Vol.5, No.12 pp.471~480 pISSN: 2287-5891 https://doi.org/10.3745/KTCCS.2016.5.12.471

Efficient Feature Selection Based Near Real-Time Hybrid Intrusion Detection System

Woosol Lee† · Sangyoon Oh††

ABSTRACT

Recently, the damage of cyber attack toward infra-system, national defence and security system is gradually increasing. In this situation, military recognizes the importance of cyber warfare, and they establish a cyber system in preparation, regardless of the existence of threaten. Thus, the study of Intrusion Detection System(IDS) that plays an important role in network defence system is required. IDS is divided into misuse and anomaly detection methods. Recent studies attempt to combine those two methods to maximize advantagesand to minimize disadvantages both of misuse and anomaly. The combination is called Hybrid IDS. Previous studies would not be inappropriate for near real-time network environments because they have computational complexity problems. It leads to the need of the study considering the structure of IDS that have high detection rate and low computational cost. In this paper, we proposed a Hybrid IDS which combines C4.5 decision tree(misuse detection method) and Weighted K-means algorithm (anomaly detection method) hierarchically. It can detect malicious network packets effectively with low complexity by applying mutual information and genetic algorithm based efficient feature selection technique. Also we construct upgraded the the hierarchical structure of IDS reusing feature weights in anomaly detection section. It is validated that proposed Hybrid IDS ensures high detection accuracy (98.68%) and performance at experiment section.

Keywords : Intrusion Detection System, Feature Selection, C4.5 Decision Tree, Weighted K-Means Algorithms

근 실시간 조건을 달성하기 위한 효과적 속성 선택 기법 기반의 고성능 하이브리드 침입 탐지 시스템

이 우 솔† · 오 상 윤††

요 약

최근 국가 기반 시스템, 국방 및 안보 시스템 등에 대한 사이버 공격의 피해 규모가 점차 커지고 있으며, 군에서도 사이버전에 대한 중요성을 인식하고 전·평시 구분 없이 대비하고 있다. 이에 네트워크 보안에서 탐지와 대응에 핵심적인 역할을 하는 침입 탐지 시스템의 중요성이 증대되고 있다. 침입 탐지 시스템은 탐지 방법에 따라 오용 탐지, 이상 탐지 방식으로 나뉘는데, 근래에는 두 가지 방식을 혼합 적용한 하이브리드 침입 탐지 방식에 대한 연구가 진행 중이다. 그렇지만 기존 연구들은 높은 계산량이 요구된다는 점에서 근 실시간 네트워크 환경에 부적합하다는 문제점이 있었다. 본 논문에서는 기존의 하이브리드 침입 탐지 시스템의 성능 문제를 보완할 수 있는 효과적인 속성 선택 기법을 적용한 의사 결정 트리와 가중 K-평균 알고리즘 기반의 고성능 하이브리드 침입 탐지 시스템을 제안하였다. 상호 정보량과 유전자 알고리즘 기반의 속성 선택 기법을 적용하여 침입을 더 빠르고 효율적으로 탐지할 수 있으며, 오용 탐지 모델과 이상 탐지 모델을 위계적으로 결합하여 구조적으로 고도화된 하이브리드 침입 탐지 시스템을 제안하였다. 실험을 통해 제안한 하이브리드 침입 탐지 시스템은 98.68%로 높은 탐지율을 보장함과 동시에, 속성 선택 기법을 적용하여 고성능 침입 탐지를 수행할 수 있음을 검증하였다.

키워드 : 침입 탐지 시스템, 속성 선택 기법, C4.5 의사 결정 트리, 가중 K-평균 알고리즘

1. 서 론

근래 국가 기반 시스템, 국방 및 안보 시스템 등이 IT 기술에 크게 의존하는 상황에서 사이버 공격과 사이버 테러의 피해 규모가 점차 커지고 있다. 특히 군에서는 전장이 기존 물리 공간에서 사이버전 공간으로 확장되고, 작전 수행 개념

※ 본 논문은 미래창조과학부 및 정보통신기술진흥센터의 ICT/SW 창의연구과정 (SW중심대학) 지원사업(R2215-15-1002) 및 2016년도 정부(교육부) 재원으로 한국 연구재단의 지원(NRF-2015R1D1A1A01059557)을 받아 수행된 연구임.
† 준 회 원 : 육군 3사관학교 컴퓨터공학과 강사
†† 종신회원 : 아주대학교 소프트웨어학과 교수
Manuscript Received : July 22, 2016
Accepted : September 1, 2016
* Corresponding Author : Sangyoon Oh(syoh@ajou.ac.kr)

이 모든 전력 요소를 유기적인 연결을 통해 통합 작전 체계를 구성하는 네트워크 중심전(NCW, Network Centric Warfare)으로 변화함에 따라 사이버전에 대한 중요성을 인식하고 전·평시 구분 없이 대비하고 있다[1]. 사이버 전장에서 지속적인 우세를 유지하기 위해서는 네트워크 작전, 사이버 방어 작전, 공세적 사이버 작전을 성공적으로 수행해야 한다. 그 중에서도 핵심 지휘통제체계와 무기체계 등을 연동시키는 네트워크를 철저하게 방어해야 사이버작전 뿐 아니라 궁극적으로 물리전에서도 성공적인 작전을 수행할 수 있다. 이를 위해서는 네트워크에 대한 적의 공격을 조기에 탐지하고 적시적으로 대응하는 것이 매우 중요한데, 네트워크 방호 시스템 중 탐지와 대응에 핵심적인 역할을 하는 것이 침입 탐지 시스템(Intrusion Detection System)이다.

일반적으로 침입 탐지 시스템은 탐지 방법에 따라서 오용 탐지(misuse detection), 이상 탐지(anomaly detection)의 두 가지 유형으로 나뉜다[2]. 오용 탐지 방식은 이미 알려진 공격 유형을 분석하여 규칙을 생성하고, 이를 기반으로 침입을 탐지하는 방식이다. 오용 탐지 방식은 알려진 공격에 대해서 빠르고 정확하게 탐지하지만 새로운 공격을 탐지할 수 없다는 특징이 있다. 이와 반대로 이상 탐지 모델은 정상 네트워크 트래픽에 대해 분석하고 패턴을 프로파일링한 것을 기반으로, 정상 트래픽의 프로필로부터 크게 벗어나는 트래픽들을 이상 행위라고 가정하여 공격을 탐지하는 방식이다. 이상 탐지 방식은 새로운 공격 유형을 탐지하는데 유용하나 알려진 공격을 탐지하는 데에는 오류율이 높아 비효율적인 방식이다.

기존의 오용 탐지 방식과 이상 탐지 방식의 문제점을 해결하기 위해서, 두 가지 방식을 혼합하는 하이브리드 탐지 방식이 제안되었다[3]. 하이브리드 탐지 방식의 초기 연구는 단순히 오용 탐지 모델과 이상 탐지 모델을 합치는 방향으로 연구되었다. 각각의 모델에서 따로 학습하고, 테스트하여 단순히 결과를 합치는 방식이었고, 자원과 성능 측면에서 비효율적이었고 이상 탐지 방식의 오류율 문제를 해결하지 못했다. 이런 문제점을 해결하기 위해서 두 모델을 위계적으로 결합한 탐지 모델이 제안되었다[4]. 이 방식은 오용 탐지 모델 구축 이후 나머지 정상 데이터들을 통해 이상 탐지 모델을 구축하는 방식으로, 중복된 데이터로 각각의 모델을 구축하지 않아 효율적으로 공격을 탐지할 수 있다는 장점이 있다. 하이브리드 침입 탐지 시스템은 아래 두 가지 측면에서의 연구가 필요하다.

