KIPS Trans. Softw. and Data Eng.
Vol.10, No.2 pp.65~72
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.2.65

Improved Network Intrusion Detection Model through Hybrid Feature Selection and Data Balancing

Byeongjun Min† · Jihun Ryu† · Dongkyoo Shin†† · Dongil Shin††

ABSTRACT

Recently, attacks on the network environment have been rapidly escalating and intelligent. Thus, the signature-based network intrusion detection system is becoming clear about its limitations. To solve these problems, research on machine learning-based intrusion detection systems is being conducted in many ways, but two problems are encountered to use machine learning for intrusion detection. The first is to find important features associated with learning for real-time detection, and the second is the imbalance of data used in learning. This problem is fatal because the performance of machine learning algorithms is data-dependent. In this paper, we propose the HSF-DNN, a network intrusion detection model based on a deep neural network to solve the problems presented above. The proposed HFS-DNN was learned through the NSL-KDD data set and performs performance comparisons with existing classification models. Experiments have confirmed that the proposed Hybrid Feature Selection algorithm does not degrade performance, and in an experiment between learning models that solved the imbalance problem, the model proposed in this paper showed the best performance.

Keywords : Intrusion Dectection, Deep Learning, Over Sampling, Feature Selection

Hybrid Feature Selection과 Data Balancing을 통한 효율적인 네트워크 침입 탐지 모델

민 병 준† · 유 지 훈† · 신 동 규†† · 신 동 일††

요 약

최근 네트워크 환경에 대한 공격이 급속도로 고도화 및 지능화 되고 있기에, 기존의 시그니처 기반 침입탐지 시스템은 한계점이 명확해지고 있다. 이러한 문제를 해결하기 위해서 기계학습 기반의 침입 탐지 시스템에 대한 연구가 활발히 진행되고 있다. 하지만 기계학습을 침입 탐지에 이용하기 위해서는 두 가지 문제에 직면한다. 첫 번째는 실시간 탐지를 위한 학습과 연관된 중요 특징들을 선별하는 문제이며, 두 번째는 학습에 사용되는 데이터의 불균형 문제로, 기계학습 알고리즘들은 데이터에 의존적이기에 이러한 문제는 치명적이다. 본 논문에서는 위 제시된 문제들을 해결하기 위해서 Hybrid Feature Selection과 Data Balancing을 통한 심층 신경망 기반의 네트워크 침입 탐지 모델인 HFS-DNN을 제안한다. NSL-KDD 데이터 셋을 통해 학습을 진행하였으며, 기존 분류 모델들과 성능 비교를 수행한다. 본 연구에서 제안된 Hybrid Feature Selection 알고리즘이 학습 모델의 성능을 왜곡 시키지 않는 것을 확인하였으며, 불균형을 해소한 학습 모델들간 실험에서 본 논문에서 제안한 학습 모델이 가장 좋은 성능을 보였다.

키워드 : 침입 탐지, 딥 러닝, 오버샘플링, 특징 선택

1. 서 론

네트워크 침입 탐지 시스템(NIDS; Network-Based Intrusion Detection)은 허가되지 않은 사용자의 침입을 제한하는 시스템으로, 트래픽을 감시하여 공격 여부를 판단한다. 현재 많이 사용되고 있는 네트워크 침입 탐지 시스템은 시그니처(Signature) 기반 분석 방법을 사용하고 있다. 이는 공격에 대한 특징을 전문가가 분석하여 패턴화 시킨 뒤, 실시간으로 들어오는 네트워크 패킷들과 매칭하여 탐지하는 방법이다. 그러나 최근 APT (Advance Persistent Threat) 공격과 같이 빠르게 변화하는 공격이 빈번히 발생하고 있으며, 이에 따라 매번 발생하는 새로운 공격에 대한 트래픽과 로그 분석 과정에서 비용적 문제와 시간적 문제가 발생한다. 시그니처 패턴은 빠르게 변화하는 공격의 비슷한 유형에 대하여 일반화된 성능을 보장하지 못하기 때문에, 기존의 시그니처 기반의 시스템의 한

※ 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다(UD190016ED).
※ 이 논문은 2020년 한국정보처리학회 춘계학술발표대회의 우수논문으로 "Hybrid Feature Selection과 Data Balancing을 통한 네트워크 침입 탐지 모델"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 세종대학교 컴퓨터공학과 박사과정
†† 종신회원 : 세종대학교 컴퓨터공학과 교수
Manuscript Received : July 8, 2020
Accepted : August 9, 2020
* Corresponding Author : Dongil Shin(dshin@sejong.ac.kr)

계점이 명확해 지고 있다. 최근 이러한 문제를 해결하기 위해 기계학습 기반의 탐지 시스템의 연구가 활발하다[1-3, 5-7].

