Journal of The Korea Institute of Information Security & Cryptology(JKIISC) ISSN 1598-3986(Print)
VOL.23, NO.3, June 2013 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2013.23.3.501

속성기반 악성코드 유사도 분류 문제점 개선을 위한 가중치 분석 연구

정 용 욱,† 노 봉 남‡
전남대학교 산학협동과정

The weight analysis research in developing a similarity classification problem of malicious code based on attributes

Yong-Wook Chung,† Bong-Nam Noh‡
Chunnam National University

요 약

악성코드를 효과적으로 분류 및 대응하기 위해서 유사도 비교를 통한 그룹화 과정이 요구된다. 기존 유사도 비교 방법에서 사용되는 기준 또는 속성만을 이용했을 경우, 미탐 및 오탐이 증가하는 문제점이 발생한다. 그러므로, 본 논문에서는 악성코드 자동분석시스템의 2차적인 휴리스틱 기반 행위분석의 문제점을 보완하기 위해 다양한 속성을 선택하여 사용하고, 속성별 가중치 적용을 위해 AHP(Analytic Hierarchy Process) 의사결정기법을 반영한 유사도 비교 방법을 제안한다. 악성코드의 유사도 비교를 통하여 탐지율과 오탐율의 최적 임계치를 설정하고, 새로운 악성코드에 대한 분류 실험으로 악성코드생성기로 생성된 그룹을 결정함을 보이므로 향후 해킹 유형 및 악성코드 근원지를 추적 할 수 있는 악성코드 그룹 정보로서 활용할 수 있기를 기대한다.

ABSTRACT

A grouping process through the similarity comparison is required to effectively classify and respond a malicious code. When we have a use of the past similarity criteria to be used in the comparison method or properties it happens a increased problem of false negatives and false positives. Therefore, in this paper we apply to choose variety of properties to complement the problem of behavior analysis on the heuristic-based of 2nd step in malicious code auto analysis system , and we suggest a similarity comparison method applying AHP (analytic hierarchy process) for properties weights that reflect the decision-making technique. Through the similarity comparison of malicious code, configured threshold is set to the optimum point between detection rates and false positives rates. As a grouping experiment about unknown malicious it distinguishes each group made by malicious code generator. We expect to apply it as the malicious group information which includes a tracing of hacking types and the origin of malicious codes in the future.

Keywords: Similarity comparison, Malicious code generator, The weight analysis

I. 서 론

오늘날 안티바이러스 업체에서는 수집된 악성코드를 자동분석 시스템을 통하여 전반적인 흐름과 특이점을 파악하고 새로운 악성코드에 대해서 역공학(reverse engineering)기법으로 수동 분석을 하고 있다. 악성코드 분석에 투입할 수 있는 기술 인력은 악성코드 증가율에 비해 매우 부족한 반면, 악성코드를 이용한 디도스(DDoS: Distributed Denial of

접수일(2013년 3월 28일),게재확정일(2013년 4월 15일)
† 주저자, kit1989@daum.net
‡ 교신저자, bbong@jnu.ac.kr(Corresponding author)

Service) 및 취약점 공격은 전 세계적으로 월평균 수백만 건씩 발생하고 있다[1].

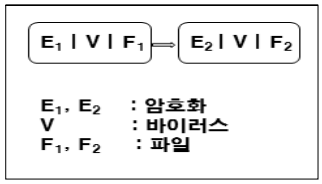
안티바이러스 업체 및 연구소에서 운영하는 악성코드 자동분석 시스템의 구성은 다음과 같다. 모니터링 및 신고를 통해 수집된 악성코드는 1차적인 악성코드 자동분석시스템에서 악성코드 시그니처 기반으로 자동분석 된다. 분석 후에는 2차적인 휴리스틱(heuristic) 기반 기술의 행위 분석을 통해 변종 악성코드를 탐지하게 된다. 악성코드 시그니처 방식의 정적분석[2]과 휴리스틱 방식의 행위분석을 거친 후에는 3차적인 모니터링을 통한 동적분석을 통하여 최종적으로 탐지 되지 않은 악성코드를 탐지한다. 추가적인 정보가 필요한 경우에는 전문 분석가에 의한 수동분석을 하는 시스템으로 구성 되어있다. 고전적 바이러스(classic virus)는 1차적 악성코드 시그니처 방식으로 탐지율이 높으나 변종 악성코드에 대한 탐지율은 낮다. 이러한 변종 악성코드는 2차적 휴리스틱 방식의 행위분석에서 탐지하게 된다. 그러나, 휴리스틱 기술은 이미 알려진 악성코드의 행위를 규칙으로 정의하고 동일한 행위에 대해서 탐지하는데 오탐율이 높은 단점을 가지고 있다. 알려지지 않은 새로운 악성코드는 3차적 모니터링 기반 동적분석에서 탐지 되어야 한다. 1차 및 2차 단계에서 탐지 되지 않는 새로운 악성코드가 증가 할 경우에는 3차적 모니터링 기반의 동적분석이 어려워지는 한계에 도달하게 된다. 또한 탐지 되지 않은 새로운 악성코드에 대한 수동분석 시간이 많이 소요되는 문제점이 발생하게 된다. 하지만, 안티바이러스 및 기존 연구에서 사용하는 악성코드 시그니처[2], 공통 함수, 제어흐름도 등으로 악성코드를 탐지하여 그룹화 하는 방법에는 오탐율 및 미탐율을 증가시키는 한계가 있다.