첫째, 이상 탐지 모델과 오용 탐지 모델의 장점을 극대화할 수 있는 효율적인 구조와 기법이 필요하다. 위계적 하이브리드 구조의 장점을 극대화 할 수 있는 적절한 기법의 조합과 순서, 분할 구조 등에 대한 연구가 필요하다.

둘째, 성능을 고려한 하이브리드 침입 탐지 시스템에 대한 연구가 필요하다. 최근 하이브리드 침입 탐지 시스템의 탐지율을 향상시키기 위해서 데이터 마이닝 및 기계 학습 기법을 적용하려는 시도가 활발히 진행 중이다. 이런 기법들은 정확한 패턴 분류와 군집화를 수행할 수 있다는 장점으로, 하이브리드 침입 탐지 시스템에 혼합 적용되어 알려진 공격의 패턴을 추출하고 정상 행위를 군집화하여 아웃라이어를 분별해내는 역할을 한다. 이런 접근 방식들은 침입 탐지 시스템의 탐지율을 향상시키고 오탐율을 줄이는데 큰 기여를 하였다. 그렇지만 보통의 데이터 마이닝, 기계 학습 알고리즘은 큰 계산 비용이 든다는 문제점이 있기 때문에 근 실시간 네트워크 환경을 고려하면 이 문제점은 침입 탐지에 핵심적인 결함이 될 수 있다. 보안 시스템에서는 탐지의 정확성뿐만 아니라 빠른 탐지와 적시적인 대응 또한 필수적이기 때문이다. 이런 문제를 해결하기 위해서는 전체 시스템 측면에서 성능을 고려한 침입 탐지 시스템을 구축해야한다.

본 논문에서는 위에서 언급한 하이브리드 침입 탐지 시스템의 성능 문제를 보완할 수 있는 효과적 속성 선택을 적용한 의사 결정 트리와 가중 K-평균 알고리즘 기반의 고성능 하이브리드 침입 탐지 시스템을 제안한다. 먼저 네트워크 프로토콜 유형별로 속성 선택(Feature selection) 기법을 적용하고, 선택된 속성에 대해서 탐지를 수행한다. 오용 탐지 모델에는 네트워크 프로토콜 유형별로 C4.5 의사 결정 트리를 적용하고 이상 탐지 모델에는 가중 K-평균 알고리즘을 적용한다. 실험을 통해 제안한 하이브리드 침입 탐지 시스템은 98.68%로 높은 탐지율을 보장함과 동시에, 속성 선택 기법을 적용하여 고성능 침입 탐지를 수행할 수 있음을 검증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 하이브리드 침입 탐지 시스템과 속성 선택 기법에 대해서 소개한다. 3장에서는 제안하는 하이브리드 침입 탐지 시스템에 대해 설명한다. 4장에서는 제안 시스템을 활용하여 기존 하이브리드 시스템과 탐지율과 성능 측면에서 비교·분석하였으며, 5장에서는 결론 및 향후 연구 과제에 대해 제시하였다.

2. 관련 연구

2.1 하이브리드 침입 탐지 시스템

Depren[3]은 하이브리드 침입 탐지 시스템 구축에 있어 오용 탐지 방식과 이상 탐지 방식을 평행하게 구축하는 접근 방식(Parallel Hybrid IDS)을 제안하였다. 제안한 하이브리드 침입 탐지 시스템 모델은 의사 결정 트리 기반의 오용 탐지 모델과 SOM (Self-organization map) 기반의 이상 탐지 모델, 의사 지원 시스템으로 구성되어 있다. 두 모델을 수평으로 배치하여 구축되어 독립적으로 데이터셋을 훈련시키는 구조다. 각각의 탐지 모델은 상호 교류가 없고 단순히 의사 지원 시스템에 의해 탐지 결과를 통합되는 방식으로 비효율적인 구조를 지니며, 같은 데이터에 대해 두 번 학습해야하기 때문에 불필요한 계산 비용이 들어간다.

G. Kim[5]는 의사 결정 트리 기반의 오용 탐지 모델과 1-class SVM 기반의 다중 이상 탐지 모델을 결합한 하이브리드 탐지 시스템(DT-SVM)을 제안하였다. 이전 연구들과는 다르게 오용 탐지 모델과 이상 탐지 모델을 위계적으로 통합한 구조로 각각 모델의 장점과 특성을 적절히 활용하였다. 먼저 전체 데이터에 대해서 의사 결정 트리를 활용하여 알려진 공격에 대한 탐지를 수행하고, 분류된 정상 데이터들에 대해서 다중 1-class SVM 모델을 형성하여 아웃라이어를 분류해 새로운 공격으로 간주한다.

그렇지만 이 연구에서 적용한 DT, SVM 모델은 성능 문제를 고려하지 않아 근 실시간 환경에 부적절한 침입 탐지 시스템이다. 침입 탐지 시스템에 사용되는 모든 특징이 탐지에 필수적인 영향을 미치지 않을 뿐 아니라, 일부는 좋지 않은 영향을 미치기도 하는데[6], 이 연구에서는 이런 현상을 고려하지 않고 데이터 셋의 모든 특성을 고려하여 모델을 구축하였다.

본 논문에서는 위에서 설명한 침입 탐지 시스템의 위계적인 구조를 적용하되, 속성 선택 기법을 적용하여 고차원 분류 문제를 해결하고자 한다. 훈련, 탐지 시간을 단축시킬 수 있으며, 특히 이상 탐지 모델에는 속성 선택 단계에서 분석된 속성의 중요도를 반영하여 더 정확한 이상 탐지를 수행한다.

2.2 속성 선택 기법을 활용한 침입 탐지 시스템

본 장에서는 제안하는 하이브리드 침입 탐지 시스템에 적용될 속성 선택 기법의 기본적인 개요에 대해서 설명하고, 다음 속성 선택 기법을 적용한 기존 침입 탐지 시스템에 대해서 소개한다.

일반적으로 속성 선택 기법은 필터(filter) 방식과 래퍼(wrapper) 방식으로 분류된다. 필터 방식은 전처리 단계에서 학습 알고리즘에 의해 독립적으로 속성을 선택하는 기법으로, 높은 순위의 속성을 우선 선택한다. 필터 기법은 연산 속도가 빨라 고차원 데이터 셋을 분석하는데 적합하다는 장점이 있다[7]. 그렇지만 필터 기법은 분류기에 최적화되있지 않고, 최적의 속성 개수를 선택하는데 추가적인 알고리즘이 필요하다는 단점이 있다[8]. 필터 방식에서 속성 연관성의 순위를 매기는 척도로 가장 일반적으로 사용되는 방식은 상관 계수(Correlation criteria), 상호 정보량(Mutual Information) 등이 있다.

래퍼 기법에 기반을 둔 속성 선택 기법은 분류기의 분류 결과를 활용하여 속성들을 반복적으로 선택하고, 평가하여 속성들을 선택하는 기법이다. 래퍼 기법은 분류기의 분류 결과를 평가하여 속성을 선택하기 때문에 분류기에 최적화된 속성을 선택할 수 있어 데이터들을 정확하게 분류할 수 있다는 장점이 있다. 하지만 속성 부분 집합을 평가하기 위해 매번 새로운 분류 모델을 생성해야하기 때문에 계산 비용이 많이 들어 보통 필터 기법에 비해 연산 속도가 느리다. 래퍼 기법에서 주로 사용하는 탐색 기법은 유전자 알고리즘(Genetic algorithm)이나 PSO(Particle Swarm Optimization) 등이 있다.