기계학습 모델은 데이터로부터 침입에 대한 판단을 내리기 위한 규칙을 모델이 학습하기에, 이를 통해 자동화된 침입탐지 시스템을 구축할 수 있다면, 앞서 언급한 시간 및 비용적 문제를 해결할 수 있다. 또한 새로운 공격 패턴들에 대해 일반화된 성능을 보장할 수 있다. 하지만 입력으로 사용할 수 있는 모든 속성들을 사용하는 것은 기계학습 모델의 성능의 저하와 학습 시간의 낭비를 가져올 수 있다[4]. 따라서 실시간 탐지 위해 사용 가능한 많은 속성 중에서 학습과 관련 있는 특징들을 선별하는 것도 중요하게 다뤄지고 있다[5]. 이 외에도 현실 세계에서 수집되는 많은 데이터들은 클래스간 균형이 완벽하지 않은 환경이 대부분으로, 특히 침입 탐지 문제에서는 전체 데이터 중 침입 데이터의 비율이 약 1%로 알려져 있다[6]. 기계학습에서 이러한 소량의 침입 데이터로 정상학습을 하는 것은 매우 어려우며[4, 5], 이를 극복하기 위한 방법으로 오버샘플링(Over Sampling)과 언더 샘플링(Under Sampling) 기법을 활용할 수 있다[7].

본 논문에서는 기계학습 모델에 불필요한 속성들과 중첩 속성들을 제거하기 위해서 HFS (Hybrid Feature Selection) 기법을 제안하며, 이를 심층 신경망에 이용한 HFS-DNN (Deep Neural Network) 모델을 제안한다. 기존의 많은 입력 속성 값들을 모두 학습에 사용하는 것과 달리 HFS 기법을 통해 32% 규모의 입력만을 사용해 동일한 학습 효과를 보장하는 것을 실험을 통해 검증하며, 학습에 사용된 NSL-KDD 데이터 셋의 불균형 문제로 인한 소수 클래스(Minor Class)들의 저조한 탐지율을 개선하고자 SMOTE (Synthetic Minority Over sampling Technique)[10]기법과 RUS (Random Under Sampling) 기법들을 활용하여 불균형 문제를 다루었다. 실험에 사용된 평가 지표는 Accuracy, Precision, Recall, F1-Score이며, 실험 결과는 Decision Tree, Random Forest, KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Multinomial Naive Bayes 모델들과 비교하였다.

2. 관련 연구

2.1 기계학습 기반 네트워크 침입탐지 시스템

기계 학습에서 특징 선택이란 학습에 필요한 특징을 제거하여 간결한 특징 집합을 만드는 것이다[8]. 강승호 외[1]는 NSL-KDD 데이터로부터 Pearson 상관계수 기반의 특징 선택 알고리즘을 제안하였다. 주어진 임계치 이상의 상관계수를 갖는 특징 집합을 그래프 자료구조로 표현한 뒤, 최소 지배 집합(Minimum dominating set)문제로 정의하였으며 이를 해결하는 휴리스틱 알고리즘을 제안하였다. 최희수 외[2]는 NSL-KDD 데이터로부터 특징들의 빈도수와 평균값을 통한 새로운 특징 선택 기법 AR (Attribute Ratio)을 제안하였다. Nutan 외[3]는 Hybrid Feature Selection 방법을 제안하였다. 서로 다른 특징 선택 알고리즘으로부터 중복 제거 합집합으로 표현하여 학습에 사용하였다.

불균형 데이터란 각 클래스별로 데이터들의 비율이 고르지 않고 치우쳐져 있는 것을 의미한다. 소수 클래스(Minor class)의 비율이 10% 이하인 경우는 심각한 불균형 데이터로 분류할 수 있다[9]. SMOTE (Synthetic Minority Oversampling Technique)는 이러한 불균형 데이터를 해결하기 위한 기법으로 소수 클래스들의 중 임의샘플을 중심으로 KNN 알고리즘을 활용해 k개의 샘플을 합성하여, k 샘플들 사이에 새로운 가공 데이터를 생성하는 방법이다[10]. ROS (Random Over Sampling)는 단순히 동일한 데이터를 복제하여 데이터를 늘리는 방법으로, SMOTE는 이와 달리 합성 데이터를 생성해 낼 수 있다. Tesfahun 외[7]는 SMOTE를 이용해 NSL-KDD 데이터 셋의 불균형 문제를 해결하는 연구를 진행하였다. 또한 신경망 기반의 생성 모델을 통해 이러한 가공 데이터를 만들어 내는 연구가 활발히 진행되고 있다. Yanqing Yang 외[11]은 생성모델 ICVAE (Improved Conditional Variational AutoEncoder)를 제안하였으며, 이를 통해 NSL-KDD 데이터 불균형을 해소하여 심층 신경망 모델을 학습시켰다.

2.2 NSL-KDD 데이터 셋

NSL-KDD 데이터 셋은 1999년 DARPA 침입탐지 평가 프로그램을 통해 만들어진 KDD CUP 99 데이터 셋을 M. Tavallaee 외 [12]가 개선하여 제안한 데이터 셋으로, 미 공군의 네트워크를 모델링하여 38가지의 네트워크 침입 탐지 공격 시뮬레이션을 통해 만들어 졌다. M. Tavallaee 외 [12]은 KDD CUP 99 데이터 셋의 규모가 지나치게 크며, 많은 중복 레코드 등을 포함하는 문제점이 있다고 지적하였다. 이는 발생 빈도가 높은 공격에 데이터가 매우 치우쳐져 있음을 의미한다. 이러한 문제를 해결한 NSL-KDD 데이터 셋 또한 클래스간 불균형 문제는 여전히 존재한다.