본 논문에서는 암호화 난독화에 대한 정적분석 문제점과 가상환경과 디버깅 분석도구에 대한 동적분석 문제점 등을 속성으로 포함시켜 유사도 비교를 하게된다. 또한 악성코드 자동분석 시스템에서 휴리스틱 기반 행위분석의 단점을 보완하고자 속성 가중치를 사용한 임계치 방식의 유사도 비교 방법을 제안하고자 한다.

II. 악성코드 유사도 관련연구

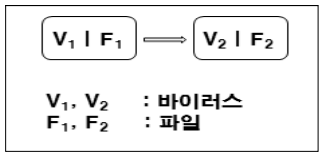
2.1 악성코드 유형

변종 악성코드는 유형에 따라 다형성 바이러스와 메타모픽 바이러스로 구분된다. 다형성 바이러스는 상이한 복호화 루틴을 포함하고 있으며, 동일한 함수를 채택하면서 상이한 명령어 순서로 동작하는 속성을 가지고 있다. 암호화 루틴을 생성하는 엔진으로 MtE, TPE, NED, DAME 등으로 암호화 후에는 일정한 바이러스 패턴을 가지게 된다.



[그림 1] 다형성 바이러스 구조

메타모픽 바이러스는 암호화 대신 바이러스 자체가 시스템에 내제된 컴파일러에 의해 매번 생성되는 속성을 가지고 있다. 메모리에 바이러스 코드가 들어 있지 않아 탐지가 어려우며 복호화를 제공하지 않는다.



[그림 2] 메타모픽 바이러스 구조

다형성 바이러스와 메타모픽 바이러스의 판단기준은 복호화 가능성에 있다. 변형 함수를 A, 바이러스 소스코드를 S, 변형된 악성코드를 N 이라고 할 때, 원형 바이러스와 변형 바이러스 사이에는 N=A(S)라는 관계가 성립한다. 다형성 바이러스와 메타모픽 바이러스 모두 동일한 형태로 보일 수 있지만 위의 2가지 형태의 바이러스 차이는 생성된 N에 있다. 예를 들면, 생성된 N으로부터 원본 S로 돌아갈 수 있는 역함수 $A(S)^{-1}$이 존재한다면 이 변형함수 N은 다형성 바이러스이며, 역함수 $A(S)^{-1}$이 존재하지 않으면 메타모픽 바이러스로 판단하게 된다. 즉, 변형된 바이러스에서 원형 바이러스로 추출이 가능하면 다형성 바이러스, 불가능하면 메타모픽 바이러스로 결정된다.

2.2 악성코드 유사도 분석 연구

악성코드 탐지 및 분류 기술에는 체크섬(check-sum), 문자열 검사, 스마트 검사(빠른 검사), X-Ray 검사(전체 검사), 코드 가상실현, 기하학적

탐지(뜻하지 않게 기하학적인 전파), 휴리스틱(폴더 파일쓰기, 레지스트리 생성 명령어를 시그니처화), 제너릭 (공통된 코드영역) 탐지 등이 있다. 악성코드 탐지 및 분류 관련연구에는 시스템 함수간의 코드와 데이터 흐름을 비교, 변종 악성코드에 전형적으로 사용된 난독화를 제거하는 코드를 정규화[10], 내부 절차 제어 흐름도가 일치하는 서브그래프로 공통 변형 악성코드를 분류, 공통된 난독화(정크삽입, 코드녹화, 패킹)를 복구, 중복값과 필요없는 명령어를 탐지하여 제거, 중복되는 코드를 식별하는 동일성탐지와 컴파일러 최적화 등이 있다. [표 1]은 악성코드 탐지 및 분류 연구 중에서 유사도를 이용한 연구들을 보여주며, 악성코드의 다양한 특징을 추출하여 정적·동적분석 방법으로 유사도 비교를 하고 있다. 정적분석 방법은 악성코드의 사전 행위를 예측할 수 있으며[4,6,7]. 동적분석 방법은 행위결과를 확인할 수 있다[8,9]. 정적분석에서 예측된 행위를 동적분석을 통하여 결과를 확인할 수 있는 관련 연구도 있다[5]. 악성코드 속성(원격 동적라이브러리, 제어 및 명령, 국가코드, 컴파일러 환경, Manifest 정보, 디버깅 정보, 안티디버그, 타임라인, 암호화/난독화, 파일경로, ASCII 문자)을 추출하여 변종 악성코드에 대한 유사도 비교를 하는 연구도 있다[10]. 기존 유사도 관련 연구들은 악성코드 전체를 대상으로 하거나, 악성코드의 일부 속성으로 한정되어 있어 미탐율이 높다. 다양한 11가지 속성들에 대해서도 연구하였으나 유사도 비교에 대한 탐지율이 낮았다. 또한, 디도스 및 취약점 공격에 사용되는 악성코드 등의 속성추출에 대한 연구가 부족하여 이에 대한 지속적인 연구가 필요하다. 본 장에서 악성코드 유사도 관련 연구 및 기존 유사도 연구 문제점들에 대해 살펴보았다. 이를 통해 악성코드의 유사도 탐지에 대한 연구는 계속 발전하였으나 악성코드의 유사도 분류 속성을 사용한 연구는 많이 진행되지 않았음을 파악 하였다.