침입 탐지 시스템 분야에서도 고차원 네트워크 데이터에서 악성 네트워크 패킷을 탐지하기 위해 속성 선택을 적용하는 연구들이 진행되고 있다. 침입탐지시스템에서 속성 선택을 활용하여 얻을 수 있는 장점은 다음과 같다[9]. 첫째, 선택한 부분 속성 집합만을 활용하여 침입 탐지를 수행하기 때문에 탐지 학습 및 분류의 계산량이 줄어든다. 둘째, 고속 네트워크 환경에서 수집해야 하는 속성 정보의 양이 줄어든다. 셋째, 분류된 네트워크 패킷 데이터에 대한 이해도를 높인다. 넷째, 중복된 데이터와 상관성이 낮은 데이터, 노이지 데이터를 배제한다.

Fatem Amiri[10]은 상호 정보량 기반의 속성 선택 기법을 적용한 침입 탐지 시스템을 제안하였다. 최초 속성은 전통적인 방식의 상호 정보량을 기준으로 선택하고, 다음 속성부터는 상호 정보량을 기반으로 한 목적 함수를 두어 그 값이 최대가 되는 속성을 순차적으로 선택했다. 최종적으로 목적함수가 최대가 되는 부분 집합의 속성들을 선택한다. 여기서 제안한 상호 정보량 기반 속성 선택 기법은 기존 필터 기법이 지니는 속도 측면에서의 장점은 가져가면서, 최적의 속성 부분 집합 수까지 결정할 수 있는 알고리즘을 제시하였다. 그렇지만 속성 중요도를 탐지에 활용하지 않았으며, 선택된 속성 집합이 분류기 결과에 최적화된 집합이 아니라는 필터 방식의 단점은 그대로 보유하고 있다.

J. Cho[11]은 상호 정보량 기반의 필터 방식과 유전자 알고리즘 기반의 래퍼 방식 두 가지를 융합한 속성 선택 기법을 제안하였다. 먼저 상호 정보량을 계산하여 순위를 매겨 초기 후보 집단을 선택하고, 유전자 알고리즘으로 분류기에 최적화된 속성들을 선택한다. 여러 데이터들을 적용하여 실험한 결과 분류의 정확성 측면에서 다른 속성 선택 기법에 비해 우수함을 증명하였다. 그렇지만 초기 속성 집합을 선택하는데 있어 속성수를 직접 결정하지 못하기 때문에 직접 속성수를 지정해주어야한다는 단점이 있고, 제안한 속성 선택 기법의 속도에 대한 우수성을 증명하지 않았다.

본 논문에서는 J. Cho[11]에서 제안한 속성 선택 기법과 유사하게 필터 방식과 래퍼 방식을 혼합한 방식을 적용한다. 상호 정보량을 통해 초기 후보 집단을 선택하고, 유전자 알고리즘으로 분류기에 최적화된 속성들을 선택한다. 그렇지만 초기 후보 집단을 선택하는데 있어 Fatem Amiri[10]에서 제안한 상호 정보량 기반 속성선택 기법을 적용하여 최적화된 초기 후보 집단을 선택한다. 단순 상호 정보량과 유전자 알고리즘을 적용하여 속성을 선택하는 것보다 빠르게 최적화된 후보 속성들을 선택할 수 있다.

3. 제안하는 하이브리드 침입 탐지 시스템

본 장에서는 제안하는 하이브리드 침입 탐지 시스템에 대해서 설명한다. 침입 탐지 시스템은 속성 선택 단계, 전처리 단계, 탐지 단계의 세단계로 이루어진다. 본 논문에서 제안하는 전체 시스템 구조는 아래 Fig. 1과 같다.

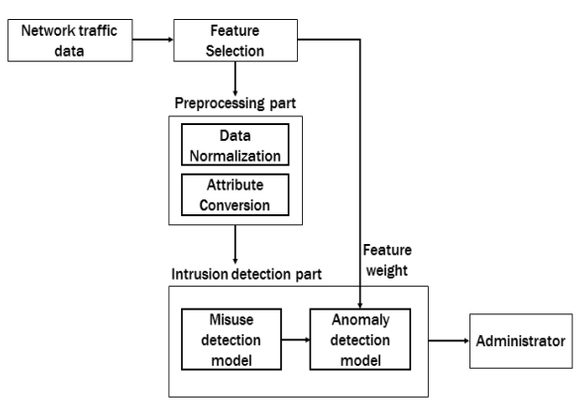

Fig. 1. Structure of Proposed Hybrid Intrusion Detection System

3.1 속성 선택 단계

제안하는 하이브리드 침입 탐지 시스템에서는 상호 정보량과 유전자 알고리즘을 혼합한 속성 선택 기법을 적용한다. 본 장에서는 먼저 상호 정보량과 유전자 알고리즘에 대해 각각 설명하고 제안하는 속성 선택 기법에 대해서 설명한다.

1) 상호 정보량

새논(Shannon)의 정보 이론에서는 랜덤 변수들 사이의 연관 정도를 나타내는 상호 정보량을 소개하였다. 상호 정보량은 기본적으로 랜덤 변수의 무질서도를 나타내는 엔트로피(Entropy) 개념을 기반으로 한다. 만약 랜덤 변수 X가 $p(x)=\Pr\{X=x\}, x\in\lambda$의 소스 알파벳 λ를 가진다면 X의 엔트로피는 아래 Equation (1)과 같다.

$$H(X)=-\sum_{x\in\lambda}p(x)\log p(x) \tag{1}$$

X, Y 사이의 연관성을 의미하는 상호 정보량은 아래 Equation (2)로 계산할 수 있다. X, Y가 상호 독립적이라면 이 값은 0이 되고, 연관성이 클수록 더 큰 값을 지닌다. 여기서 $p(x,y)$는 X, Y의 결합 확률 밀도 함수(joint probability density function)를 의미한다.

$$I(X;Y)=-\sum_{x\in\lambda}\sum_{y\in\delta}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

또한 여러 속성이 존재 할 때 속성 F와 전체 클래스 C와의 상호 정보량 $I(F;C)$를 계산할 수 있는데, 그 수식은 아래 Equation (3)과 같다.

$$I(F;C)=-\sum_{f\in\lambda}\sum_{c\in y}p(f,c)\log\frac{p(f,c)}{p(f)p(c)} \tag{3}$$

$I(F;C)$ 값이 클수록 전체 클래스에서 속성 F가 갖는 정보량이 많다는 것을 의미하고, 작을수록 속성 F가 갖는 정보량이 적다는 것을 의미한다.

2) 유전자 알고리즘

유전자 알고리즘은 자연의 진화과정에 기반으로 한 모델로 존 홀랜드(John Holland)가 개발한 최적화 탐색 기법이다. 유전자 알고리즘은 문제의 후보 해들을 유전자 형태의 구조로 표현하고 저장한 다음, 목적함수에 의해 점차적으로 진화시킴으로써 우수한 해 집합을 생성해낸다. 기본적으로 초기화, 적합도 평가, 재생산, 교배, 변이 단계로 이루어진다. 먼저 무작위로 해의 초기 집단을 선택하고, 선택된 집단에 대해 적합도 평가를 수행한다. 적합도 평가는 미리 설정한 목적함수 값의 비교를 통해 평가한다. 적합도가 우수한 해의 개체들을 확률적으로 더 많이 선택하고, 교배를 통해 재생산함으로써 해의 우수성을 유지한다. 또한 일부 변이를 통해 최적화된 해를 탐색한다.

3) 제안하는 속성 선택 기법

본 논문에서는 Fatem Amiri[10]에서 제안한 상호 정보량 기반의 기법과 유전자 알고리즘을 혼합한 속성 선택 기법을 제안한다. 먼저 상호 정보량을 기반으로 한 목적함수를 바탕으로 순차적으로 속성을 선택하고, 최종 속성 집합을 다음 유전자 알고리즘의 초기 속성 집합으로 선택한다. 제안하는 속성 선택 기법의 순서도는 아래와 같으며, 1차 속성 집합을 선택하는 목적함수와 유전자 알고리즘의 목적함수는 아래 Equation (4), (5)와 같다.