NSL-KDD 데이터 셋은 정답(Label)을 포함하여 42개의 속성으로 구성되며, 공격은 실제 개별 38개의 공격을 모두 분류하는 것이 아닌, Table 1에서 제시된 4개의 공격유형과 1개의 정상상태로, 총 5개의 클래스를 분류하는 것을 목표로 한다. 또한 학습 데이터 셋과 테스트 데이터 셋을 따로 구분해서 제공하고 있으며, 학습 데이터 셋에는 24가지 공격 유형만이 포함되어 있다. 이는 두 데이터 셋의 간극이 크다고 할 수 있다.

Table 1. Attack Type of NSL-KDD Dataset

Type	Description
Normal	normal traffic
DoS	Denial of Service
Probe	Pre-operation for vulnerability analysis before intrusion
U2R	Unauthorized access to take over root authority
R2L	Attempting unauthorized access from remote

3. HFS-DNN

3.1 Data Preprocessing

심층 신경망에서 입력 데이터의 정규화(Normalization)는 학습 속도를 장려하며, 지역 최적점(Local Optimum)에 빠지는 것을 방지하는 것으로 알려져 있다. NSL-KDD 데이터 셋 또한 학습 전 정규화 과정을 진행하며, 모든 속성값들을 0과 1사이의 값으로 변경한다. 정규화는 데이터 형식에 따라 달리 진행하였으며, NSL-KDD 데이터 셋의 데이터 형식은 nominal, numeric, binary 3가지로 구분할 수 있다[2]. nominal type 데이터들은 범주형 문자 데이터들로 신경망의 입력으로 사용할 수 없는 형태이다. 따라서 모두 정수형으로 인코딩 한 뒤 one-hot 벡터로 변환하였다. numeric type 데이터들에 대해서는 속성 값들의 범위의 차이를 왜곡하지 않고 공통 스케일로 변경하기 위해 최소 최대 정규화(Min-max Normalization)를 진행하였으며, binary type 데이터들의 경우 모두 0과 1로 구성되기 때문에 별다른 전처리 과정을 수행하지 않았다. 이를 통해 41 입력차원에서 122 입력차원으로 변환되었으며, nominal 데이터들의 one-hot 표현에 따라 입력차원이 증가하였다. 이후 데이터를 관측한 결과 num_outbound_cmds 특징은 표준편차가 0으로 모든 데이터의 값이 동일하기 때문에 학습에 불필요하다 판단되어 사전에 제거하였다. 이를 통해 데이터는 최종적으로 총 121 입력차원으로 변환된다.

3.2 Hybrid Feature Selection

대용량 네트워크 트래픽을 실시간 탐지하기 위해서는 학습 성능을 보장하면서도, 더 적은 하위 속성 집합을 찾을 수 있어야 한다. 본 논문에서는 HFS (Hybrid Feature Selection) 기법을 제안하며, 단일 특징 선택 알고리즘들에 비해 더 적은 하위 속성 집합으로 학습 모델의 정확도를 유지할 수 있음을 보인다. HFS 기법은 단일 특징 선택 알고리즘들의 출력 결과인 각 하위 속성 집합들을 구한 뒤, 이들의 교집합을 사용하는 것으로 비교적 간단한 방법으로 효율성을 증대할 수 있다. HFS 기법은 중첩 특징(Irrelevant Feature) 및 학습에 무관한 특징(Redundant Feature)을 제거하는 2 가지 목적에 따라 아래 제시된 3가지 특징 선택 알고리즘을 Fig. 1의 Feature Selection 파트와 같이 분류하고 있다.

- Pearson Based Feature Selection
- Feature Importance Based Feature Selection
- Attribute Ratio Based Feature Selection

1) Pearson Based Feature Selection

Pearson 상관계수를 이용한 특징 선택 방법은 높은 상관계수를 가지는 두 가지 특징을 중첩관계로 보고, 이 중 하나만을 사용하는 것이다. Pearson 상관계수는 Equation (1)로 정의되며 -1과 1사이의 값으로 나타난다. cov는 공분산을 의미하며, σ_X 모집단 X의 표준편차를 σ_Y 는 모집단 Y의 표준편차를 나타낸다. 1에 가까울수록 양의 상관관계에 있다고 할 수 있으며, -1에 가까울수록 음의 상관관계에 가까움을 의미한다. 0에 근접할 경우는 상관관계가 없음을 의미한다.

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Pearson 상관계수는 연속형 자료들간의 상관관계를 나타냄에 따라서 NSL-KDD 데이터 셋에서는 numeric 속성들에 대해서만 특징 선택을 진행하였다. 0.9 임계값을 사용하여 특징들의 관계를 분석하였으며, 이들의 관계를 무방향 그래프 자료구조로 표현하였다. 이렇게 표현된 그래프 안에서 최