[표 1] 유사도 관련연구

연구 주체	분석기준	분석 방법	내 용
한양 대학교 컴퓨터 공학과	API 화이트 리스트	정적	응용 프로그램의 주요 API 함수를 화이트리스트로 작성, 악성코드와 공통으로 포함하는 API 함수를 제외한 후 나머지 함수의 순차적 특징을 이용함. 정성적 범위의 오탐율 발생함[4]
고려 대학교 정보 경영 공학과	악성 시그니처	정적 동적	취약한 API 함수 분류기준과 압축 코드에 대한 동적분석으로 악성 시그니처를 추출하여 변종 악성코드 탐지 실험을 통해 증명. 오탐율이 높고 비교샘플에 한정된 결과를 사용함[5]
Black Hat 그룹	악성코드 키워드	정적	Vilo 검색방법으로 n-grams, n-perms을 사용하여 악성코드 키워드를 임계치 필터로 탐지하며 코사인 함수로 유사도 검사. 오탐율과 미탐율로 평가함[6]
Univer sityof Bonn	플래그 값과 상수값 일치도	정적	악성코드에 포함된 플래그 값과 상수값은 동일함을 사용하여 새로운 변종 악성코드를 탐지하고 공통그룹으로 분류함[7]
Micros oft Acade mic Search	계층도	동적	가상환경에서 악성코드 공통함수에 대한 유사도를 계산, 가상환경을 우회하는 악성코드의 유사도 비교가 어려움[8]
Univer sity of Mann heim	안티 바이러스 기반 기계학습 기법	동적	가상환경에서 악성코드 행위를 모니터링하고 안티바이러스에 기반한 기계학습기법으로 악성코드를 그룹 분류. 제한된 악성코드로 지속적인 모니터링이 어려움[9]

III. 악성코드 분류를 위한 유사도 분석

유사도 비교는 악성코드의 문자열이 출현하는 빈도순에 따라 정리한 후, 상위 N%의 문자열과 하위 M%의 문자열을 제거하고 남은 문자열로 전체적인 내용을 상호비교 하는 방식과 악성코드에서 추출된 문자열을 출현빈도로 정리한 뒤, 상위 N개의 문자열로 특정 내용을 상호비교 하는 방식이 있다[11].

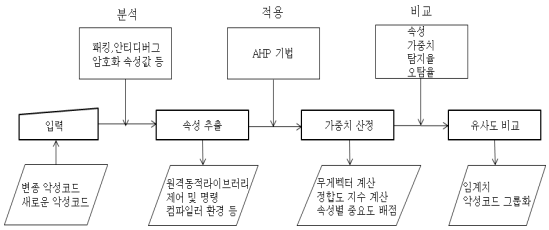
본 논문에서는 후자의 방법을 선택하였으며 속성, 가중치, 자카드 계수, 임계치 등으로 새로운 악성코드가 악성코드 그룹 분류를 통해서 공통 그룹인지 여부를 판단한다. 다형성 바이러스 및 메타모픽 바이러스와 같은 악성코드는 관련정보가 공개되지 않아 엄선된

[표 2] 속성(MP_n)

속성	내용	속성	내용
MP_1	원격동적 라이브러리	MP_7	안티디버그
MP_2	제어명령	MP_8	타임라인
MP_3	국가코드	MP_9	암호화 난독화
MP_4	컴파일러 환경	MP_{10}	파일경로
MP_5	Manifest 정보	MP_{11}	ASCII 문자열
MP_6	디버깅 정보		

악성코드를 분류하기 위해서 관련 연구[12]에서 제안했던 [표 2]의 11가지 속성을 기반으로 가중치를 산정한다.

악성코드 분류를 위한 유사도 분석을 위해 구성도는 [그림 4]와 같으며, 악성코드 입력, 속성 산출, 가중치 산정, 유사도 비교에 대한 전체적 유사도 분석 등으로 구성된다.



[그림 3] 유사도 분석 구성도

3.1 가중치 산정

악성코드 자동분석시스템 결과와 전문 분석가의 수동 분석결과는 미탐(false negative) 및 오탐(false positive)을 최소한으로 줄이기 위해서 병행되고 있다. 독자적 성향을 띤 악성코드는 전문 분석가의 주관적이지만 능률적인 결과에 의존하게 된다. 본 논문에서는 대표적인 주관적 방식이며 능률적인 결과를 가져올 수 있는 AHP 기법으로 다양한 의사결정을 통한 속성에 대한 최적의 가중치를 선정한다. AHP 기법의 이용으로 악성코드에서 측정이 어려운 주관적인 요소와 안티바이러스 결과로 측정이 용이한 객관적 요소들을 의사 결정과정에서 효과적으로 융합시키게 된다. 다음은 악성코드 속성의 가중치에 대한 의사결정을 위하여 아래의 4단계로 진행한다. 단계 1에서는 쌍대비교를 정의하고, 단계 2에서 설문자료를 수집한다. 단계 3에서 유효성 평가를 하여 단계 4에서는 속성에 대한 가중치를 부여하게 된다.

4개 기준범위에 대한 쌍대비교에서 객관성을 위하여 그룹에 의한 AHP 의사결정 방법을 사용한다. 4개 기준의 상대적인 중요도 비교를 위하여 쌍대비교행렬을 만들게 된다. m이 가로축 4개 기준범위이며 n은 세로축 4개 기준범위 일 때, 쌍대비교행렬은 $a_{mn} = \frac{1}{a_{mn}}$이 되도록 한다. 상호 비교되는 기준들의 중요도는 1부터 2까지의 수치로 평가된다. 반대로 m보다 n이 더 중요한 경우에는 역수를 사용한다. 「m행의 기준은 n열의 기준보다 중요하다」이면 2점을 부여하고, 「m행의 기준이 n열의 기준과 같은 수준으로 동일하다」이면 1점을 부여한다. 「m행의 기준이 n열의 기준보다 중요하지 않다」이면 0.5로 부여하기로 할 때, *Pseudo* 코드는 다음과 같다.