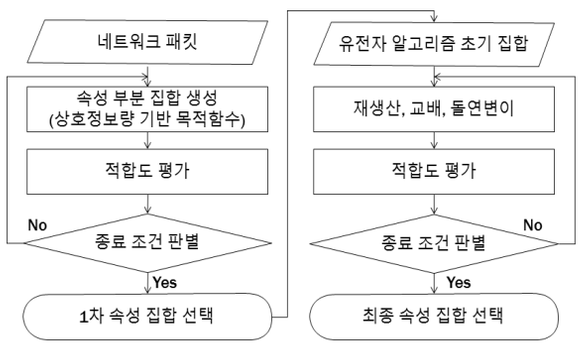

Fig. 2. Flowchart of Proposed Feature Selection

$$fitness\,function_i=I(f_i;y)-\beta/|s|\sum_{f_s\in S}I(f_i;f_s) \tag{4}$$

($I(f_i;y)$: 속성 i와 선택된 전체 클래스 y와의 상호 정보량, β : 가중 변수, s : 선택된 속성수, S: 선택된 속성 집합)

$$fitness\,function(s)=acc(s)-\frac{numS}{num\,Total} \tag{5}$$

($acc(s)$: 분류기의 정확도, $numS$: 속성 집합의 속성 수, $num\,Total$: 전체 속성 수)

여기서 분류기의 정확도는 전체 데이터에서 정확하게 분류된 데이터의 비율을 의미한다. 기존 래퍼 방식의 유전자 알고리즘으로 탐색할 경우에는 랜덤으로 초기 집합을 선정하지만, 제안 방식은 상호 정보량 기반으로 우수한 집합을 먼저 선정하여 더 적은 진화를 수행하여 탐색 시간을 효과적으로 줄일 수 있다. 필터 기법에 비해서는 느린 탐색 시간으로 속성을 선택하지만, 기존 래퍼 방식보다 빠른 탐색 시간으로 분류기에 최적화된 속성 부분 집합 선택이 가능하다.

3.2 전처리 단계

네트워크 패킷 데이터를 침입 단계의 분류 기법에 적용하여 분석하기 위해서는 분류 기법에 적절한 데이터 형식으로 전환해야 한다. 제안하는 하이브리드 탐지 기법의 가중 K-평균 알고리즘은 유클리디안 거리를 측정하여 군집화하기 때문에 이산형 값들을 연속형 값으로 전환시켜야 하고, 값들의 범위를 통일시켜야 한다. 이를 위해 탐지 단계 이전에 전처리 단계로 데이터 정규화 및 속성 값 치환을 실시한다.

1) 데이터 정규화

네트워크 패킷 데이터의 속성 값들은 속성별로 범위가 다르기 때문에 정규화가 필수적이다. K-평균 알고리즘에서 정규화하지 않고 거리 값을 계산할 경우 거리는 큰 값을 갖는 속성에 의존하게 된다. 보통 데이터 범위를 0 ~ 1 사이로 통일시키기 위해서 다음 Equation (6)을 적용해 정규화한다.

$$x_{after} = \frac{x_{before} - MIN}{MAX - MIN} \tag{6}$$

(x_{before}, x_{after} : 전, 후 속성 값, MIN : 해당 속성 차원의 값 중 가장 작은 값, MAX : 해당 속성 차원의 가장 큰 값)

2) 속성 값 치환

네트워크 패킷 데이터의 속성들은 이산형, 연속형 속성으로 나뉘며 속성 값 치환은 이런 이산형 속성 값들을 K-평균 알고리즘에 적용하기 위해서 연속형 속성 값으로 치환하는 과정이다.

이산형 속성 값을 연속형 속성 값으로 치환하는 방식 중 해당 속성의 값들을 0 ~ 1 사이에 동일한 간격으로 임의 분할시키는 방식을 고려할 수 있다. 그렇지만 이런 방식은 값에 따라 거리 측정에 오류를 범할 수 있다.

그러므로 여기서는 공격의 위협 정도를 고려해서 치환하고자 하는 속성 값에 대해서 그 값이 전체 클래스에서 침입 패킷이 얼마나 존재하는지에 대한 비율을 계산하고 그 값을 새로운 속성 값으로 치환한다. 이 치환 방식을 적용하면 공격의 위협 정도에 따라서 거리가 측정되기 때문에 침입을 탐지할 확률이 더 높아진다. 치환 수식은 다음 Equation (7)과 같다.

$$x_c = P(attack|x_d) \tag{7}$$

(x_c : 연속형 속성 값, x_d : 이산형 속성 값,
$P(attack|x_d)$: x_d 값이 전체 클래스에서 침입 패킷을 보유한 비율)

3.3 침입 탐지 단계

본 장에서는 침입 탐지 단계의 오용 탐지 모델과 이상 탐지 모델에 적용될 C4.5 의사 결정 트리와 가중 K-평균 알고리즘에 대해서 각각 설명하고 통합 모델인 하이브리드 침입 탐지 모델을 설명한다.

1) C4.5 의사 결정 트리

의사 결정 트리는 데이터 마이닝에서 일반적으로 사용되는 분류 알고리즘이다. 의사 결정 트리의 목표는 입력 변수를 바탕으로 목표 변수의 값을 예측하는 모델을 생성하는 것이다. 입력되는 훈련 데이터가 특정 클래스에 속하도록 구성될 때까지 분할 정복 방식을 재귀적으로 수행한다. 의사 결정 트리 기법 중 가장 보편적으로 사용되는 기법은 Quinlan[12]가 제안한 C4.5 의사 결정 트리 알고리즘이다. C4.5 알고리즘에서는 속성 중 정보 이익 비율(Information gain ratio)이 가장 높은 속성값을 선택하여 노드를 분리한다. 데이터 셋 X로부터 S가 n개의 부분 집합 S_1, S_2, ..., S_n으로 분리된다면 정보 이익(Information gain)은 아래 Equation (8)과 같다. 이를 통해 Equation (9), (10)으로 정보 이익 비율을 구할 수 있다.

$$Gain(X) = Entropy(S) - \sum_{i=1}^{n} |S_i|/|S| \times Entropy(S_i) \tag{8}$$

$$Gain ratio(X) = Gain(X)/Split\ Info(X) \tag{9}$$

$$Split\ \ Info(X) = \sum_{i=1}^{n} |S_i|/|S| \times \log_2(|S_i|/|S|) \tag{10}$$

위에서 설명한 C4.5의 분류 기준인 정보 이익 비율은 기존 노드 분류 기준이었던 정보 이익에 비해서 더 정확한 분류 결과를 제시한다[12].

본 논문에서는 하이브리드 침입 탐지 시스템의 오용 탐지 모델에 C4.5 의사 결정 트리를 적용한다. 네트워크 프로토콜 타입 별로 선택된 속성만을 사용하여 전처리된 훈련 데이터 셋을 학습하여 알려진 공격과 정상 데이터를 분류한다. 마지막 잎 노드에는 결정 트리에 의해 규칙이 정립된 공격 데이터들과 정상 데이터들의 부분 클래스로 분류된다.

2) 가중 K-평균 알고리즘

가중 K-평균 알고리즘을 설명하기에 앞서 그 이전 기본 버전인 K-평균 알고리즘을 설명한다. K-평균 알고리즘은 주어진 데이터를 K로 분할하는 군집화 기법으로 각 클러스터와 유클리디안 거리 차이의 분산을 최소화하는 방식으로 동작한다. 먼저 알고리즘은 임의의 K개의 중심(centroids)를 선정하고, 아래 두 가지 단계가 수렴할 때까지 진행된다.

