퍼지 의사결정 나무를 이용한 네트워크 증거 분석

이용현*, 이승용*, 김동국**, 노봉남*
*전남대학교 정보보호협동과정
**전남대학교 전자컴퓨터공학부
dazzilee@lsrc.chonnam.ac.kr

Network Forensic using Fuzzy Decision Tree

Yong-Hyun Lee*, Seong-Yong Lee*, Dong-Kook Kim**,
Bong-Nam Noh*
*Interdisciplinary Program of Information Security, Chonnam National University,
**Dvision of Electronics Computer Engineering, Chonnam National University

요 약

컴퓨터의 생활 전반에 걸친 영향으로, 컴퓨터는 우리 생활 속에서 빼놓을 수 없는 하나의 정보 매체로 자리 매김 되었다. 하지만 그 이면에는 컴퓨터를 이용한 전산망 침해 행위, 전자기록 위·변조, 각종 음란물 유통, 바이러스 제작 유포 등 많은 위험들이 우리를 위협하고 있다. 그래서 컴퓨터를 사용한 범죄 행위를 탐지하는 방법에 대한 관심이 높아지고 있다. 또한 각종 범죄 행위는 인터넷을 통한 범죄가 늘고 있어, 네트워크 정보를 통한 포렌식에 관한 연구가 활발하다. 하지만, 매일 많은 양의 패킷을 분석하는 것은 많은 전문 인력과 비용이 소요된다. 본 논문에서는 의사결정나무를 이용한 패킷분석을 통하여 네트워크 포렌식의 정보를 추출하는 방법을 제안한다.

1. 서론

인터넷의 보급과 급속한 성장은 정보통신의 발달과 함께, 우리 생활 속에서 빼놓을 수 없는 하나의 정보 매체로 자리매김 되었다. 하지만, 매우 높은 발전과 더불어, 컴퓨터를 이용한 많은 범죄가 늘어나고 있다. 표 1은 경찰청 사이버 테러대응센터에 발표한 사이버 범죄 현[1]황를 표로 나타낸 것이다. 표에서 알 수 있듯이 5년 사이에 2배이상의 범죄 행위가 발생하고 있다[5].

또한 그 범죄는 컴퓨터의 기술의 증가에 따라 더 정교해 지고 있다. 디지탈 포렌식은 법정에서 증거로 쓰일 수 있게 정보의 추출 및 발견을 하는 것이다. 특히, 네트워크 포렌식은 네트워크 환경에서의 범죄에 사용된 증거를 추출하는 것이다.

* 본 연구는 정보통신부 대학 IT 연구센터 육성, 지원사업의 연구결과로 수행되었습니다.

<표 1> 경찰청 사이버테러대응센터

구분	사이버테러형 범죄	일반사이버 범죄
2006	20,186	62,000
2005	21,389	67,342
2004	15,390	61,709
2003	14,241	54,204
2002	14,159	45,909
2001	10,638	22,651

하지만, 매일 많은 양이 쌓이는 네트워크 정보에서 증거 획득 한다는 것은 많은 인원과 비용이 소요되는 어려움이 있다. 그러므로 데이터 마이닝 기법을 이용하여 정교하고 빠른 네트워크 정보를 평가할 수 있는 방법이 요구 된다.

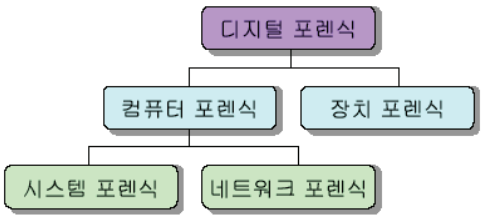
본 논문에서는 퍼지 의사결정 나무를 이용한 디지털 포렌식을 분석하는 네트워크 포렌식의 방법 소개한다. 이 방법은 데이터 마이닝의 한 기법인 의

사 결정나무를 사용 하는 것으로, 수직적 경계 분류 방식을 사용하여 경계점 근처에서의 예측오류의 발생을 줄일 수 있는 방법이다. 2장에서는 네트워크 포렌식의 정의와 기존의 의사 결정 나무를 소개한다. 3장에서는 퍼지이론을 이론적으로 설명한다. 4장에서 실험 및 평가를 하며, 마지막으로 결론으로 구성된다.

2. 관련 연구

2.1 디지털 포렌식의 분류

포렌식은 사전적 의미로 경찰이 범죄를 풀어나가기 위하여 증거를 조사하는 과학자들의 작업을 말한다. 디지털 포렌식은 그림 1 과 같이 분류할 수 있다[1].