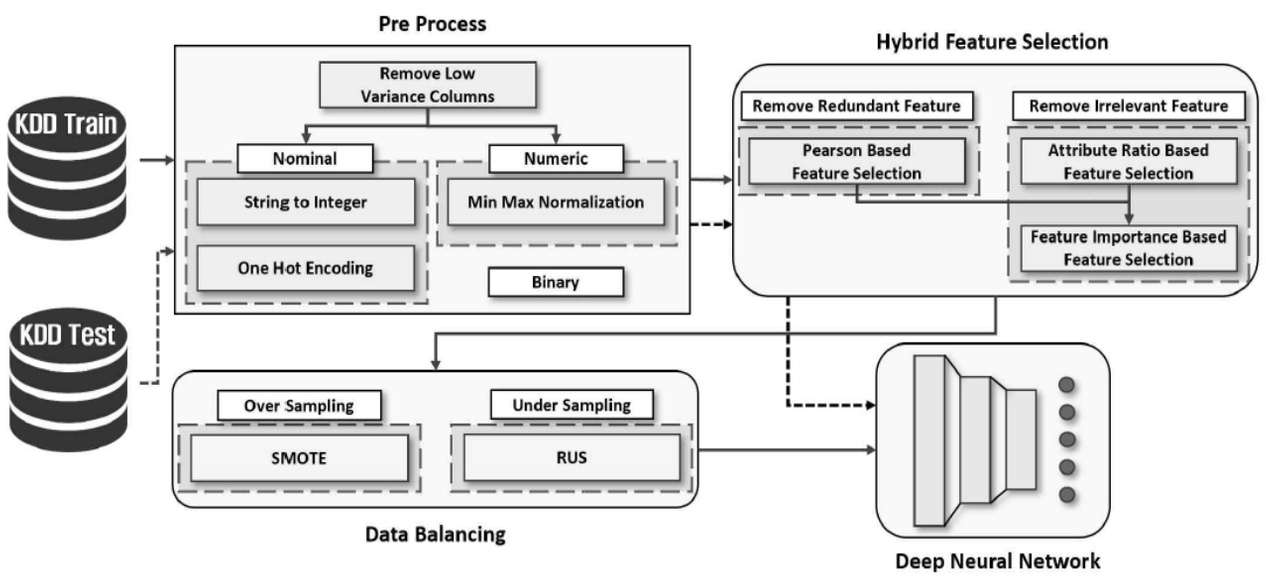

Fig 1. Proposed Network Intrusion Detection Model Structure

Table 2. Features with a Pearson Correlation of 0.9 or Higher

Complete graph node	
1	dst_host_srv_count(31), dst_host_same_srv_rate(32)
2	rerror_rate(19), srv_rerror_rate(18), dst_host_rerror_rate(20), dst_host_srv_rerror_rate(17)
3	serror_rate(26), srv_serror_rate(25), dst_host_serror_rate(27), dst_host_srv_serror_rate(24)
4	num_comproimised(15), num_root(28)

소수 하위 특징 집합을 선택할 경우 입력 특징 집합의 크기를 최소화 할 수 있으며, 이는 최소 지배 집합 문제(Minimum dominating set)로 귀결된다[1]. 최소 지배 집합 문제는 NP-Hard에 해당 하는 문제지만[13], 본 실험에서는 Table 2와 같이 4쌍의 완전 그래프 결과를 획득하였으며, 이에 따라 어떠한 속성만을 한 가지 사용하여도 최소수가 보장되는 것을 알 수 있다. 따라서 본 논문에서는 Table 2의 각 행의 맨 앞 4가지 속성만을 사용하며, 나머지 중첩 특징들은 제거한다. 이를 통해 113개의 특징을 선출하였다.

2) Feature Importance Based Feature Selection

특징 중요도(Feature Importance)를 이용한 특징 선택 방법에서는 의사결정트리 모델을 학습 시킨 뒤, 정보 획득량(Information Gain)에 기반 하여 학습에 사용된 각 특징들의 중요도를 파악할 수 있다. 의사결정트리는 정보 획득량을 최대화하는 특징을 기준으로 노드를 우선 분할한다. 이는 노드의 중요도 값이 클수록 해당 노드에서의 불순도가 크게 감소하는 것을 의미한다. 본 논문에서는 Random Forest 학습 모델을 통해 이러한 특징 중요도를 추출한 뒤 정렬하여, 임계값 0.0001값을 통해 상위 55개의 특징을 선출하였다.

3) Attribute Ratio Based Feature Selection

AR (Attribute Ratio) 기반 특징 선택 방법은 위의 두 방법과 달리 일반적이지 않은 새로운 접근 방법으로, 특징의 빈도수와 평균값을 통해 특징 중요도를 계산하며, NSL-KDD 데이터 셋의 학습 과정에서 매우 적은 특징수를 가지고도 학습이 잘 되는 것으로 보고하고 있다[2]. 본 논문에서는 0.1 임계값을 통해 상위 50개의 특징을 선출하였다.

Table 2는 Pearson 상관 계수가 0.9 이상의 관계를 가진 특징들을 나타내고 있다. 각 특징들의 옆에는 Random Forest 학습 모델을 통해 선출된 특징 중요도의 순위를 표시되어 있는데, 이를 참조하면 높은 상관관계를 가진 특징들끼리는 모두 유사한 순위를 가지는 것을 알 수 있으며, 또한 모두 b)절에서 선출한 상위 55개에 포함되어 있는 것을 알 수 있다. 이는 특징 중요도만을 통해 특징 선택을 할 경우 이러한 중첩 특징들을 고려할 수 없음을 의미한다. 따라서 55개의 특징 집합에서 추가적으로 8개의 특징을 제거할 수 있음을 알 수 있다. 이는 이들 특징 선택 기법들의 결과에 해당하는 하위 집합들의 교집합임을 알 수 있다. 본 논문에서 제시하는 HFS는 이러한 두 가지 관점의 특징들을 모두 걸러낼 수 있다. 이와 같이 위에 제시된 3가지 특징 선택 기법의 교집합을 통해, 단일 특징 선택에 비해 더 작으면서 학습 성능을 저해하지 않는 입력 특징 집합을 획득하였다. 최종적으로 Table 3에서 제시된 39개의 특징을 사용한다. 이는 기존 121 입력차원 대비 32% 규모의 크기를 가진다.