- 단계 1 : 각각의 개체 (point)는 가장 거리가 가까운 중심으로 배정된다. 이 때 개체 사이의 거리 계산은 유클리디안 거리로 측정한다.
- 단계 2 : 배정된 개체를 포함한 클러스터의 중심을 다시 계산한다.

여기서 클러스터의 중심은 $\vec{\mu}$라 하며, 아래 Equation (11)로 정해진다.

$$\vec{\mu}(\omega) = \frac{1}{|w|} \sum_{\vec{x} \in \omega} \vec{x} \tag{11}$$

(ω : 클러스터에 속한 패턴 집합, $\vec{x}$: 클러스터에 속한 특정 패턴)

K-평균 알고리즘에서 목적 함수는 중심과 모든 패턴들에 대한 거리의 제곱 합이며, 이 값을 최소화하는 방향으로 군집화를 수행한다. 목적함수를 RSS (Residual Sum of Squares)라고 하며 이것을 구하는 Equation (12), (13)은 아래와 같다.

$$RSS_k = \sum_{\vec{x} \in \omega_k} |\vec{x} - \vec{\mu}(\omega_k)|^2 \tag{12}$$

$$RSS = \sum_{k=1}^{K} RSS_k \tag{13}$$

본 논문에서는 위에서 설명한 K-평균 알고리즘의 수정 버전인 가중 K-평균 알고리즘을 적용하여 이상 탐지 모델을 구축한다. 수정된 목적함수는 아래 Equation (14)와 같다.

$$RSS_k = \sum_{x \in \omega_k} \omega_v^{\beta} |\vec{x} - \vec{\mu}(\omega_k)|^2 \quad (14)$$

(ω_v : 속성 v에 대한 가중치 값, β : 사용자 변수)

속성 선택 단계에서 분석된 속성별 중요도를 가중 K-평균 알고리즘 목적함수의 유클리디안 거리에 가중치로 부여함으로써, 정상 데이터와 악성 데이터의 유형별로 정확한 군집화를 돕는다.

3) 하이브리드 침입 탐지 모델

침입 탐지 모델은 먼저 네트워크 유형(TCP, UDP, ICMP)에 따라 각각 선택된 속성차원에서 C 4.5 의사 결정 트리 기반의 오용 탐지 모델을 구축한다. 훈련 데이터 셋으로 결정 트리를 구축하여 정상 데이터와 알려진 공격 데이터의 부분 집합들로 분류함으로써, 알려진 공격에 대해서 결정 트리로 탐지하고, 새로운 공격은 이상 탐지 모델에서 나누어진 부분 집합들을 군집화 시켜 탐지한다.

다음으로 오용 탐지 모델의 마지막 잎 노드에 의해서 분류된 각각의 정상 데이터 부분 집합들에 대해 가중 K-평균 알고리즘 기반의 이상 탐지 모델을 구축한다. 이상 탐지 모델은 해당 네트워크 유형에서 분석된 속성들의 가중치를 반영하여 정상 데이터 부분 집합을 군집화 시키고 정상 데이터 클러스터 외에 다른 클러스터가 형성될 경우 이를 새로운 공격으로 간주한다. TCP 유형에 대한 침입 탐지 모델의 탐지 과정은 Fig. 3과 같다.

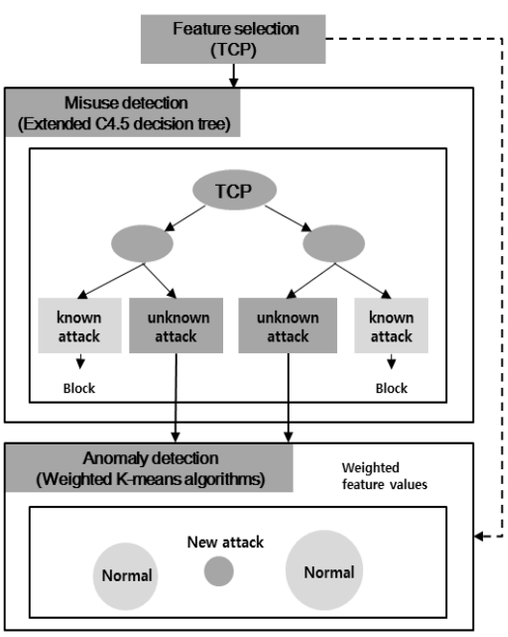

Fig. 3. Intrusion Detection Procedure of Proposed HIDS (TCP)

제안하는 하이브리드 구조가 갖는 장점은 다음과 같다. 첫째, 이상 탐지 모델의 훈련 시간과 테스트 시간을 줄일 수 있다. 보통 기존의 이상 탐지 시스템들이 오프라인 분석 모델로 구축된 주요 원인은 훈련 학습과 테스트의 과도한 계산량과 메모리 오버헤드 때문이었다. 이와 더불어 악성 데이터의 탐지 이후에 실시간적으로 탐지 모델을 업데이트하기 위해서는 훈련 학습 시간을 최소화 시켜야한다. 그러므로 실시간 침입 탐지 시스템 구축을 위해서는 이상 탐지 모델의 훈련 학습과 테스트 시간을 최소화 시키는 것이 필수적이다. 제안한 이상 탐지 모델은 전체 데이터 셋에 대해서 군집화 시키는 것이 아니라, 결정 트리에 의해서 속성 특성이 유사하게 가분류된 부분집합을 군집화 시킨다. 이미 일정 범위 내에 있는 데이터들로 분류되어 있기 때문에 K-평균 알고리즘의 군집화는 더 빨리 수렴한다. 궁극적으로 제안한 하이브리드 구조는 훈련 시간과 테스트 시간을 줄임으로써 실시간 침입 탐지 시스템에 적합한 구조라고 할 수 있다.

둘째, 제안 구조는 이상 탐지 모델의 긍정 오류율을 낮춤으로써 탐지율을 높일 수 있다. 기존 하이브리드 침입 탐지 방식들은 정상 행위 규칙들을 규정하기 위해서 대체로 하나의 이상 탐지 모델만을 형성하였다. 그렇지만 정상 데이터들의 패턴들은 매우 복잡하고 다양하기 때문에, 하나의 이상 탐지 모델만으로 전체 정상 데이터를 분석해 군집화 시키는 것은 오분류를 일으킬 수 있고 복잡도를 높인다. 제안 구조를 통해 분류된 부분 집합들은 전체 데이터 셋에 비해 적은 수를 분석하며, 패턴이 비슷한 데이터끼리 분류되어 있어 정확하게 탐지할 수 있다.

4. 성능 평가

본 장에서는 성능평가에 사용된 데이터 및 환경에 대해 설명하고, 제안한 하이브리드 침입 탐지 시스템 성능을 평가한다. 성능 평가에 앞서 실험 과정의 정확한 이해를 위해서 TCP 프로토콜 유형에 속하는 네트워크 패킷에 대한 실제 탐지 수행 과정의 예시를 Fig. 4와 같이 제시하였다.

