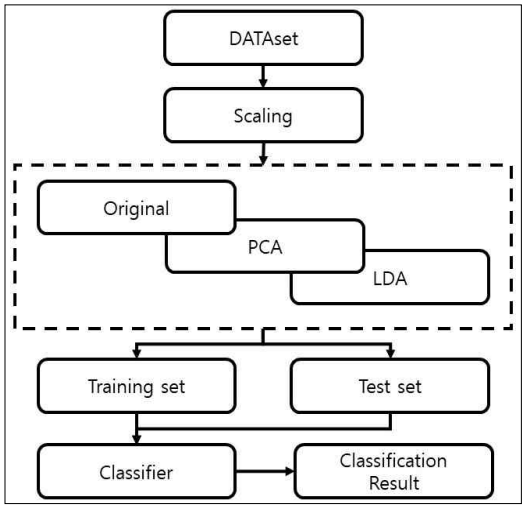
'Ransomware Dataset'[8]은 11종의 랜섬웨어를 샌드박스 환경에서 동적 분석하여 획득한 시그니처를 수집하였다. API Call, 레지스트리 호출, 파일 실행 등의 정보를 학습 특성으로 포함하고 있으며, 실행 여부를 1과 0으로 표시하였다.

<표 1> 실험에 사용된 데이터세트의 세부사항

Dataset	Feature Dynamic API	Ransomware Dataset
총 개체	5,811	1,524
정상파일	4,008	942
랜섬웨어	1,803	582
클래스	9	12
학습 특성	286	30,970

4.2 실험방법

각각의 데이터세트에 대하여 정규화를 수행한 뒤 PCA와 LDA를 통해 특성을 추출하여 학습 차원을 축소했다. 원본 데이터세트와 차원이 축소된 데이터세트를 각각 학습 데이터세트와 검증 데이터세트로 나누고 총 3가지 분류 모델에 적용하여 평가를 수행하였다. 실험과정은 <그림 3>에서 보는 바와 같다.



(그림 3) 실험 진행 과정

'Feature Dynamic API'의 값은 특성마다 범위가 다르므로 모델의 성능을 향상하기 위해서 Scaling 과정이 필요하다. 실험에서는 Min-Max 정규화 방식을 채택하여 데이터세트의 모든 특성을 0에서 1 사이의 값으로 변환하였다. Standardization 정규화도 실험하였으나 Min-Max 정규화 방식이 모델의 성능을 올리는 데 적합하였다.

실험에 사용할 분류 기법은 [9], [10]의 연구를 참고하여 LightGBM으로 선정하였으나 여러 가지 기계학습 기법을 동시에 비교한 [7], [11]의 연구 방법론을 채택하여 LightGBM과 같이 부스팅 알고리즘을 사용하는 XGBoost, CatBoost 기법까지 총 3개를 선정하였다.

추출하는 특성의 개수도 중요한 변수이다. PCA는 주성분의 개수를 조절하여 새로운 특성에 기존 데이터의 분산도를 반영할 수 있다. 본 연구에서는 PCA를 통한 분산도를 40%부터 시작하여 90%까지 5% 단위로 각각 반영하여 실험하였으며, 이 중 가장 좋은 성능을 보인 분산도 75%의 결과를 실험 결과로 선정하였다. LDA는 특성 개수를 클래스 수보다 하나 작게 설정하면 100%에 가까운 분산도를 가진다. 본 연구에서 LDA를 통해 얻은 새 데이터세트의 특성 개수(8개, 11개)는 PCA를 적용한 새 데이터세트의 특성 개수(48개, 66개)와 비교하였을 때 크게 작은 숫자이기 때문에 LDA는 분산도 100%를 기준으로 특성 개수를 선정하였다.

실험 환경은 <표 2>에서 보는 바와 같다.

<표 2> 실험 환경

운영체제	Windows 11 Pro
CPU	11th Gen Intel i5-11300H
RAM	40GB
개발언어	Python 3.10

4.3 실험 결과 및 분석

실험은 이진 분류와 다중 분류를 모두 수행하였으나, 이진 분류의 경우 특성 차원 축소기법이 원본 특성을 모두 사용한 경우보다 유의미한 성능 향상 효과를 발휘하지는 못하였다. 이진 분류의 결과는 <표 3>과 같다.