Table 3. Set of Features Selected through HFS

Selected Features (39)
count, diff_srv_rate, dst_bytes, dst_host_count, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_count, dst_host_srv_diff_host_rate, duration, flag_REJ, flag_RSTR, flag_S0, flag_SF, hot, is_guest_login, logged_in, num_compromised, num_failed_logins, num_file_creations, protocol_type_icmp, protocol_type_tcp, protocol_type_udp, rerror_rate, root_shell, same_srv_rate, serror_rate, service_domain_u, service_eco_i, service_ftp, service_ftp_data, service_http, service_other, service_private, service_smtp, service_telnet, src_bytes, srv_count, srv_diff_host_rate, wrong_fragment

3.3 Data Balancing

불균형 데이터란 소수 클래스(Minor class)에 데이터 수가 다수 클래스(Major class)에 포함된 데이터 수와 비교해 현저히 적은 데이터를 의미한다. 기계 학습 모델의 성능은 데이터에 의존적이며, 이를 해소하지 않은 채 학습 모델에 적용할 시 분류 성능의 저하를 야기할 수 있다[9]. 특히 소수 클래스들의 탐지율은 크게 떨어지게 되는데, 이는 소수 클래스의 범주가 다수에 클래스에 의해 침범 당하기 때문이다[14]. 본 논문에서 학습에 사용하는 NSL-KDD 데이터 셋 또한 Table 4와 같이 클래스 간 샘플 수의 차이가 매우 크다. 이러한 문제를 해결하기 위해서 본 논문에서는 오버샘플링(Over sampling) 기법과 언더 샘플링 기법(Under sampling)을 통해 불균형 문제를 해소한다.

데이터의 절반에 해당하는 Normal 클래스는 다수 클래스로 RUS (Random Under Sampling) 기법을 통해 데이터를 샘플 수를 축소하였으며, 상대적으로 소수 클래스에 해당하는 Probe, U2R, R2L은 SMOTE (Synthetic Minority

Table 4. Percentage and Number of Samples in NSL-KDD Train Data Set

	KDD Train		Balanced KDD Train	
Normal	67343	(53%)	45000	(19.99%)
DoS	45927	(37%)	45927	(20.04%)
Probe	11656	(9.11%)	45000	(19.99%)
U2R	52	(0.04%)	45000	(19.99%)
R2L	995	(0.85%)	45000	(19.99%)
Total	125973		225927	

Table 5. The Neural Network Configuration used in the Experiment

DNN Parameters	
Layer	39-256-512-512-5
Activation	Relu, Softmax
Initializer	He Uniform
Regularizer / Strength	L2 / 0.0001
Optimizer / Learning rate	Adam / 0.0005
Loss	Cross Entropy

Over Sampling Technique) 기법을 통해 샘플 수를 비슷한 수준의 크기로 늘려주었다. Table 4를 참조하면 불균형을 해소한 데이터 셋의 샘플수와 비율을 확인할 수 있으며, 분류 모델의 소수 클래스들의 탐지율 개선을 기대할 수 있다.

3.4 Deep Neural Network

본 논문에서는 네트워크 침입 탐지 분류 모델로 심층 신경망을 사용한다. Table 5를 통해 제안된 심층 신경망 네트워크의 구조를 확인할 수 있으며, 은닉 계층(Hidden layer)의 활성 함수로 relu를 사용하였다. 신경망의 입력 계층은 3.1절의 전처리 과정과 3.2절의 특징 선택 과정을 통해 39개이 입력 크기를 가지며, Fully Connected Network를 사용하였다. 신경망의 학습에서 초기 가중치 설정은 매우 중요한 역할을 하는데, 이는 기울기 소실(Gradient vanishing)과 같은 문제로 이어질 수 있기 때문이다. 일반적으로 많이 사용되는 Xavier Initializer는 relu 함수와 같이 사용할 경우 레이어의 깊이가 깊어질수록 출력값이 0에 가까워지는 문제가 발생한다[15]. 따라서 이러한 문제를 해결한 He Initializer를 사용하여 신경망의 초기값을 설정한다. 또한 학습 데이터에 과적합이 되는 것을 방지하기 위해서 L2 규제를 사용하였다.

4. 실험 및 결과

본 연구에서는 실험을 위해 NSL-KDD 데이터 셋에서 학습과 테스트를 위해 제공되는 KDDTrain+, KDDTest+ 데이터를 사용한다. 학습 데이터는 3.3절에서 소개된 Data Balancing 단계를 거쳐 Table 4의 Balanced KDD Train과 같이 변경 후 사용하며, 전체 학습 데이터 중 30%를 검증용 데이터(Validation set)로 사용하여 심층 신경망 모델의 학습 과정에서 과적합 여부를 판단한다. 이를 통해 검증 오류율(Validation loss)이 10 epoch 이상 증가할 경우 조기 멈춤하였으며, 이후 학습에서 가장 검증 오류율이 낮았던 모델을 테스트에 사용하였다. 전체 epoch은 50을 사용하였으며, batch size는 64를 사용하였다.