<1. 속성 선택 단계>에서는 TCP 프로토콜 유형 데이터 집합을 분석한 결과 19개의 속성 부분 집합이 선택되었다. 각 속성은 속성 선택 시 계산되었던 상호 정보량 값을 가중치 값으로 갖는다. 다음 <2. 전처리 단계>에서 먼저 연속형 속성 중 속성 값 범위가 0~1 사이가 아닌 속성들에 대해 <2-1. 정규화>를 수행한다. 예시에서는 duration 속성에 대한 정규화 과정을 나타내었다. 이어서 <2-2. 속성 값 치환> 단계를 수행하는데, 예시에서는 service 속성의 telnet의 공격 위협 정도를 계산하여 연속형 속성으로 치환하였다. 전처리가 끝난 네트워크 패킷은 <3. 오용 탐지 모델>의 의사 결정 트리에 의해서 마지막 리프들 중 37번 리프로 분류된다. 마지막으로 <4. 이상 탐지 모델>에서는 37번 리프에 분류된 네트워크 패킷에 대하여 가중치를 적용한 K-평균 알고리즘을 수행하는데, 예시에서는 service 속성에 대한 상호 정보량 기반 가중치 값을 보여주었다. 각각 속성들의 가중치를 적용하여 K-평균 알고리즘을 수행하면 값의 범위에 따라 최종적으로 정상 또는 이상 군집에 속하는지 결정되고, 탐지를 마치게 된다.

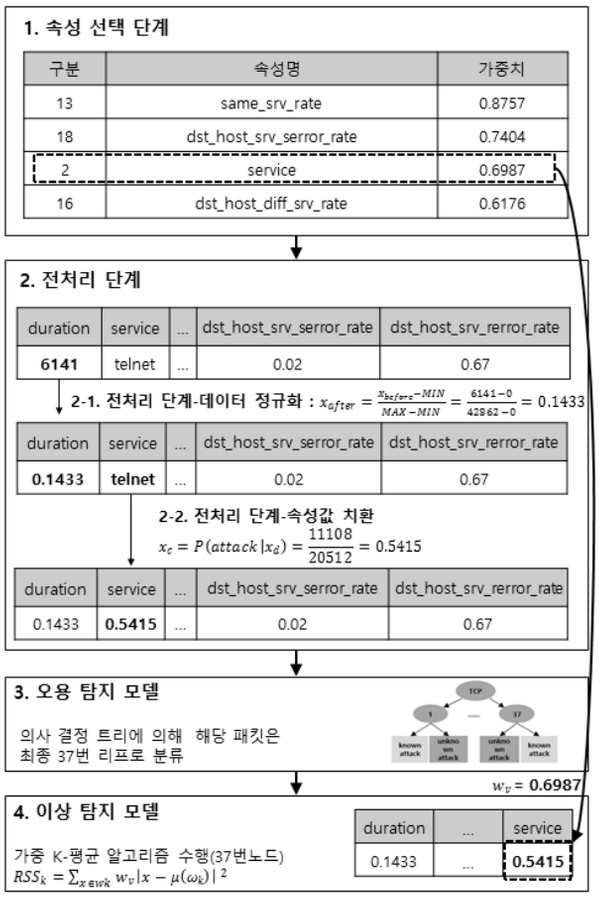

Fig. 4. Example of Intrusion Detection Procedure

4.1 실험 환경

본 논문이 제시한 하이브리드 침입 탐지 시스템의 성능을 측정하기 위해, 실험 데이터로는 침입 탐지 시스템 성능 평가 데이터 셋으로 잘 알려진 KDD' 99 (Knowledge Discovery & Data mining Cup)의 수정 버전인 NSL-KDD 데이터 셋[13]을 사용하였다. KDD Cup' 99 데이터는 미 국방부 산하 연구기관인 DARPA에서 침입 탐지 평가를 위해서 만든 데이터이다. 기존 KDD' 99 데이터 셋은 중복 데이터 문제로 정확한 실험이 어려웠으나, 근래에는 중복 데이터 문제를 개선한 NSL-KDD 데이터 셋이 제안되었으며, 침입 탐지 시스템의 실험 데이터로 활용되고 있다[14]. 또한 적용 기법은 데이터 마이닝 소프트웨어인 WEKA[15]를 통해 실험하였다. Table 1은 성능평가를 위한 실험환경 정보이다.

Table 1. Experimental Environment

	환 경
CPU	Intel i7-5930K
	3.50GHz
RAM	16GB
OS	Windows7
Data	NSL-KDD [13]

4.2 실험 결과

먼저 제안하는 하이브리드 침입 탐지 시스템의 속성 선택 기법의 우수성을 검증하기 위해, 훈련 데이터에서 선택한 속성의 탐지율과 오류율, 소요 시간을 타 기법들과 비교한다. 타 속성 선택 기법은 아래 속성 평가 지표와 속성 부분 집합 탐색 기법을 혼합 적용한 기법이며, 설명은 아래와 같다.

- 속성 평가 지표

- CfsSubsetEval[16] : 상관 관계 계수(Correlation relation coefficient)를 기반으로 속성 부분 집합을 평가하는 방식이다.
- ConsistencySubsetEval[17] : 속성 부분 집합의 전체 클래스에 대한 일관성 (Consistency) 수준을 측정한다. 일관성은 확률적 접근 방식으로, 전체 속성 집합에 대한 부분 속성 집합의 일관성을 측정하는 척도이다.
- ChiSquaredAttributeEval : 각 속성들의 전체 클래스에 대한 카이 제곱 테스트 (chi-squared test)를 수행하여 카이 제곱 통계 값을 측정하여 속성을 평가한다.
- GainRaioAttributeEval : 각 속성들의 전체 클래스에 대한 정보 이익 비율 (Information gain ratio)를 측정하여 속성을 평가한다.
- InfoGainAttributeEval : 각 속성들의 전체 클래스에 대한 정보 이익 (Information gain)을 측정하여 속성을 평가한다.

- 속성 부분 집합 탐색 기법

- PSOSearch[18] : PSO (Particle Swarm Optimization) 알고리즘을 활용하여 속성 부분 집합을 탐색하는 기법이다. PSO 알고리즘은 진화형 계산 기법의 일종으로 각 Particle이 집단 전체에서 발견한 목적함수 F, 해의 위치벡터를 공유하고 Particle들이 이동하며 집단 그룹을 진화시키는 방식으로 집합을 탐색한다.
- GeneticSearch : 유전자 알고리즘을 활용하여 속성 부분 집합을 탐색한다.
- BestFirst : Greedy hillclimbing, backtracking 방식으로 속성 부분 집합을 탐색하는 방식이다. 공집합부터 탐색하는 방식(hillclimbing) 전체 집합에서 탐색하는 방식(backtracking) 중 선택할 수 있다.
- Ranker : 속성 평가 지표에 따른 각 속성 별 순위를 매겨 속성을 선택한다.

비교 분석을 위해 각각의 속성 선택 기법들을 통해 훈련 데이터를 분석하여 속성을 선택했으며, 각 기법 별 20회씩 실험을 진행하였다. 20회의 속성 선택을 통해 결정된 20개의 속성 부분 집합에서 15회 이상 선택된 속성들을 최종 부분 집합으로 결정하였다. 또한 실험 시간의 경우 20회의 평균 탐지 시간을 최종 시간 값으로 결정하였다. Table 2에서도 볼 수 있듯이 실험 결과 제안 기법은 다른 기법에 비해 높은 탐지율과 낮은 오류율을 지님을 알 수 있다.