AHP를 적용하여 유사도 속성에서 정한 4개 기준범위에 대한 11개 속성의 배점을 제안한다. 4개 기준범위를 11개 속성으로 세분화함으로써 악성코드 유사

【단계 1】 쌍대비교 정의
4개의 기준범위인 네트워크, 개발정보, 포렌식, 이진파일로 나누어 AHP 레벨 1에 대한 쌍대비교를 하게 된다. AHP 레벨 2에서는 네트워크 기준에 원격동적 라이브러리, 제어 및 명령에 대한 쌍대비교를 하며, 개발정보 기준에 국가코드, 컴파일러 환경, Manifest 정보, 디버깅 정보에 대한 쌍대비교를 하며, 포렌식 기준에 안티 디버그, 타임라인, 암호화 · 난독화에 대한 쌍대비교를 하며, 파일경로 기준에 대한 ASCII 문자열에 대한 쌍대비교를 하도록 구성한다.

【단계 2】 설문자료 수집
침해사고 대응 분석가, 안티바이러스 업체 악성코드 분석가, 국가기관 악성코드 전문 연구원 등에게 설문조사를 실시해 의사결정 요소들 간의 쌍대비교로 판단할 수 있는 설문자료를 수집한다.

【단계 3】 유효성 평가
고유치(eigen-value)방법을 사용하여, 의사결정요소의 상대적 가중치를 구하고 정합성 평가인 C.I.(Consistency Index)가 0.1 이하인지 유효성 여부를 평가한다.

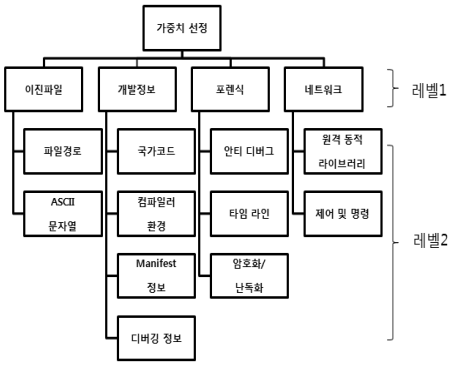
【단계 4】 가중치 부여
기준 범위 및 유사도 속성들에 대한 종합순위를 얻기 위하여 의사결정 요소들의 AHP 레벨 1과 레벨 2의 상대적인 가중치를 종합화 한다. 속성에 대하여 각각의 정성적인 배점기준을 중요도 순으로 정하고 이에 따른 점수를 부여한다. 각 속성별 주어진 점수를 만점으로 총 합계가 100점 기준으로 점수를 부여한다[3]. 악성코드생성기의 해쉬값(hash value)과 체크섬(checksum)을 비교하여 동일하면, 유사한 악성코드이며, 100%이상의 유사도를 보여주는 것으로 정의한다. 변종 악성코드 정적분석에서는 위와 같은 가중치에 근..거하여 악성코드 생성기들에 대한 유사성을 비교한다. 유사성 비교를 통하여 동일한 유형별로 분류가능하며, 알려지지 않은 변종 악성코드에 대해서는 신규그룹으로 형성하게 된다.

[그림 4] 속성 가중치에 대한 의사결정 4단계

if m행 is important than n열, then 2,
if m행 is equal to n열, then 1
if m행 is not important than n열, then 0.5

[그림 5] 쌍대비교행렬 *Pseudocode*

도의 정확성을 높이고자 한다. 또한 AHP로 최종 항목 순위를 구하여 유사도 속성의 중요도 순서를 계산 하는 것을 목표로 한다. 지정 기준 배점을 위한 계층구조는 〔그림 6〕과 같다. 레벨 1은 4개 기준범위를 나타내며, 레벨 2는 11개 속성을 나타낸다. 4개 기준범위를 먼저 비교하고, 세부 속성에 대한 비교를 개별적으로 하여 레벨 2에 해당하는 추가적인 세부 기준이 발생하더라도 세부 기준이 속하게 되는 4개 기준의 상대적인 중요도에 따라 세부 기준의 중요도가 결정되도록 한다.



〔그림 6〕 유사도 계층구조

악성코드 기준범위의 쌍대비교에 대한 설문 참여자의 응답 결과를 보정하여 〔표 3〕에 따라 쌍대비교행렬을 구하였다. 행렬의 각 요소 a에 대한 가로축 m, 세로축 n에 대한 a_{mn}를 행렬 A로 표현하면, 다음과 같다.

$$A=\begin{bmatrix} a_{11}\,a_{12}\,a_{13}\,a_{1n} \\ a_{21}\,a_{22}\,a_{23}\,a_{2n} \\ a_{31}\,a_{32}\,a_{33}\,a_{3n} \\ a_{41}\,a_{42}\,a_{43}\,a_{4n} \end{bmatrix}=\begin{bmatrix} 1 & 1.05 & 0.42 & 0.60 \\ 0.95 & 1 & 1.05 & 0.48 \\ 2.36 & 0.95 & 1 & 1.05 \\ 1.67 & 2.10 & 0.95 & 1 \end{bmatrix}$$

3.1.1 무게벡터 계산

단계 1에서 쌍대비교 행렬 열에 대한 합을 구하고 단계 2에서는 쌍대비교 행렬을 열의 합계로 나눈다. 단계 3에서는 행 요소의 수로 나눈 행 요소의 합인 무게벡터를 구한다.

