Proceedings of KFIS Spring Conference 2006 Vol. 16, No. 1.

# 침입탐지시스템에서 하이브리드 특징 선택에 관한 연구

# A Study on Hybrid Feature Selection in Intrusion Detection System

한명묵[1]

[1] 경기도 성남시 경원대학교 소프트웨어대학
E-mail: mmhan@kyungwon.ac.kr

**요 약**

네트워크를 기반으로 한 컴퓨터 시스템이 현대 사회에 있어서 더욱 더 불가결한 역할을 하는 것에 따라, 네트워크 기반 컴퓨터 시스템은 침입자의 침입 목표가 되고 있다. 이를 보호하기 위한 침입탐지시스템(Intrusion Detection System : IDS)은 점차 중요한 기술이 되었다. 침입탐지시스템에서 패턴들을 분석한 후 정상/비정상을 판단 및 예측하기 위해서는 초기단계인 특징추출이나 선택이 매우 중요한 부분이 되고 있다. 본 논문에서는 IDS에서 중요한 부분인 feature selection을 Data Mining 기법인 Genetic Algorithm(GA)과 Decision Tree(DT)를 적용해서 구현했다.

**Key Words** : Intrusion Detection System, Data Mining, Genetic Algorithm, Decision Tree, Feature Selection

## 1. 서 론

네트워크를 기반으로 한 컴퓨터 시스템이 현대 사회에 있어서 더욱 더 불가결한 역할을 하는 것에 따라, 네트워크 기반 컴퓨터 시스템은 적과 범죄에 의한 침입 목표가 되었다. 그러므로 기밀에 관련되는 정보가 저장된 것과 온라인 조작되고 있는 것에 따라 네트워크 시스템의 안전은 더욱 더 중요하게 되고 있다. 침입탐지시스템은 이와 같이 우리 시스템을 보호하는 것을 돕기 위한 중요한 기술이 되었다.

침입 탐지 기술은 비정상행위 탐지와 오용탐지로 분류할 수 있다. 비정상행위 탐지 시스템은 정상적인 사용자 프로파일로부터 크게 벗어나는 활동에 주목한다. 정상적인 시스템 사용에 관한 프로파일에서 벗어나는 행위들을 탐지한다. 오용 참지는 시스템의 알려진 취약점들을 이용한 공격 행위들에 대한 공격 특징 정보를 통해 침입을 탐지한다.

대부분의 침입탐지시스템(Intrusion Detection System:IDS)은 전문 지식의 수동의 기호화에 의해 개발되는 수제 서명에 의거한다. 이 시스템은 공격의 알려진 서명에 감시되고 있는 시스템 위에서 활동에 필적한다. 이 접근에 관한 주요한 문제는 시스템이 새로운 공격을 찾기 위해 개괄할 수 없다라는 것이다. 최근, 관심이 되는 것이 IDS의 발견 모델을 형성한다 것에 데이터 마이닝을 기반으로 둔 것이 있었다. 이 방법은 알려진 공격과 보통의 행동의 그들의 모델을 알려지지 않은 공격을 찾기 위해 일반화할 수 있으며, 그들은 또한 도메인 전문가에 의해 감사 자료의 어려운 분석을 필요로 하는 수동으로 코드화 되었던 모델보다 더 빠르고 더 자동화되었던 방법으로 생성될 수 있다. 침입을 찾는 효과적인 몇 개의 데이터 마이닝 기술은 개발되었다[1][2][3][4].

이 논문에서는 IDS에서 중요한 특징 선택에 관한 방법에서 Data Mining 기법 중 GA와 DT를 활용한 hybrid방법을 활용해서 적용한다. 본 논문의 구성은 2장에서 특징선택 문제에 대해서 서술하며 3장에서 이 논문의 기본이 되는 GA, DT를 간략히 설명한 후 Hybrid 시스템에 대해서 논하고 4장에서는 실험과 그 결과를 분석한다. 마지막으로 5장에서 결론을 맺는다.

## 2. 특징선택 문제

분류과정에서의 첫 번째 단계 중에 하나는

좀 더 많고 원래의 특징 집합 중에서 작은 특징의 부분집합을 선택하는 특징 선택이다. 이러한 특징 선택 방법은 무엇을 평가에 사용되는 지에 따라 필터(filter)와 래퍼(wrapper)접근 방법으로 나눈다. 필터 접근 방법은 분류과정 전에 예제들 사이에서 어떤 거리의 측정을 기반으로 특징의 부분집합을 선택하며, 래퍼 접근 방법은 분류과정에서 분류의 결과를 기반으로 특징의 부분집합을 선택한다.

통계학[5], 기하학[6], 기계학습 등을 포함한 여러 방법을 통해서 많은 연구자들이 다양한 특징 추출 방법들을 개발해 왔다.