테스트에 사용되는 데이터 셋은 Table 10의 support를 참고할 경우 클래스간 불균형을 확인할 수 있으며, 이러한 불균형 데이터를 평가하는데 일반적으로 사용되는 Accuracy만으로 평가하는 것은 적절하지 않다. 이는 전체 샘플 중 옳게 평가된 샘플 수의 비율을 의미하기에 모델의 출력이 모두 다수 클래스로 예측하더라도 높은 결과를 보이기 때문이다. 따라서 Recall, Precision, F1-Score와 같은 불균형 데이터에서의 많이 사용되는 지표들과 함께 성능을 비교한다. 학습 모델의 실험 결과는 Decision Tree, Random Forest, KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Multinomial Naive Bayes 모델과 비교 및 분석한다.

4.1절에서는 3.2절에서 제안한 HFS의 효율성을 검증하기 위한 실험을 진행하며, 4.2절에서는 3.2절에 이어 3.3절에서 서술하고 있는 내용을 추가하여 최종적 실험 결과를 보고한다. 실험에 사용된 심층 신경망 모델은 모두 3.4절에서 서술한 네트워크를 동일하게 사용한다.

4.1 특징 선택 기법에 따른 성능 비교

Table 6은 3.2절에서 언급한 특징 선택 기법들을 통해 선택된 특징 하위 집합들의 크기와 그에 따른 성능을 비교하고 있다. 학습에 사용된 심층 신경망은 Table 5에서 언급한 네트워크와 동일한 구조를 가지고 있으며, 오버 샘플링을 하지 않았기에 데이터 수가 적은 것을 감안하여 20% 만을 검증용 데이터로 사용하여 조기 멈춤을 실행하였다. 각 특징 선택들의 임계값은 3.2절에 언급된 값을 사용하였으며, 전체 121 입력차원을 사용한 것(Vanilla model)과 큰 차이를 보이지 않는 것을 확인할 수 있다. 심층 신경망에서 입력 차원의 축소는 학습해야 될 가중치가 더 줄어든 것을 의미하며, 이는 계산 효율성과 연결된다. 이러한 결과를 통해 본 논문에서 제안한 HFS 알고리즘이 단일 특징 선택 알고리즘들과 비교해, 실시간 탐지를 위한 효율적인 입력 특징 집합을 찾을 수 있는 것을 확인하였다.

하지만 특징 선택 기법들을 통해 입력 차원을 축소하는 것에는 성공하였으나, Table 7을 참고하면 소수 클래스들의 분류 성능은 여전히 개선되지 않을 것을 확인할 수 있다. 특히 대부분의 모델들이 소수 클래스들의 recall 점수는 매우 저조하며, precision은 상대적으로 높은데, 이는 학습모델이 매우 낮은 비율로 소수 클래스라고 예측을 시도하며, 전체 소수 클래스의 분류에 실패하였음을 알 수 있다. 따라서 학습 모델이 다수 클래스에 편향되어 학습 되었음을 알 수 있으며, 위 결과를 통해 3.3절에서 언급한 Data Balancing 파트의 필요성을 확인할 수 있다.

Table 6. Performance Comparison of DNN between Feature Selection Techniques

	acc	pre	recall	f1
Vanilla(121)	0.78	0.79	0.78	0.75
Pearson(113)	0.77	0.78	0.77	0.73
RF Importance(55)	0.78	0.78	0.79	0.76
Attribute Ratio(50)	0.78	0.83	0.78	0.75
Hybrid(39)	0.77	0.82	0.77	0.75

Table 7. Comparison of Minority Class Detection Performance of DNN between Feature Selection Techniques

	precision	recall	f1	support
Vanilla DNN				
R2L	0.68	0.11	0.19	2754
U2R	0.75	0.1	0.18	200
Pearson FS DNN				
R2L	0.68	0.03	0.05	2754
U2R	0.63	0.11	0.19	200
RF Importance FS DNN				
R2L	0.47	0.13	0.20	2754
U2R	0.69	0.10	0.17	200
Attribute Ratio FS DNN				
R2L	0.97	0.14	0.24	2754
U2R	0.8	0.02	0.04	200
Hybrid FS DNN				
R2L	0.92	0.1	0.19	2754
U2R	0.67	0.07	0.13	200

Table 8. Performance Comparison between Classification Models with HFS

	acc	pre	recall	f1
Decision Tree	0.75	0.79	0.75	0.72
Random Forest	0.76	0.82	0.77	0.73
KNN	0.75	0.78	0.75	0.71
SVM	0.74	0.79	0.74	0.71
MultinomialNB	0.72	0.77	0.72	0.70
DNN	**0.77**	**0.82**	**0.77**	**0.75**

Table 8은 Hybrid Feature Selection을 통해 구해진 특징 집합을 통해 비교군 모델들과 학습 성능을 비교한 결과이다. 4가지 성능 지표 모두 본 논문에서 제안한 심층 신경망 모델이 가장 좋은 것을 확인하였으며, 눈에띄는 성능 상의 차이는 보이지 않고 있다. 이 외에 심층 신경망을 제외한 나머지 결과에서는 Random Forest 모델이 비교적 우수한 결과를 보였다.