<표 3> 각 데이터세트의 이진 분류 결과

데이터세트	최고 성능 조합	정확도	F1 Score
Feature Dynamic API	원본 - LightGBM	0.9957	0.9931
	PCA - LightGBM	0.9880	0.9808
	LDA - CatBoost	0.9751	0.9603
Ransomware Dataset	원본 - LightGBM	0.9836	0.9784
	PCA - XGBoost	0.9770	0.9692
	LDA - Catboost	0.9999	0.9999

따라서 본 절에서는 유의미한 성능 향상 효과가 확인된 다중 분류 결과만을 분석한다. 성능평가지표는 정확도와 F1 Score를 활용한다. 데이터세트에서 각 클래스의 샘플 개수가 불균형하므로, F1 Score를 계산하는 방법론 중 micro average를 적용한 방법을 채택하였다.

4.3.1 특성 차원이 작은 데이터세트

학습 특성이 적은 'Feature Dynamic API' 데이터세트는 원본을 기반으로 분류한 결과가 특성 추출기법을 적용했을 때보다 전반적으로 더 좋은 성능을 보였다. 그러나 각 모델끼리 비교했을 때, 특성 추출기법을 적용한 모델과 원본 기반 모델의 정확도와 F1 Score는 차이는 최대 1.63%p로 유의미한 차이라고 보기는 어려웠다. 같은 데이터세트를 사용한 기존 연구 [9]와 비교하였을 때도 성능 향상이 보이지 않았다. PCA와 LDA를 적용한 모델만을 비교하였을 때는 LDA를 적용한 모델의 성능이 근소하게 우수하였다. Dynamic API'의 실험 결과는 <표 4>에서 보는 바와 같다.

<표 4> Feature Dynamic API 다중 분류 결과

구분	모델	특성 수	정확도	F1 score
원본	XGBoost	286	0.7773	0.7773
	LigntGBM		0.9828	0.9828
	CatBoost		0.9768	0.9768
PCA	XGBoost	48	0.7730	0.7730
	LigntGBM		0.9665	0.9665
	CatBoost		0.9656	0.9656
LDA	XGBoost	8	0.7721	0.7721
	LigntGBM		0.9733	0.9733
	CatBoost		0.9751	0.9751

4.3.2 특성 차원이 큰 데이터세트

학습 특성이 많은 'Ransomware Dataset'는 모든 모델에서 LDA를 적용한 경우가 나머지 경우보다 좋은 성능을 보였다. 특히 LightGBM, CatBoost 기반 모델의 경우 원본 및 PCA 적용 데이터세트와 비교하여 정확도, F1 Score에서 13.77%p ~ 14.76%p 우수한 성능을 보였다. 같은 데이터세트를 사용한 기존 연구[7]에서 100개 특성을 선택한 실험 결과는 93.3%의 정확도를 보였으나 본 연구에서 LightGBM, CatBoost기반 모델에 LDA를 적용한 경우 약 1/10 수준인 11개의 특성 만을 가지고도 98.69%의 정확도를 보이며 약 5%p 이상 우수한 성능을 나타냄을 확인하였다.

이를 통해 특성 차원이 큰 랜섬웨어를 분류할 때는 클래스에 대한 이해가 없이 특성을 추출하는 PCA보다 기존 데이터세트의 클래스를 기반으로 특성을 추출하는 LDA가 더욱 적합하다는 것을 알 수 있었다. 'Ransomware Dataset'의 실험 결과는 <표 5>에서 보는 바와 같다.

<표 5> Ransomware Dataset 다중 분류 결과

구분	모델	특성 수	정확도	F1 score
원본	XGBoost	30,970	0.6393	0.6393
	LigntGBM		0.8754	0.8754
	CatBoost		0.8459	0.8459
PCA	XGBoost	66	0.6426	0.6426
	LigntGBM		0.8492	0.8492
	CatBoost		0.8393	0.8393
LDA	XGBoost	11	0.6426	0.6426
	LigntGBM		0.9869	0.9869
	CatBoost		0.9869	0.9869

5. 결 론

랜섬웨어 탐지를 위해 기계학습 기반 탐지모델이 학습해야 하는 특성 차원의 크기는 점점 커지고 있으나, 특성 차원이 커지면 모델의 성능이 저하될 수 있으므로 차원 축소가 선행되어야 한다.

본 논문에서는 기계학습 기반 랜섬웨어 공격 탐지 모델의 성능을 향상하기 위해 PCA와 LDA중 어떤

특성 추출기법이 랜섬웨어 탐지에 더욱 적합한지 비교 분석하였다. 기계학습 기반 모델에서 주로 사용되는 주성분 분석과 분류에 좋은 성능을 보이는 것으로 알려진 선형 판별 분석 중 어떤 기법이 랜섬웨어 공격 탐지에 더욱 적합한지 알아보기 위해 특성 수에 큰 차이가 있는 두 개의 랜섬웨어 데이터세트에 각각의 특성 추출기법을 적용한 후 XGBoost, LightGBM, CatBoost 기계학습 모델을 기반으로 랜섬웨어 공격 탐지를 시도하였다.