통계학 방법에서는 forward 와 backward stepwise multiple regression(SMR)이 특징을 선택하는데 사용되어져 왔다. 특히 forward 방법이 backward 방법보다 계산의 복잡도가 적기 때문에 forward 방법이 선호되어 왔다.

기하학 방법에서는 탐색공간에서의 예제들의 위치들이 결정트리를 위한 특징들을 선택하기 위해서 IDG 알고리즘에 입력이 된다. 다른 클래스로부터 경계 예제가 분리된 규칙들은 보상을 받고, 같은 클래스로부터 경계 예제가 분리된 규칙들은 벌칙을 받는다.

기계학습 방법에서는 Sequential Forward Search(SFS) 와 Sequential Backward Search(SBS) 그리고 이것들의 여러 변형된 방법들이 사용되어져 왔다. SFS는 비어있는 집합에서 시작하여 국지적으로 가장 좋은 특징이 집합에 첨가된다. SBS는 완전한 집합에서 시작하여 국지적으로 가장 나쁜 특징이 집합에서 제거된다. 교사학습 알고리즘으로 학습된 뉴럴 네트워크는 misuse 탐지에 적용이 되고, 비교사학습 알고리즘으로 학습된 뉴럴 네트워크는 anomaly 탐지에 적용이 된다. 퍼지 집합이론을 활용해서 FuzzyARTMAP이 적용이 되었고, 러프 집합이론을 활용해서 PRESET이 이진 집합들을 선택하기 위해서 특징의 의존도를 결정한다. GA는 특징 집합에서 특징의 존재 여부를 나타내는 1과 0을 가진 비트열로서 특징 집합을 encoding함으로써 사용되어진다.

## 3. Genetic Algorithm과 Decision Tree를 이용한 Hybrid 시스템

### 3.1 Genetic Algorithm

GA는 유전적 계승과 다윈적 생존 경쟁이라는 자연 현상을 모델링한 확률적인 탐색방법으로 유전검색이 불가능할 정도로 큰 후보해 공간을 갖는 최적화문제에 적용할 수 있다. 즉, 해가 될 가능성이 있는 개체집단을 유지함으로써 여러 방향의 탐색을 실행하고 이들 방향간의 정보형성과 교환을 행한다. 개체집단은 진화과정을 모방하는데, 각 세대에서 비교적 우량한 해들이 재생산되고, 반면에 비교적 불량한 해들은 소멸된다. 또한 다른 해들간의 차이를 구별하기위해 환경의 역할을 수행하는 목적함수를 사용한다. 이러한 유전자 알고리즘은 특정한 문제에 대해 다섯 가지의 요소를 가져야만 한다. 유전자적 표현방법, 초기 개체집단을 만들어 내는 방법, 목적함수, 유전 연산자, 그리고 여러 가지 매개변수의 값이다.

어떤 개체집단을 초기화하기 위해서는 단순히 개체집단의 염색체를 비트단위로 임의로 설정 할 수 있다. 혹은 가능한 최적값들의 분포에 관한 지식을 가지고 있다면 초기의 해집합을 배열하는 데 그 정보를 이용할 수 있다.

알고리즘의 나머지 부분은 각 세대에서 각각의 염색체를 평가하고, 적합도값에 기초한 확률분포에 의하여 새로운 개체집단을 선택하며, 돌연변이와 교배연산자에 의하여 새로운 개체집단의 염색체들을 변화시킨다. 여러 세대 후에 더 이상의 개선이 없으면, 그 세대의 가장 좋은 염색체가 최적해를 나타낸다. 선택과정에서는 적합도에 비례해서 가장 좋은 염색체는 더 많이 복제되고, 보통 염색체는 비슷하게 남아 있으며, 최악의 염색체는 소멸된다. 교배연산자는 교배연산확률을 토대로 두개의 염색체에 적용되서 새로운 두 개의 자손을 생산하며, 마지막으로 돌연변이 연산자가 돌연변이 확률에 의해 비트별로 적용된다. 이러한 선택, 교배, 그리고 돌연변이를 한 후에 새로운 개체집단은 다음 평가를 받는다.

### 3.2 Decision Tree

결정 트리는 분류화 작업과 예측을 하는데 있어서 강력하고 가장 많이 사용하는 도구이다. 트리를 기반으로 한 방법인 결정트리나 뉴럴 네트워크의 경우는 규칙(rules)으로 표현이 되는데, 이런 규칙은 읽을 수 있도록 영어로 표현이 된다. 따라서 접하는 사람들은 쉽게 이해할 수 있거나 혹은 데이터베이스에 접근하는 언어인 SQL로 표현이 되는데, 이런 레코드들은 특별한 카테고리를 회수하는 것에는 상당히 약하다. 이러한 응용에서는 분류화나 예측에서의 유일한 문제는 바로 정확성이다.