4.2 데이터 불균형을 해소한 학습 모델의 성능 비교

본 절에서는 4.1절에서 아직 해소하지 못한 불균형 문제를 Table 4에 제시된 밸런싱을 진행한 데이터 셋을 통해 실험을 진행하고, 다른 기계학습 모델들과 성능을 비교하여 보고한다. Fig. 2와 Table 9를 참조하면 각 모델들의 성능을 비교할 수 있으며, 각 모델들이 불균형 문제를 해소하였음에도 불구하고 Decision Tree, Random Forest, Multinomial NB 알고리즘에서는 precision과 recall의 조화 평균인 F1-score 기준으로 평가할 시 유의미한 성능 개선이 이루어지지 않은 것을 알 수 있다. 하지만 KNN, SVM, DNN 모델은 모두 F1-score 기준으로 6~8% 정도의 큰 폭의 성능 개선이 이루어졌으며, 본 논문에서 제시한 심층 신경망이 가장 좋은 성능을 보이고 있다.

Table 9. Performance Comparison between Classification Models Learned through Balanced Data Sets

	acc	pre	recall	f1
Decision Tree	0.73	0.75	0.73	0.71
Random Forest	0.76	0.82	0.76	0.72
KNN	0.79	0.79	0.79	0.78
SVM	0.79	0.82	0.79	0.79
Multinomial NB	0.73	0.79	0.73	0.74
DNN	**0.82**	**0.84**	**0.82**	**0.81**

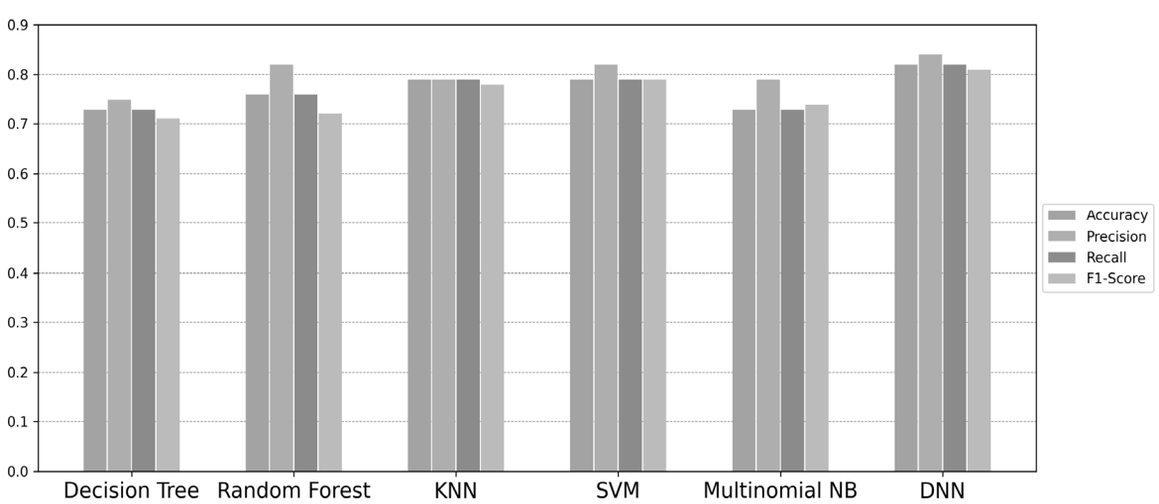

Fig. 2. Performance Comparison through HFS and Data Balancing

Table 10. The Detection Performance of the Proposed Learning Model in NSL-KDD Dataset

HFS-DNN				
	precision	recall	f1	support
DoS	0.96	0.86	0.90	7458
Probe	0.86	0.72	0.78	2421
R2L	0.85	0.36	0.51	2754
U2R	0.22	0.14	0.17	200
Normal	0.75	0.97	0.84	9711
Total	0.84	0.82	0.81	22544

Table 10은 본 논문에서 제안한 HFS-DNN의 각 클래스별 성능 지표로, Table 7의 Hybrid FS DNN 항목과 비교시 소수 클래스들의 성능 개선을 확인할 수 있다. precision 점수는 불균형 해소 전과 비교해 낮아졌지만 recall 점수가 오른 것을 통해 모델의 출력이 소수 클래스라 예측하는 빈도가 비교적 상승했으며, 전체 소수 클래스들의 탐지 성능이 개선되었음을 알수 있다. 또한 R2L 클래스의 탐지 성능의 경우 F1-score 기준 32%의 큰 상승폭을 가지는 것을 알 수 있다.

그럼에도 불구하고 U2R 클래스의 분류 성능은 여전히 낮은 것을 확인할 수 있는데, 이는 학습 데이터 셋에서 제공되는 데이터양이 52개인 반면에 테스트 데이터 셋에서 제공되는 양은 200개로 학습에 사용되는 샘플 수가 더 적은 것을 알 수 있다. 따라서 학습 데이터 셋의 샘플들과 테스트 데이터 셋의 샘플들 사이의 간극이 큰 것으로 판단되며, 이러한 유형의 문제는 오버샘플링 기법을 통해 극복하기에 부적절하다. 이는 NSL-KDD 데이터 셋이 가진 문제점으로 분석하고 있으며, 정상적인 학습을 위해서는 학습과 테스트 데이터 셋의 U2R 샘플을 서로 교환하여 학습을 진행하는 것이 바람직하다 판단된다.