이진 분류에서는 특성 차원의 크기와 상관없이 특성 추출기법이 유의미한 성능 향상으로 이어지지 않았다. 다중 분류에서는 특성이 286개인 'Feature Dynamic API' 데이터세트의 경우 특성을 추출하지 않고 원본 데이터세트를 그대로 이용한 모델이 가장 우수한 성능을 보였다. 그러나 PCA, LDA를 적용한 모델과의 정확도와 F1 Score의 차이는 최대 1.63%p로 유의미한 차이라고 보기 어려웠다.

특성이 30,970개인 'Ransomware Dataset'의 경우 LDA를 적용한 모델이 우수한 성능을 보였다. 원본, PCA를 적용한 모델과 비교하였을 때 정확도와 F1 Score는 최대 14.76%p 우수하였다.

실험을 통해 학습 특성의 수에 따라 차원 축소를 적용하지 않는 것이 오히려 모델의 성능을 향상할 수 있으나, 특성의 수가 아주 크다면 클래스와 관계없이 특성을 추출한 PCA보다는 기존 데이터세트의 클래스를 기반으로 특성을 추출한 LDA를 적용하는 것이 모델의 성능을 높일 수 있음을 확인하였다.

향후 연구에서는 더욱 최신의 랜섬웨어를 자체적으로 수집하고 이를 동적으로 분석하여 랜섬웨어 데이터세트를 구축한 후, LDA를 적용한 모델을 연구해 나갈 예정이다.

참고문헌

[1] Emisoft, "The State of Ransomware in the US: Report and Statistics 2022 (2023.1.2)", Retrieved Feb. 11, from: https://www.emsisoft.com/en/blog/43258/the-state-of-ransomware-in-the-us-report-and-statistics-2022/.

[2] Korea Anti Ransomware Alliance, "KARA 랜섬웨어 동향 보고서(2022.9.20), Retrieved Feb. 11, from: https://www.skshieldus.com/kor/support/download/report.do.

[3] Donoho, D. L., "High-dimensional data analysis: The curses and blessings of dimensionality.", AMS conference on Math Challenges of the 21st Century, pp.1-32, 2000.

[4] Hotelling, H., "Analysis of a complex of statistical variables into principal components.", Journal of Educational Psychology, 24(6), pp.417-441, 1933.

[5] Fisher, R. A., "The use of multiple measurements in taxonomic problems.", Annals of Eugenics, 7(2), pp.179-188.

[6] Almashhadani, Ahmad O., et al. "A multi-classifier network-based crypto ransomware detection system: A case study of locky ransomware." IEEE access 7, pp.47053-47067, 2019.

[7] Sgandurra, D., Munoz-Gonzalez, L., Mohsen, R., and Lupu, E. C., "Automated Dynamic Analysis of Ransomware: Benefits, Limitations and use for Detection.", arXiv preprint arXiv:1609.03020, 2016.

[8] Gyu Bin Lee, Jeong Yun Oak, Eul Gyu Im, "Method of Signature Extraction and Selection for Ransomware Dynamic Analysis.", KIISE Transactions on Computing Practices, 24(2), pp.99 - 104., 2019.

[9] Nguyen Duc Thang, Soojin Lee. "LightGBM-based Ransomware Detection using API Call Sequences." International Journal of Advanced Computer Science and Applications 12.10, 2021.

[10] Ji-Gu Lee, Soo-Jin Lee, "IoT Attack Detection Using PCA and Machine Learning.", Proceedings of the Korean Society of Computer Information Conference 30(2), pp. 245-246, 2022

[11] Şahin, D. Ö., Kural, O. E., Akleylek, S., & Kılıç, E. "Permission-based Android malware analysis by using dimension reduction with PCA and LDA." Journal of Information Security and Applications, 63, 102995, 2021.

[12] Datti, R., & Lakhina, S., "Performance comparison of features reduction" techniques for intrusion detection system. vol, 3, 4., 2012.

〔저자소개〕

김 한 석 (Han-Seok Kim)
2014년 3월 육군사관학교 학사
2021년 3월 ~ 현재
국방대학교 국방과학학과 석사 과정

email : 14.10083a@gmail.com

이 수 진 (Soojin Lee)
1992년 3월 육군사관학교 학사
1996년 2월 연세대학교 석사
2006년 2월 한국과학기술원 박사
2006년 3월 ~ 현재
국방대학교 국방과학학과 교수

email : cyberkma@gmail.com