결정 트리는 전형적으로 루트, 즉 위에서 아래로 향하는 형태를 가지게 된다. 트리의 루트 노드에 레코드를 집어넣게 되면 어떤 자식 노드로 분기를 할 것인지를 결정을 하게 된다. 이러한 절차를 리프 노드에 도달할 때까지 계속적으로 이루어지게 된다. 또한 각각의 리프

노드로 가는 길은 유일하며 그 길은 레코드를 분류하는 규칙으로 표현이 되게 된다.

결정 트리를 구축하는데 있어서 많은 알고리즘이 존재하는데, 그 중에 가장 일반적인 것은 C4.5 이다. C4.5와 같은 계통인 ID3는 잠재적인 분기를 비교하여 information gain이라고 불리는 특징을 사용한다. 이것의 중요한 개념은 특정 상황이나 가능한 결과들의 집합의 크기에 의존적인 결과를 많은 수의 비트로 표현하는 것을 요구된다. 만일 여덟 개의 동일하게 예상되는 클래스가 있다면, 어떤 특정한 것을 식별하기 위해서는 세 개의 비트 혹은 log2(8)이 소요된다. 반면에, 표현된 클래스가 4개만 있다면, 그들 중 하나를 식별하는데는 log2(4) 혹은 두 개의 비트만이 필요하다. 그래서, 노드에서의 분기는 여덟 개의 클래스의 예와 평균 4개의 클래스로 분기하는데, 각각의 클래스는 한 비트의 information gain을 갖는다. 분기 평가 특징으로써 information gain은 매우 무성한 결정 트리로 향하는 높은 성향을 가지고 있다.

2 단계의 ID3 결정트리는 그림 1과 같다.

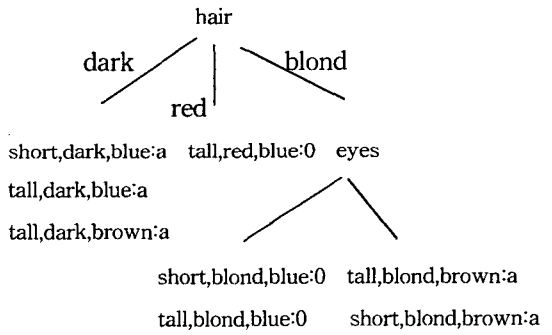


그림 1. 2단계의 ID3 결정트리

### 3.3 Hybrid System

제안하는 hybrid 시스템의 기본 개념은 GA를 활용해서 적은 계산량과 높은 분별력의 특징 부분집합을 발견하기 위해 주어진 전체 특징 집합에서 가능한 모든 부분집합의 공간을 효율적으로 탐색하는 것이다. 이를 달성하기 위해 적응 평가는 크기와 분류 성능에 직접적으로 관여해야만 한다.

특징들의 초기집합은 측정되어진 특징 벡터들의 훈련 집합하고 같이 주어진다. GA는 좀 더 작은 특징 집합을 활용해서 좀 더 나은 분류 성능을 달성하는 특징 부분집합의 모든 공간을 탐색한다. 선택되어진 각 특징 부분집합은 C4.5에 의해 생성되어진 결정트리를 테스트함으로써 평가되어진다. 이 과정은 진화단계에서 반복적으로 이루어지며, 가장 우수한 특징 부분집합은 패턴 분류 시스템의 실질적인 설계에 사용되어진다.

GA가 매우 큰 공간을 효율적으로 탐색하기 위해, 표현 방식과 평가함수의 선정에 주의를 기울여야 한다. 이 논문에서는 특징 집합의 모든 가능한 부분 집합들의 공간을 자연적으로 표현하는 방법을 사용하였다. 즉, 고정된 이진 스트링 표현에서 i번째의 0은 그 특징이 포함되어 있지 않으며 1은 특징이 포함된 것을 의미한다. 따라서 GA 집단의 각 individual은 주어진 특징 집합의 부분 집합을 표현하는 고정된 길이의 이진 스트링으로 구성된다.

GA 집단의 각 individual은 진화 과정에 적응도를 얻기 위하여 평가되어지는 특징 부분 집합들의 경쟁 형태를 표현한다. 이는 특정의 특징 부분 집합과 훈련 집합을 포함하는 C4.5를 생성함으로써 얻어진다. 그 후 C4.5에 의해 생성된 결정 트리는 알려져 있지 않은 평가 데이터에서 분류의 정확도를 위해 테스트된다. 정확도는 특징 부분집합의 크기와 더불어 GA 평가 측정으로써 사용되어진다.