(그림 1) 디지털 포렌식의 종류

디지털 포렌식은 컴퓨터 포렌식과 장치 포렌식으로 나뉘며, 컴퓨터 포렌식은 범위에 따라 시스템 포렌식과 네트워크 포렌식으로 나누어진다. 시스템 포렌식과 네트워크 포렌식의 특징은 다음과 같다[3].

1) 시스템 포렌식

일반적인 컴퓨터 포렌식 분야를 나타낸다. 컴퓨터 범죄 순간에 포착된 데이터로부터 증거를 수집하는 일련의 작업으로써, 저장매체 조사, 지워진 파일 복구, 지스러기영역(slack space), 자유영역 검색, 법적 대응을 위한 수집된 자료 보존등을 수행한다.

2) 네트워크 포렌식

네트워크 포렌식은 90년대 초반 Marcus Ranum 의해 소개가 되었다. 네트워크 포렌식은 침해사고 확인을 시발점으로 침해와 관련된 네트워크 이벤트를 수집, 분석, 저장하는 일련의 과정이다. 네트워크 포렌식의 목적은 침해당시의 상황 재구성을 통해 증거자료에 대한 신뢰성을 제공하는데 있다.

고난도 기술이 필요한 분야로 범위가 매우 넓고, 광범위하며 복잡한 네트워크상에 분산되어 있는 디지털 증거를 찾아낸다.즉, 침입탐지 시스템과 같이 모든 정보를 실시간 탐지할 수 있는 시스템이 필요하다. 네트워크 포렌식은 일반적으로 두 가지 형태로 분류 된다[2].

가) Catch-it-as-you-can systems

내부 네트워크로 진입하는 모든 패킷을 저장소에 저장하고 저장소의 정보를 이용하여 포렌식 분석서버에서 분석을 진행한다. 이 시스템은 많은 저장 공간을 필요로 한다.

나) Stop, look and Listen Systems

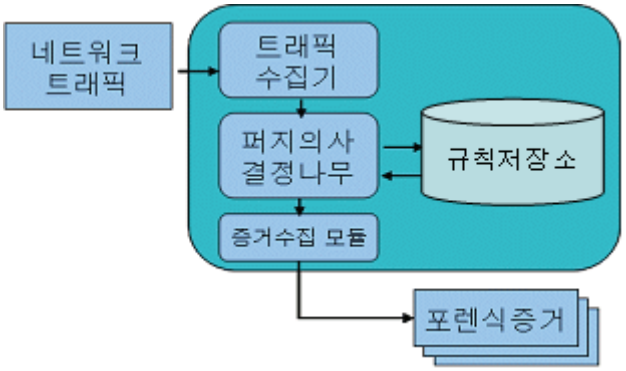
내부 네트워크로 진입하는 패킷을 메모리로 가져와 먼저 분석을 진행하고 침입과 관련된 정보들만 저장소에 저장하는 방법이다. 적은 저장 공간을 필요로 하나 실시간 패킷 분석을 위해 빠른 처리가 필요하다.

2.2 의사결정 나무

의사결정 규칙을 나무 구조로 도표화하여 관심있는 대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 의사결정 나무는 노드와 가지로 구성되어있는데, 각각의 루트노드에서 터미널 노드까지의 경로는 규칙으로 해석이 가능하며, 데이터 마이닝에 유용하다.

Id3와 C4.5는 Quinlan 에 의해 제안된 의사결정 나무로써 가장 널리 사용되고, 이들은 정보 획득을 근거로 한 엔트로피를 이용하여 트리를 구성한다[1]. 의사결정 나무는 연속형 변수를 비연속적인 값으로 취급한다. 여기서 분리 경계를 나눌 때, 수직적 경계 분류 방식을 사용하여 경계점 근처에서의 예측오류를 발생한다.

3. 퍼지의사 결정 나무를 이용한 네트워크 포렌식 증거 추출



(그림 2) 퍼지 의사결정 나무 네트워크 포렌식

제안된 퍼지 의사 결정 나무 네트워크 포렌식 시스템은 그림 2 와 같이 트래픽 수집기, 퍼지의사 결정나무, 증거수집모듈, 규칙 저장소로 이루어져 있다. 각 모듈의 동작은 디지털 포렌식의 절차와 견주어 볼 수 있으며, 시스템의 전체적인 동작과정은 다음과 같다. 먼저, 트래픽 수집기는 네트워크 패킷들을 수집하여 의사 결정 나무 모듈로 보낸다. 다음으로, 의사 결정 나무 모듈은 트래픽 수집기가 보내온 패킷에 대하여 룰 데이터베이스의 규칙에 따라 사건을 식별한다. 마지막으로, 증거수집 모듈은 사건식별 모듈이 식별한 사건에 대한 필요한 증거를 수집하여 증거를 생성한다. 각 모듈의 세부동작은 다음과 같다.