제안하는 속성 선택 기법은 4, 5, 6번 기법과 훈련 데이터에서 유사한 탐지율을 갖지만, 실험 데이터에서 2 ~ 6% 정도 우수한 탐지율을 보인다. 이것은 제안하는 속성 선택 기법이 분류 기법에 최적화 되어 있으며, 중요한 속성을 정확히

Table 2. Performance Comparison of Feature Selection Algorithms

구분	속성 선택 기법	개수	훈련 데이터			실험 데이터		
			DR (%)	ER (%)	Time (s)	DR (%)	ER (%)	Time (s)
1	CfsSubsetEval+ PSO	6	99.270	0.008	0.22	75.066	0.197	0.31
2	CfsSubsetEval + BestFirst	8	99.467	0.006	0.27	75.594	0.193	0.34
3	InfoGain+ Ranker	18	99.573	0.004	0.57	75.310	0.194	0.63
4	GainRatio + Ranker	18	99.688	0.003	0.56	78.269	0.171	0.67
5	ConsistencySubsetEval + GeneticSearch	19	99.683	0.003	0.54	80.540	0.154	0.56
6	ChiSquaredAttributeEval + Ranker	22	99.559	0.005	0.66	76.632	0.184	0.85
7	Proposed	17	99.719	0.003	0.34	82.301	0.141	0.40
8	All features	41	99.630	0.004	1.23	81.054	0.150	1.27

선택했다고 볼 수 있다. 또한 제안 기법은 4, 5, 6번 기법에 비해 탐지 시간을 단축할 수 있었다.

또한 전체 속성을 선택하여 탐지한 것과 6번 기법(22개)들로 탐지한 것보다 4, 5, 제안 기법의 속성으로 탐지한 것이 더 높은 탐지율을 보였는데(훈련 데이터), 이것은 속성의 수가 분류 문제에서 중요한 것이 아니며 의미 있는 속성을 선택하는 것이 중요하다는 것을 보여준다. 또한 탐지 시간을 분석해보면 각 기법들의 선택된 속성 개수에 비례하여 증가한다는 것을 알 수 있다.

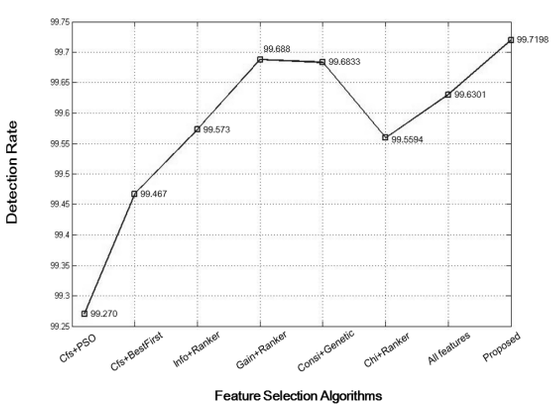

Fig. 5. Detection Rate Comparison of Feature Selection Algorithms (Training Data)

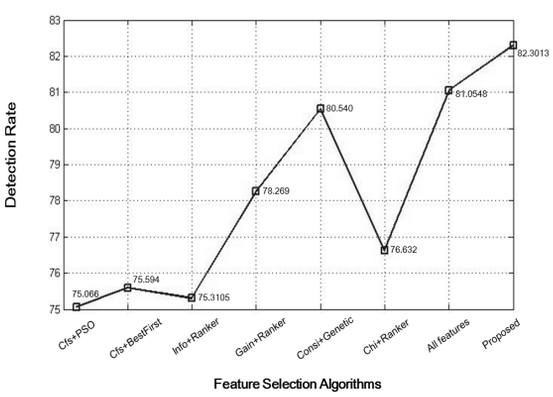

Fig. 6. Detection Rate Comparison of Feature Selection Algorithms (Testing Data)

다음은 제안하는 하이브리드 침입 탐지 하이브리드 시스템의 성능을 탐지 정확성 측면, 속도 측면에서 평가한다. 비교군은 단순히 두 모델을 결합한 하이브리드 방식(C 4.5+K-Means), 단순 결합 모델에 속성 선택 기법을 적용한 하이브리드 방식(FS+C4.5+K-Means), 제안 방식에서 속성 선택을 제외한 하이브리드 방식(Proposed-FS)이며, 가중치의 영향을 평가하기 위해 제안 방식에서 가중치를 적용하지 않은 방식(Proposed-FW)를 비교하였다. 또한 탐지 정확성을 평가하기 위해 탐지율, 정확도, 오류율을 비교하였고, 속도 측면에서의 성능을 평가하기 위해 훈련 시간과 실제 실험 시간을 측정하여 비교한다. 위 속성 선택 기법 시간 측정 방식과 마찬가지로 각 모델별 20회씩 수행하여, 20회 수행한 평균 훈련 시간과 실험 시간을 최종 시간 값으로 결정하였다. 성능 평가 결과는 Table 3과 같다.

Table 3. Performance Comparison of HIDSs

구분	탐지 모델	DR (%)	ACC (%)	ER (%)	Training time(s)	Testing time(s)
1	C 4.5 + K-Means	92.281	94.533	6.088	30.27	4.80
2	FS+C 4.5+K-Means	90.122	93.213	7.102	14.12	2.25
3	Proposed-FS	99.353	99.421	0.822	20.51	4.51
4	Proposed-FW	98.454	98.632	1.361	11.10	1.79
5	Proposed	98.685	99.081	1.326	11.57	1.87

제안하는 하이브리드 침입 탐지 시스템은 단순 결합 모델인 1, 2에 비해 높은 탐지율과 정확도, 낮은 오류율을 지니며, 이것은 제안 시스템이 탐지 정확성 측면에서 단순 결합 모델보다 더 우수한 탐지 모델임을 알 수 있다. 속성 선택 기법의 영향을 알아보기 위해 Proposed (5번)와 Proposed-FS (3번)을 비교하면, 제안 시스템이 탐지율은 3번 시스템에 비해 0.668% 낮았지만 훈련 시간에서 3번 시스템이 20.51초 인데 비해 제안 시스템은 11.57초로 8.94초 빠르게 모델을 구축하였다. 또한 실험시간에서도 3번 시스템이 4.51초임에 비해 제안 시스템은 1.87초로 2.64초 빠르게 탐지를 수행했다. 이것은 제안 시스템의 구조에 속성 선택 기법을 적용할 경우 탐지율은 소폭 감소할 수 있지만, 모델 구축 시간과 실제 탐지