〔표 3〕 4개 기준범위의 쌍대비교 결과

질문	이진파일	개발정보	포렌식	네트워크
이진파일	1	1.05	0.42	0.60
개발정보	0.95	1	1.05	0.48
포렌식	2.36	0.95	1	1.05
네트워크	1.67	2.10	0.95	1

【단계 1】 쌍대비교 행렬 열에 대한 합계

쌍대비교 행렬의 각 열의 요소를 합한다. 〔표 14〕에서 열의 합계는 5.98, 5.1, 3.42, 3.13 이 된다. $S=a_{mn}$ 열의 합으로 계산식은 다음과 같다.

$$S=\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \end{bmatrix}=\begin{bmatrix} a_{11}+a_{21}+a_{31}+a_{41} \\ a_{12}+a_{22}+a_{32}+a_{42} \\ a_{13}+a_{23}+a_{33}+a_{43} \\ a_{14}+a_{24}+a_{34}+a_{44} \end{bmatrix}=\begin{bmatrix} 5.98 \\ 5.1 \\ 3.42 \\ 3.13 \end{bmatrix}$$

【단계 2】 열의 합계로 나눈 쌍대비교 행렬

쌍대비교 행렬의 요소를 열의 합계로 나눈다. $Z=\frac{a_{mn}}{S}$ 일 때, 계산식은 다음과 같다.

$$Z=\begin{bmatrix} \frac{1}{S1} & \frac{1.05}{S1} & \frac{0.42}{S1} & \frac{0.60}{S1} \\ \frac{0.95}{S2} & \frac{1}{S2} & \frac{1.05}{S2} & \frac{0.48}{S2} \\ \frac{2.36}{S3} & \frac{0.95}{S3} & \frac{1}{S3} & \frac{1.05}{S3} \\ \frac{1.67}{S4} & \frac{2.10}{S4} & \frac{0.95}{S4} & \frac{1}{S4} \end{bmatrix}=\begin{bmatrix} 0.167\,0.205\,0.122\,0.191 \\ 0.158\,0.196\,0.307\,0.153 \\ 0.394\,0.186\,0.292\,0.335 \\ 0.279\,0.411\,0.277\,0.319 \end{bmatrix}$$

【단계 3】 행 요소의 수로 나눈 행 요소의 합

단계 2에서 구해진 행렬에서 각 행의 요소의 합을 각 행의 요소의 개수(n)으로 나눈다. $W=\frac{a_{mn}\text{ 행의 합}}{n}$ 일 때, 계산식은 다음과 같다.

$$W=\begin{bmatrix} W1 \\ W2 \\ W3 \\ W4 \end{bmatrix}=\begin{bmatrix} \frac{a_{11}+a_{12}+a_{13}+a_{13}}{n} \\ \frac{a_{21}+a_{22}+a_{23}+a_{24}}{n} \\ \frac{a_{31}+a_{32}+a_{33}+a_{34}}{n} \\ \frac{a_{41}+a_{42}+a_{43}+a_{44}}{n} \end{bmatrix}=\begin{bmatrix} 0.172 \\ 0.204 \\ 0.302 \\ 0.322 \end{bmatrix}$$

본 결과가 무게벡터가 된다. “이진파일”($W1$)의 중요도는 0.172, “개발정보”($W2$)의 중요도는 0.204, “포렌식”($W3$)의 중요도는 0.302, “네트워크”($W4$)의 중요도는 0.322 가 된다.

$$W=\begin{bmatrix} W1 \\ W2 \\ W3 \\ W4 \end{bmatrix}=\begin{bmatrix} 0.172 \\ 0.204 \\ 0.302 \\ 0.322 \end{bmatrix}$$

무게벡터에 따라 4개 기준은 “네트워크” 〉 “포렌식” 〉 “개발정보” 〉 “이진파일” 순으로 중요함을 알 수 있다.

3.1.2 정합도 지수 계산

정합도 지수 계산의 단계 1에서 무게벡터의 요소를 곱한 합을 구하고, 단계 2에서는 단계 1의 합을 요소의 중요도로 나눈다. 단계 3에서 단계 2에 대한 평균값을 구하고, 단계 4에서는 정합도 지수 $C.I.$를 계산한다.

【단계 1】 무게벡터의 요소를 곱한 합 X를 계산

〔표 14〕 쌍대비교행렬의 각 열에 앞에서 구한 무게벡터의 각 요소를 차례로 곱하고 그 합을 구한다. 다음과 같이 계산한다. $X = W \times a_{mn}$ 열의 합 일 때, 계산식은 다음과 같다.

$$X = W1 \times \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \end{bmatrix} + W2 \times \begin{bmatrix} a_{21} \\ a_{22} \\ a_{23} \\ a_{24} \end{bmatrix} + W3 \times \begin{bmatrix} a_{31} \\ a_{32} \\ a_{33} \\ a_{34} \end{bmatrix} + W4 \times \begin{bmatrix} a_{41} \\ a_{42} \\ a_{43} \\ a_{44} \end{bmatrix}$$

$$X = \begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \end{bmatrix} = \begin{bmatrix} 0.706 \\ 0.840 \\ 1.244 \\ 1.321 \end{bmatrix}$$

【단계 2】 X를 요소의 중요도로 나눈 I를 계산

단계 1에서 구한 각 요소의 계산결과를 앞에서 구한 각 요소의 중요도로 나눈다. 계산 결과는 다음과 같다. $I = \frac{X}{W}$ 일 때, 계산식은 다음과 같다.