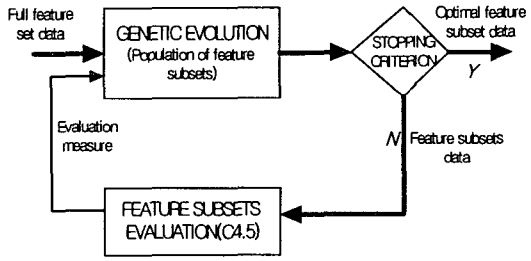


그림 2. GA와 DT를 활용한 Hybrid 시스템

## 4. 실험 및 고찰

본 장에서는 실험에 사용된 데이터를 분석하고 정리한다. 실험에서 사용된 데이터는 1999년 "KDD'99 Competition: Knowledge Discovery Contest"에서 제공된 것을 활용하였다. 1998년 DARPA 침입탐지 개발 프로그램은 MIT Lincoln Labs에서 준비되었고 관리되어져 왔다. 여기서 제공되어진 데이터는 군사 네트워크 환경에서 실험되어진 방대하고 다양한 침입들을 포함하고 있는 표준 감사 데이터 집합(data set)들이다. 이후 1999년 "KDD Intrusion Detection contest"는 바로 이 데이터 집합을 활용하여 진행되었다.

이 데이터의 속성은 총 41개로 구성되어있

다. 하나의 레코드에 41개의 속성 값으로 구성되어 있으며 이는 정상적인 데이터와 비정상적인 데이터에 동일하게 구성되어져 있다. 실험에서 이러한 데이터 집합을 training data와 test data로 구분한 후 decision tree의 error rate를 GA의 fitness value 값으로 정해서 적절한 특징들의 부분집합들을 구했다. 그 결과는 그림 3과 같다.

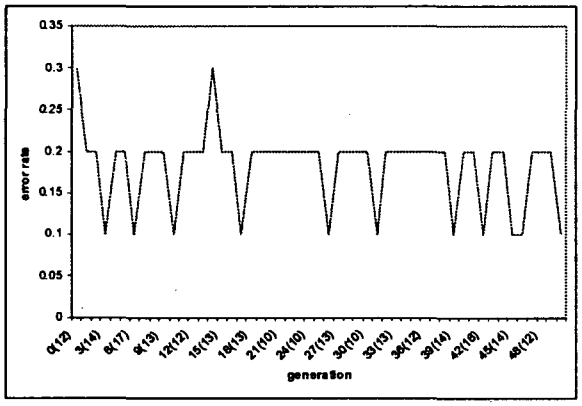


그림 3. Hybrid 시스템의 결과

## 5. 결 론

본 논문에서는 침입 탐지 시스템의 성능 향상을 위해 데이터 마이닝을 이용한 특징 선택을 하였다.

데이터 마이닝의 주요 기법 중 genetic algorithm과 decision tree를 사용하여 주어진 침입 데이터 집합에서 총 41개 특징 중 14개의 특징으로 분류를 할 수가 있었다.

주어진 14개의 특징들은 decision tree의 error rate에 의해서 평가를 하였으며, 41개를 활용한 결과와 비슷한 결과를 얻었다. 이로써 다양하고 서로 성질이 다른 공격들에 대한 속성들을 선택함으로써 정상적인 사용자의 행위를 규칙화 할 때, 필수적인 속성들만 연산에 포함됨으로 연산시간을 단축시키며, 보다 정확한 규칙 생성에 중요한 정보로 제공되어 침입 탐지 시스템의 성능 향상에 큰 도움을 줄 것이다.

향후 연구과제로는 다른 분류 시스템을 적용함으로써 더욱 향상된 방법에 대한 연구가 이루어 질 것이다.

## 참 고 문 헌

[1] Wenke Lee, Salvatore J. Stolfo, Data Mining Approaches for Intrusion Detection, Proceedings of the 7th USENIX Security Symposium, San Antonio, Jan. 1998.

[2] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, A Data Mining Framework for Building Intrusion Detection Models, IEEE Symposium on Security and Privacy, 1999.

[3] Terran D. Lane, Machine Learning Techniques For The Computer Security Domain Of Anomaly Detection, A thesis Submitted to the Faculty of Purdue University. August 2000.

[4] Wenke lee, Salvatore J.Stolfo, A Framework for Construction Features and Models for Intrusion Detection Systems, Proceedings of the1999 IEEE Symposium on Security and Privacy and the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,1999.

[5] Kittler, J., Mathematical Methods of Feature Selection in Pattern Recognition, International Journal of Man-Machine Studies, 7, pp. 609-637, 1975.

[6] Elomaa, T., and E. Ukkonen, A Geometric Approach to Feature Selection, In Proceedings of the European Conference on Machine Learning, pp.351-354, 1994.