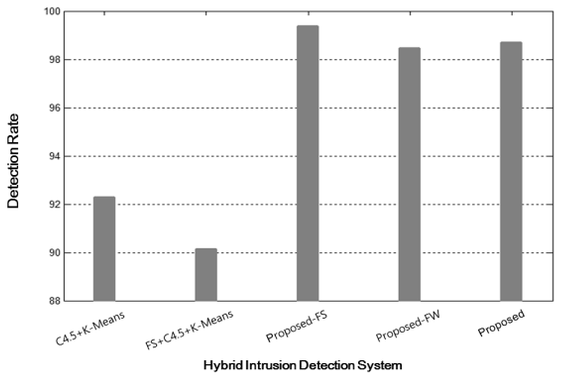

Fig. 7. Detection Rate Comparison of HIDS

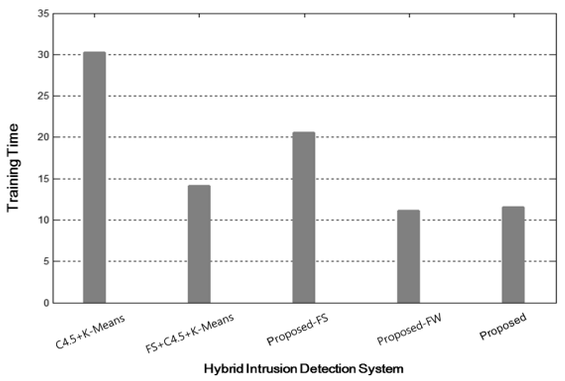

Fig. 8. Training Time Comparison of HIDS

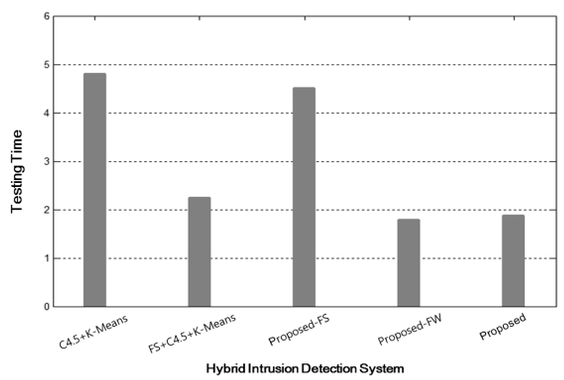

Fig. 9. Testing Time Comparison of HIDS

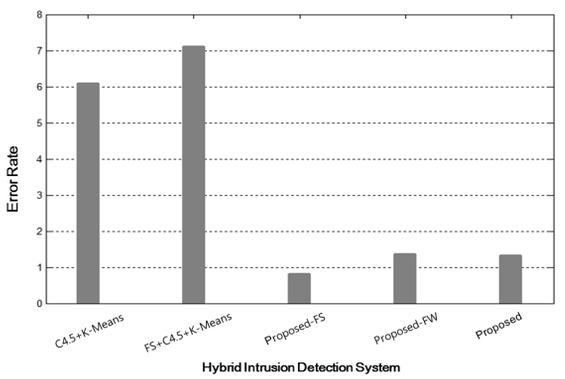

Fig. 10. Error Rate Comparison of HIDS

시간 측면에서 큰 성능 향상을 보임으로써, 제안 시스템이 비교적 높은 탐지율을 보장함과 동시에 고성능 침입 탐지를 수행할 수 있다는 것을 의미한다.

그리고 가중치의 영향을 분석하기 위해 Proposed (5번)와 Proposed-FW (4번)을 비교하면, 두 시스템은 탐지 정확도와 속도 측면에서 유사한 결과를 보이나 탐지 정확도 측면에서는 제안 시스템이, 속도 측면에서는 4번 시스템이 근소하게 우수한 것을 볼 수 있었다. 더 정확한 의미를 판단하기 위해 모델 구축을 20% 훈련 데이터로 구축하고 테스트 했을 때, 4번은 98.127%, 제안 시스템은 97.153%의 탐지율을 보였다. 이것은 분류 모델을 구축할 때 속성의 가중치를 반영하기 위해서는 속성의 정확한 가중치를 판단할 수 있는 충분한 양의 훈련 데이터와 우수한 기법이 필요함을 의미한다.

5. 결 론

본 논문에서는 기존의 하이브리드 침입 탐지 시스템의 성능 문제를 보완할 수 있는 효과적인 속성 선택 기법을 적용한 의사 결정 트리와 가중 K-평균 알고리즘 기반의 고성능 하이브리드 침입 탐지 시스템을 제안하였다. 속성 선택 기법을 적용하여 침입을 더 빠르고 효율적으로 탐지할 수 있으며, 오용 탐지 모델과 이상 탐지 모델을 위계적으로 결합하여 구조적으로 고도화된 하이브리드 침입 탐지 시스템을 제안하였다. 실험을 통해 제안한 하이브리드 침입 탐지 시스템이 탐지 정확도 측면에서 우수함과 동시에 속도 측면에서도 뛰어난 성능을 지님으로써, 제안한 시스템이 근 실시간 네트워크 환경에 적합한 침입 탐지 시스템임을 검증하였다.

향후 연구로는 제안한 하이브리드 침입 탐지 시스템에서의 가중치에 대한 다양한 추가 실험을 진행하고, 발전 요소를 도출하고자 한다. 침입 탐지 시스템에 적합한 가중치 추출 기법과 각각의 속성들의 중요도를 정확하게 가중치로 반영할 수 있는 기법에 대한 연구가 필요하다. 또한 전체 침입 탐지 시스템 구조를 병렬 및 분산화 하여 성능을 향상시키는 연구가 필요하다. 이를 통해 실제 네트워크 환경에서 실시간 침입 탐지를 수행할 수 있는 빠른 성능을 보유하고, 높은 탐지율과 낮은 오류율을 지닌 침입 탐지 시스템을 구축할 수 있을 것이라 기대한다.

References

[1] Dongil Seo and Hyeonsook Cho, “The present and future of Security Technology for Cyber-Warfare,” *Review of KIISE*, Vol.21, Iss.6, 2011.

[2] S. Noel, D. Wijesekera, and C. Youman, “Modern Intrusion Detection, Data Mining, and Degrees of Attack Guilt,” in Applications of Data Mining in Computer Security, Kluwer Academic Publisher, pp.1-31, 2002.

[3] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, “An intelligent intrusion detection system for anomaly and misuse detection in computer networks,” *Expert Sytems with Applications*, Vol.29, No.4, pp.713-722, 2005.

[4] J, Zhang and M. Zulkernin, “A hybrid network intrusion detection technique using random forests,” in *Proceedings of the First International Conference on Availability, Reliability and Security*, pp.262-269, 2006.

[5] Gisung Kim, Seungmin Lee, and Sehun Kim, “A novel hybrid intrusion detection method integrating anomaly detection with misuse detection,” *Expert Systems with Applications*, Vol.41, Iss.4, Part 2, pp.1690-1700, 2014.

[6] Xin Xu and Xuening Wang, “An adaptive network intrusion detection method based on PCA and support vector machines,” *International Conference on Advanced Data Mining and Application*, 2005.

[7] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, and C. Molter, “A Survey of filter techniques for feature selection in gene expression microarray analysis,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Vol.9, Iss.4, pp.1106-1119, 2012.

[8] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers and Electrical Engineering*, Vol.40, Iss.1, pp.16-28, 2014.

[9] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, “Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset,” *Expert Systems with Applications*, Vol.38, Iss.5, pp.5947-5957, 2011

[10] F. Amiri, M.M.R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, “Mutual information-based feature selection for intrusion detection systems,” *Journal of Network and Computer Applications*, Vol.34, Iss.4, pp.1184-1199, 2011

[11] Jae Hoon Cho, Dae Jong Lee, Chang Kyu Song, Yong Sam Kim, and Myung Geun Chun, “Feature Selection by Genetic Algorithm and Information Theory,” *Journal of Korean Institute of Intelligent Systems*, Vol.18, No.1, pp.94-99, 2008.

[12] J. R. Quinlan, “Learing decision tree classifiers,” *ACM Computing Surveys (CSUR)*, Vol.28, Iss.1, pp.71-72, 1996.

[13] UNB ISCX NSL-KDD Dataset [Internet], http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html.

[14] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in *Proceeding of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp.53-58, 2009.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” *ACM SIGKDD Explorations Newsletter*, Vol.11, Iss.1, pp.10-18, 2009.

[16] M. A. Hall, “Correlation-based Feature Subset Selection for Machine Learning,” Ph.D. dissertation, The University of Waikato, Canada, 1999.

[17] H. Liu and R. Setiono, “A probabilistic approach to feature selection - A filter solution,” in *13th International Conference on Machine Learning*, pp.319-327, 1996.

[18] A. Moraglio, C. Di Chio, and R. Poli, “Geometric Particle Swarm Optimisation,” in *Proceedings of the 10th European Conference on Genetic Programming*, pp.125-136, 2007.

이 우 솔

e-mail : lee99woo@gmail.com
2011년 육군사관학교 운영분석학과(학사)
2016년 아주대학교 NCW학과(석사)
2016년~현 재 육군 3사관학교 컴퓨터공학과 강사
관심분야 : 정보보호, 침입 탐지 시스템, 분산/병렬 컴퓨팅

오 상 윤

e-mail : syoh@ajou.ac.kr
2006년 미국 인디애나대학교 컴퓨터공학과 (박사)
2006년~2007년 SK텔레콤
2007년~현 재 아주대학교 소프트웨어학과 교수
관심분야 : 분산/병렬 시스템, 고성능컴퓨팅, Large Scale Software System, Semantic Web