$$I = \begin{bmatrix} I1 \\ I2 \\ I3 \\ I4 \end{bmatrix} = \begin{bmatrix} X1/W1 \\ X2/W2 \\ X3/W3 \\ X4/W4 \end{bmatrix} = \begin{bmatrix} 0.706/0.172 \\ 0.840/0.204 \\ 1.244/0.302 \\ 1.321/0.322 \end{bmatrix} = \begin{bmatrix} 4.107 \\ 4.120 \\ 4.106 \\ 4.112 \end{bmatrix}$$

【단계 3】 I에 대한 평균 λ_{max}를 계산

단계 2에서 구한 각 요소의 계산결과를 평균한다. 이 결과가 표의 쌍대비교행렬의 최대 고유치 $= \lambda_{max}$ 이다. 계산 결과는 다음과 같다.

$$\lambda_{max} = \frac{I1 + I2 + I3 + I4}{n} = 4.11096$$

【단계 4】 정합도 지수 $C.I.$를 계산

단계 3에서 구한 λ_{max}로부터 정합도 지수 $C.I.$를 계산하면 n = 각 행의요소의 수일 때, 다음과 같다.

$$C.I. = \frac{\lambda_{max} - n}{n-1} = \frac{4.11096 - 4}{4-1} = 0.0369871 < 0.1$$

$C.I.$의 값이 0.1보다 작으므로 간이계산법에 의한 지정 기준의 상대비교는 유효하다.

3.1.3 속성별 중요도 및 배점

4개 기준에 대한 상대적 중요도를 구한 것과 같은 방법으로 4개 기준별 속성 중요도를 구할 수 있다. 첫 번째로 이진파일「기준 1」에서 상대적 중요도의 이진파일에 따른 속성을 AHP 설문지로 실시한 결과는 〔표 4〕와 같다.

〔표 4〕 이진파일「기준 1」에서의 쌍대비교 벡터

질문	파일경로	ASCII 문자열
파일경로	1	0.57
ASCII 문자열	1.74	1

$\lambda_{max} = 2.000000$ $C.I. = 0.0000$

쌍대비교행렬의 최대 고유치는 $\lambda_{max} = 2.000000$ 이다. 이를 정규화한 무게벡터는 $w_1^T = (0.363, 0.637)$ 로서 이진파일에 따른 속성 가중치이다.

두번째로 개발정보「기준 2」에서 상대적 중요도의 개발정보에 따른 속성 쌍대비교를 AHP 설문지로 실시한 결과는 〔표 5〕와 같다.

〔표 5〕 개발정보 「기준 2」에서의 쌍대비교 벡터

질문	국가코드	컴파일러 환경	Manifest 정보	디버깅 정보
국가코드	1	0.82	0.92	0.84
컴파일러 환경	1.22	1	0.79	0.71
Manifest 정보	1.09	1.27	1	0.79
디버깅 정보	1.20	1.40	1.27	1

$\lambda_{max} = 4.018258$ $C.I. = 0.0062$

쌍대비교행렬의 최대 고유치는 $\lambda_{max} = 4.018258$ 이다. 정규화한 무게벡터는 $w_2^T = (0.222, 0, 226, 0.253, 0.299)$ 로서 개발정보에 따른 세부 기준별 가중치이다.

세번째로 포렌식「기준 3」에서 상대적 중요도의

안티디버그에 따른 속성을 AHP 설문지로 실시한 결과는 〔표 6〕과 같다.

〔표 6〕 포렌식 「기준 3」 에서의 쌍대비교 벡터

질문	안티디버그	타임라인	암호화 · 난독화
안티디버그	1	0.86	0.47
타임라인	1.16	1	0.78
암호화 · 난독화	2.13	1.27	1

$\lambda_{\max} = 3.015026 \quad C.I. = 0.0070$

쌍대비교행렬의 최대 고유치는 $\lambda_{\max} = 3.015026$ 이다. 정규화한 무게벡터는 $w_3^T = (0.238, 0.312, 0.451)$로서 포렌식에 따른 속성 가중치이다.

네 번째로 네트워크「기준 4」에서 상대적 중요도의 네트워크에 따른 속성 쌍대비교를 AHP 설문지로 실시한 결과는 〔표 7〕과 같다.

〔표 7〕 네트워크 「기준 4」 에서의 쌍대비교 벡터

질문	원격동적 라이브러리	제어명령
원격동적 라이브러리	1	0.54
제어명령	1.85	1

$\lambda_{\max} = 2.000000 \quad C.I. = 0.0000$

쌍대비교행렬의 최대 고유치는 $\lambda_{\max} = 2.000000$ 이다. 이를 정규화한 무게벡터는 $w_4^T = (0.351, 0.649)$로서 네트워크에 따른 속성 가중치이다.

3.1.4 속성 가중치 환산

속성별로 가중치를 정규화하고, 기준 범위별로 중요도에 따라 가중치를 점수로 환산한다. 레벨 1에서 구한 무게벡터에 따라 이진파일 「기준 1」의 무게를 정규화 하면, 다음과 같다.

$$0.172 \times w_1^T = (0.06, 0.11)$$

레벨 1에서 구한 무게벡터에 따라 개발정보「기준 2」의 무게를 정규화 하면, 다음과 같다.

$$0.204 \times w_2^T = (0.05, 0.05, 0.05, 0.06)$$

레벨 1에서 구한 무게벡터에 따라 포렌식「기준 3」의 무게를 정규화 하면, 다음과 같다.

$$0.302 \times w_3^T = (0.07, 0.09, 0.14)$$

레벨 1에서 구한 무게벡터에 따라 네트워크「기준 4」의 무게를 정규화 하면, 다음과 같다.

$$0.322 \times w_4^T = (0.11, 0.21)$$

기준범위별 중요도에 따른 점수 배정은 4개 기준별 11개 속성 중요도를 100점 기준으로 환산하여 〔표 8〕과 같다.