5. 결 론

본 논문에서는 네트워크 침입 탐지의 성능 개선을 위해 Hybrid Feature Selection 기법을 제안하였으며, 이를 심층 신경망에 이용한 HFS-DNN (Deep Neural Network) 모델을 제안하였다. 또한 학습에 사용된 NSL-KDD 데이터 셋의 불균형 문제를 해소하기 위해 SMOTE와 RUS 기법을 사용한다. 제안된 Hybrid Feature Selection을 통해 32% 규모로 입력 차원을 축소할 수 있었으며, 실험을 통해 축소된 입력 차원으로 기존과 동일한 성능을 보장하는 것을 확인할 수 있었다. 또한 오버 샘플링 기법을 통해 소수 클래스의 탐지율 개선을 실험을 통해 확인할 수 있었다. 하지만 U2R 클래스는 탐지율이 개선되지 않았는데, 이는 학습 데이터 셋과 테스트 데이터 셋의 간극이 너무 큰 것으로 분석된다.

본 논문에서 제안된 HFS-DNN의 성능은 Hybrid Feature Selection과 Data Balancing을 적용할 경우 F1-score 기준 6% 성능 개선이 있었으며, Decision Tree, Random Forest, K-Nearest Neighbor, SVM, Multinomial Naive Bayes 모델들과 비교해 4가지 지표 모두 앞서는 것으로 나왔다. 또한 본 논문에서 제안한 두 가지 특징 선택과 데이터 불균형 해소 방법은 SVM, KNN 모델에서도 좋은 성능 개선 효과를 보였다. 향후 연구로는 VAE, GAN과 같은 생성모델들을 통해 오버 샘플링을 진행하는 연구를 진행할 수 있으며, 이는 SMOTE 알고리즘을 대체할 수 있을 것으로 사료된다. 또한 다른 네트워크 침입 탐지 데이터 셋에 대해서 실험을 확장시킬 수 있다.

References

[1] S. H. Kang, I. S. Jeong, and H. S. Lim, "A feature set selection approach based on pearson correlation coefficient for real time attack detection," *Convergence Security Journal*, Vol.18, No.5_1, pp.59-66, 2018.

[2] H. S. Chae, B. O. Jo, S. H. Choi, and T. K. Park, "Feature selection for intrusion detection using NSL-KDD," *Recent Advances in Computer Science*, pp.184-187, 2013.

[3] N. F. Haq, A. R. Onik, and F. M. Shah, "An ensemble framework of anomaly detection using hybridized feature selection approach (HFSA)," In *2015 SAI Intelligent Systems Conference*, pp.989-995, 2015.

[4] R, Longadge and S, Dongre, "Class imbalance problem in data mining review," arXiv preprint arXiv:1305.1707, 2013.

[5] T. H. Kim, S. H. Kang, "An Intrusion Detection System based on the Artificial Neural Network for Real Time Detection," *Convergence Security Journal*, Vol.17, No.1, pp.31-38, 2017.

[6] J. Song, H. Takakura, Y. Okabe, and Y. Kwon, "Correlation analysis between honeypot data and IDS alerts using one-class SVM," *Intrusion Detection Systems*, pp.173-192, 2011.

[7] A. Tesfahun and D. L. Bhaskari, "Intrusion detection using random forests classifier with SMOTE and feature reduction," *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*, pp.127-132, 2013.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, Vol.3, pp.1157-1182, 2003.

[9] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," IEEE international joint conference on neural networks, pp.1322-1328, 2008.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, Vol.16, pp.321-357, 2002.
[11] Y. Yang, K. Zheng, C. Wu, and Y. Yang, "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," *Sensors*, Vol.19, No.11 pp.2528. 2019.
[12] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp.1-6, 2009.
[13] H. Jiang, J. Nagra, and P. Ahammad, "Sok: Applying machine learning in security-a survey," arXiv preprint arXiv:1611.03186, 2016.
[14] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, Vol.26, No.2, pp.405-425, 2012.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings of the IEEE International Conference on Computer Vision*, pp.1026-1034, 2015.

민 병 준
https://orcid.org/0000-0002-8858-2003
e-mail : bang@sju.ac.kr
2017년 호서전문학교(공학사)
2019년 세종대학교 컴퓨터공학과(석사)
2019년 ~ 현 재 세종대학교 컴퓨터공학과 박사과정
관심분야 : 강화 학습, 침입 탐지, 분산 처리

유 지 훈
https://orcid.org/0000-0001-7516-3005
e-mail : yoojihoon@sju.ac.kr
2016년 호서전문학교(공학사)
2019년 세종대학교 컴퓨터공학과(석사)
2019년 ~ 현 재 세종대학교 컴퓨터공학과 박사과정
관심분야 : 분산 처리, 데이터 마이닝, 딥 러닝

신 동 규
https://orcid.org/0000-0002-2665-3339
e-mail : shindk@sejong.ac.kr
1986년 서울대학교 계산통계학과(학사)
1992년 Illinois Institute of Technology 컴퓨터과학과(석사)
1997년 Texas A&M University 컴퓨터과학과(박사)
1998년 ~ 현 재 세종대학교 컴퓨터공학과 교수
관심분야 : 정보 보안, 기계 학습, 데이터 마이닝, 생체 데이터 처리

신 동 일
https://orcid.org/0000-0002-8621-715X
e-mail : dshin@sejong.ac.kr
1988년 연세대학교 컴퓨터과학과(학사)
1993년 Washington State University 컴퓨터과학과(석사)
1997년 North Texas University 컴퓨터과학과(박사)
1998년 ~ 현 재 세종대학교 컴퓨터공학과 교수
관심분야 : 정보 보안, 기계 학습, 데이터 마이닝, 생체 데이터 처리