3.1 트래픽 수집기

트래픽 수집기 모듈은 디지털 포렌식의 절차 중 수집단계에 해당한다. 네트워크 포렌식 시스템이 설치되어있는 네트워크의 트래픽을 수집하기 위하여 트래픽 수집기는 pcap 라이브러리를 이용하여 구현되었다. pcap 라이브러리는 패킷 수집을 위한 고 수준 인터페이스를 제공하여 주는 라이브러리로 네트워크 인터페이스 카드를 무차별모드(promiscuous mode)로 설정하여 해당 네트워크의 트래픽을 수집할 수 있다.

3.2 퍼지 의사 결정 모듈

퍼지 의사 결정 모듈에서는 트래픽 수집기로부터 받은 패킷정보를 사건규칙 룰베이스의 정보와 비교하여 사건을 판단한다. 퍼지 의사 결정 모듈의 기능은 디지털 포렌식의 절차 중 식별단계와 조사단계에 해당한다. 특히 이 모듈에서는 네트워크 포렌식에서의 성능 고려사항을 반영하여 증거에 요구된 외의 데이터를 추려내는 기능을 수행하게 된다.

3.3 규칙 저장소

규칙 저장소는 사건식별 모듈에서 사건 즉, 공격을 판단하는데 사용된다. 표 2 는 실제 학습을 통해 생성된 규칙 예이다.

<표 2> 규칙의 예

```
case xl;
  switch num_shells;
  case zero;
    switch root_shell;
    case on;
      switch d_h_s_s_p_r;
      case xs;
        class guess_passwd.;
      case s;
        class buffer_over.;
      case m;
        class buffer_over.;
```

3.4 증거수집 모듈

증거수집 모듈의 기능은 디지털 포렌식의 절차 중에서 보존단계와 수집단계에 해당한다. 증거수집 모듈에서는 사건식별 모듈로부터 사건정보를 받아 증거를 수집하게 된다. 증거는 사건식별 모듈에서 결정된 사건유형을 반영하여 형식이 결정된다. 제안된 시스템에서는 공격유형을 Probe, U2L(User toRoot), R2L(Remote to Local), DoS(Denial of Service)의 네 가지로 분류하여 증거형식을 결정한다[8].

4. 실험 및 평가

4.1 실험 데이터

제안된 네트워크 포렌식 탐지를 평가하기 위하여, 1998년 DARPA 데이터의 섹션을 변형시킨 KDD-cup 99 데이터 셋을 사용하였다. DARPA 데이터는 MIT Lincoln Labs에서 공격탐지를 평가하기 위해서 만든 것이다. 이 데이터는 41개의 독립적인 필드와 1개의 라벨로 구성되어 있다. 각각의 독립적인 필드는 9개의 기호형 속성을 가지고, 32개의 숫자형 속성을 가진다. 각각의 공격은 4가지의 범주로 나누어지고 37개의 공격으로 구성되어 있다[4].

표 3 은 실제 사용된 데이터의 실제 값이다.

<표 3> KDD-cup 99 데이터 실제 값

```
,udp,private,SF,105,146…...1.00,0.00,0.00,255,254,1.......0.
00,normal.
0,udp,private,SF,105,1461….
0.00,255,254,1.....0,0.00,snmpgetattack
0,icmp,ecr_i,SF,1032,0,0,504,504,1…..,0.00,255,255,...
1,0.00,smurf.
4,tcp,pop_3,SF,30,93,0,0,1,1,,0.00,255,218,0.85...,0.00,gue
ss_passwd.
```

4.2 실험 방법

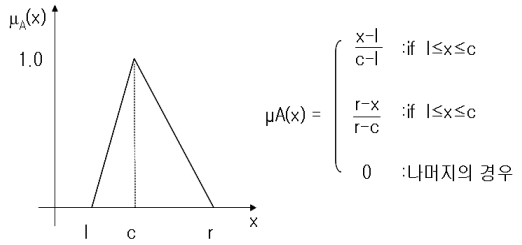
전체 데이터의 10%를 이용하여 학습을 시키고, 다시 전체 데이터를 이용하여 퍼지 의사결정 나무의 성능을 평가 한다.