〔표 8〕 속성에 대한 가중치

속성		가중치
네트워크「기준4」	①원격 동적 라이브러리	11
	②제어 및 명령	21
	합 계	32
개발정보「기준2」	③국가코드	6
	④컴파일러 환경	5
	⑤Manifest 정보	5
	⑥디버깅 정보	5
	합 계	21
포렌식「기준3」	⑦안티 디버그	7
	⑧타임 라인	9
	⑨암호화 · 난독화	14
	합 계	30
이진파일「기준1」	⑩파일경로	6
	⑪ASCII 문자열	11
	합 계	17
총 합계		100

4개 기준범위의 「기준 1」이 17점, 「기준 2」가 21점, 「기준 3」은 30점, 「기준 4」는 32점으로 배점 되었다. 중요성이 가장 높은 기준범위는 네트워크 부분이며 제어 및 명령이 21점으로 높게 배정되었다. 가장 낮은 범위는 이진파일 부분이며 컴파일러 환경, Manifest 정보, 디버깅 정보가 5점으로 낮게 배정되었다.

3.2 악성코드 유사도 비교

자카드 계수(Jaccard coefficient)를 적용하여, 악성코드 속성 가중치에 의한 유사도 비교를 한다. 유사도 비교를 위해서 악성코드생성기에서 생성된 변종 악성코드와 오탐이 높은 웹 하드용 프로그램으로부터 탐지율과 오탐율을 계산하여 임계치를 구하게 된다.

3.2.1 유사도 함수 적용

자카드 계수는 중복된 패턴이 출현하였을 때 공통 개수로서만 유사도를 계산한다. 악성코드 속성의 중요

도에 따라 중복되는 경우에는 가중치를 반영하여 유사도 비교를 함으로써 우수한 결과를 가져오도록 자카드 계수의 일부를 개선하였다. 따라서 공통 속성(*SMP*)은 악성코드 속성에 속성별 가중치를 부여하여 속성의 중요도를 포함 시키는 향상된 결과를 가져 오는 속성을 말한다. 자카드 계수는 0~1사이의 최대값과 최소값을 사용하며 패턴의 출현여부로 유사도를 나타낸다. 이러한 유사도 계수에 공통 속성값을 가진 악성코드 속성의 가중치를 부여하여 계산하면, 공통 속성에 대한 빈도 문제점을 보완 하게 된다. 악성코드의 탐지율과 오탐율에 대한 유사도 비교를 통해 임계치를 구하고, 새로운 악성코드에 대해 유사도 비교 실험을 한다. 이때 새로운 악성코드가 공통 악성코드 그룹으로 정상 분류되고 있는지를 확인함으로서 유사도의 정확성을 검증하게 된다.

〔표 9〕 속성(MP_n)

코드	속성	코드	속성
MP_1	원격동적 라이브러리	MP_9	암호화 난독화
MP_2	제어명령	MP_{10}	파일경로
MP_3	국가코드	MP_{11}	ASCII 문자열
MP_4	컴파일러 환경	*MDI*	훈련데이터 I
MP_5	Manifest 정보	*MDII*	훈련데이터 II
MP_6	디버깅 정보	*TD*	실험데이터
MP_7	안티디버그	*AD*	정상데이터
MP_8	타임라인	*SMP*	공통속성

〔표 10〕 가중치(*MW*)

속성	MP_1	MP_2	MP_3	MP_4	MP_5	MP_6	MP_7	MP_8	MP_9	MP_{10}	MP_{11}
가중치	11	21	6	5	5	5	7	9	14	6	11

MDI_{MPm}을 훈련데이터 I 의 속성, $MDII_{MPn}$을 훈련데이터 II의 속성이라 할 때,

$$MDI_{MP_m} = \{MP_1, MP_2, MP_3, MP_4, MP_5, MP_6, MP_7, MP_8, MP_9, MP_{10}, MP_{11}\}$$
$$MDII_{MP_n} = \{MP_1, MP_2, MP_3, MP_4, MP_5, MP_6, MP_7, MP_8, MP_9, MP_{10}, MP_{11}\}$$

이 되며,

자카드 계수를 사용한 유사도 함수공식S_J(Similarity Jaccard)은 다음과 같이 정의한다.

$$S_J = \frac{N(MDI_{MPm} \cap MDII_{MPn})}{N(MDI_{MPm}) + N(MDII_{MPn}) - N(MDI_{MPm} \cap MDII_{MPn})}$$