위와 같은 방법으로 하는 것은 학습 데이터를 가지고 얼마나 일반화를 할 수 있는지 평가를 하는데 도움이 된다.

41개의 독립적인 속성 값을 모두 사용하는 것은 대형 IDS 모델에서 주로 쓰이는 방식으로 높은 탐지율을 나타낼 수 있지만, 실시간 탐지에는 많은 부하가 생긴다. 하지만, 증거의 추출을 목표로 실시간 탐지가 아닌 증거 획득에 목적이 있는 네트워크 포렌식은 전체 41개의 속성 값을 모두 사용 한다.

퍼지 룰을 생성하는데 멤버쉽 함수는 매우 중요하다. 멤버쉽 함수의 정의에 따라 성능과 정확도가 달라질 수 있다. 일반적으로 사용하는 멤버쉽 함수에는 triangular, trapezoidal, gaussian과 같은 3가지 종류가 있다.

이 논문에서는 구현하기 쉬운 triangular를 이용하였다. 이 함수는 미리 정해지지 않은 데에터에 적용하는 일반적인 방법이다. 그림 3에서와 같이3개 인자 (l, c, r) 로 표현이 된다. 각각의 인자는 삼각형의 왼쪽, 중간, 오른쪽의 점을 나타낸다. 32개의 숫자형 속성값은 triangular 멤버쉽 함수를 적용한다.



(그림 3) Triangular 퍼지 멤버쉽 함수

4.3 실험 결과

표 4은 10%데이터에 의해 생성된 규칙을 가지고 전체 데이터에 적용하여 정확도를 측정한 수치이다.

공격의 탐지율(true negative)과 정상을 공격으로 잘못 분류하는 과탐율(False Positive)로 알고리즘ㅇ의 성능을 나타낸다. 시험 결과에서 볼 수 있듯이 99.47%의 높은 탐지율을 나타내었다. 이 결과는 기존의 수직적 분리 방법에서 퍼지 개념을 도입하여 경계값을 더 잘 표현해준 결과이다.

<표 4> 퍼지 의사결정 나무
실험결과

False Positive	True Negative
0.24	99.47

표 5는 학습을 통하여 생성된 퍼지 의사결정 나무 규칙 이다. 규칙의 구성은 Switch case 의 구조를 가지고 있다.

5. 결론

본 논문에서는 네트워크 포렌식에 대한 증거 수집의 자동화 방향을 제시하였다. 해석이 쉬운 의사결정 나무를 선택하였고, 기존 의사결정 나무의 단점인 분리 경계에서의 오차 값을 보완하기 위하여

<표 5> 학습을 통해 생성된 룰

```
      case xs;
        class normal.;
      case s;
        class normal.;
      case m;
        class loadmodule.;
      case xl;
        class normal.;
      end;
    case l;
      switch duration;
      case xs;
        class loadmodule.;
      case s;
        class normal.;
      end;
    case xl;
      switch num_shells;
      case zero;
        switch root_shell;
        case on;
          switch d_h_s_s_p_r;
          case xs;
            class guess_passwd.;
          case s;
            class buffer_over.;
          case m;
            class buffer_over.;
          case xl;
            switch logged_in;
            case on;
              switch num_flogins;
              case zero;
                class loadmodule.;
              case five;
                class guess_passwd.;
              end;
            case off;
              switch n_f_creation;
              case xs;
                class normal.;
              case s;
```

퍼지적 사고를 추가 하여, 실험 결과에서와 같은 높은 탐지율을 보여 주었다. 본 시스템은 많은 양의 네트워크 데이터에서 범죄에 사용된 증거 데이터만을 분류하여 포렌식 증거 획득에 사용한다면 네트워크 분석에 있어 많은 시간과 비용을 절감할 수 있다.

참고문헌

[1] Quinlan, J. R, "C4.5 : Programs for Machine Learning" Morgan Kaufmann Pulbishers, 1993

[2] 한민옥, 국가 사이버테러 대응체계 가동"디지털타임스", 16, December, 2003

[3] 황현욱, "컴퓨터 포렌식스: 시스템 포렌식스 동향과 기술", 정보보호 학회지, 제 13권

[4] KDD data set, 1999; http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[5] 경찰청 사이버 테러대응센터; http://www.netan.go.kr/

[6] Zadeh, L.A., "Fuzzy sets, Information and Control", 8, 3338-353, 1965.

[7] Marcus Ranum, Network Flight Recorder http://www.ranum.com/

[8] 1999 DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/docs/docs_index.html