다음은 유사도 함수를 통해 속성과 가중치(*MW*)를 적용한 예를 설명한다. *MDI* 그룹내 임의의 악성코드 MDI_{MPm} 속성이 *MP1, MP2, MP4, MP8, MP10, MP11* 일 때 전체 개수는 6개, *MDII* 그룹내 임의의 악성코드 $MDII_{MPn}$ 속성이 *MP1, MP2, MP4, MP6* 일 때 전체 개수는 4개이면, $MDI_m \cap MDII_n$의 공통 속성은 *MP1, MP2, MP4*로 3개가 된다. 개수를 기반으로 한 유사도 계산을 하게 되면 속성값의 동일함을 확인하지 않고 $\frac{3}{6+4-3} = 0.4285$ 의 유사도 값을 가진다. 그러나 가중치를 기반으로 한 유사도 계산을 하게 되면 공통으로 *MP1, MP2, MP4* 에 해당하는 속성값이 일치한다고 가정할 때,

$$T = \frac{N(MDI_{MPm} \cap MDII_{MPn}) \times \sum(MW)}{N(MDI_{MPm} \times \sum MW) + N(MDII_{MPn} \times \sum MW) - N(MDI_{MPm} \cap MDII_{MPn}) \times \sum MW}$$

$$= \frac{(11+21+5)}{(11+21+4+9+6+11)+(11+21+5+5)-(11+21+5)} = 0.5522$$

0.5522 유사도 값은 0.4285 유사도 값보다 높게 나왔다. 이는 공통 속성값의 가중치를 가지고 자카드 함수에 적용하면 변종 악성코드의 유사도를 향상 시킬 수 있음을 알 수 있다. 유사도 함수는 주체에 따라 사용자 기반과 항목 기반으로 나누어져 있다. 사용자 기반의 유사도는 데이터가 증가 할수록 중복 확률이 작아져 유사도 결정이 어려워 추가 작업이 필요 없는 메모리 기반 데이터 셋에 적용된다. 항목 기반의 유사도는 평가 항목을 미리 계산하고, 유사도 비교 시, 사용자 평가 항목을 보고 유사한 항목들의 가중치를 계산한다. 다음은 항목 기반의 유사도를 가지고 자카드 계수를 사용하기 위해 아래와 같이 3단계로 진행한다.

【단계 1】 비교 실험 데이터 정의

악성코드 유사도를 비교 실험 하는데 있어 훈련데이터I(*MDI*), 훈련데이터II(*MDII*), 정상데이터(*AD*), 실험데이터(*TD*)에 대한 정의가 필요하다. 훈련데이터와 실험 데이터는 TELUS Security Laboratory에서 정의한 악성코드생성기로 생성한 다형성/메타모픽 바이러스 형태의 변종 악성코드 등으로 구성한다. 정상데이터는 안티바이러스에서 오탐 할 수 있는 웹하드 업체의 전용 다운로더 프로그램과 아래아한글의 업데이트 프로그램을 포함한다.

1. 악성코드는 버전별로 동작하기 때문에 가상환경에서 윈도우 XP 표준 버전으로 실험을 한다.
2. 동일한 악성코드생성기에서 생성된 훈련데이터들은 일부 속성을 동일하게 가지기 때문에 해당 악성코드 분류 그룹과 근접한 유사도를 가진다.
3. 타임라인 속성을 가진 악성코드에서 6개월 이내의 값은 동일한 속성값으로 본다.
4. 실험데이터는 훈련데이터에 없는 새로운 속성을 포함한다.
5. 악성코드의 유사도 비교는 공통 속성을 가지면서 동일한 속성값을 가질 때 성립 된다.
6. 서로 다른 임의의 두 정수 n〉0, r〉0에 대하여 n개의 훈련데이터 그룹에서 순서를 생각하지 않고 r개의 실험데이터를 선택한다. 이때 이루어진 집합을 n개에서 r개 선택하는 조합으로 구성하고 ${}_nC_r$ 로 나타낸다.

$$ {}_nC_r = \frac{n!}{r!(n-r)!} \text{(단, } 0 \le r \le n) $$

〔그림 7〕 제약조건

【단계 2】 탐지율, 오탐율, 임계치 계산

다음은 자카드 계수에서 중복되는 개수 대신, 속성별 가중치 값을 사용하여 훈련데이터I, 훈련데이터II의 유사도 비교를 통해서 탐지율을 구하고, 정상데이터와 훈련데이터I, 훈련데이터II의 유사도 비교를 통해서 오탐율을 구한다. 탐지율, 오탐율에 대한 결과로 유사도의 정확성을 증명하기 위하여 ROC (Receiver Operating Characteristic) 그래프를 통한 절충지점의 임계치로 한다.

【단계 3】 실험데이터 분류

실험 데이터는 임계치에 의한 새로운 악성코드의 유사도 비교 시, 사용된다. 임계치에 따라 일정기준을 초과하면 악성코드로 보며, 임계치 기준을 미달하면 새로운 악성코드, 정상 프로그램으로 분류한다.

본 실험에서는 악성코드의 속성과 가중치에 대한 유사도의 정확성을 확인하기 위해 유사도로 가중치 분포에 따른 백분율을 구한다. 속성 점수를 바탕으로 실험할 때, 비교대상 및 가중치 방법에 대한 제약조건은 다음과 같다.

3.2.2 악성코드 그룹화

악성코드 유형별로 분류한 데이터를 가지고 탐지율과 탐지율을 구하고 ROC 그래프를 통해 임계치를 선택하여 악성코드 분류 그룹을 설계한다. 새로운 악성코드와 악성코드 분류 그룹간의 유사도를 비교할 수 있는 분석환경을 제안한다. 임계치 산정을 위해 TELUS Security Lab에서 분류한 악성코드 생성기 28개에서 생성한 다형성 / 메타모픽 바이러스를 바탕으로 훈련데이터I(*MDI*) ,훈련데이터II(*MDII*)를 포함한 전체 훈련데이터 (*MD*) 56개와, 국내 웹 하드 업체에서 사용하는 정상데이터(*AD*) 32개로 구성한다. 그룹 분류 시스템에서는 훈련데이터 I 28개, 훈련데이터 II 28개를 통해 악성코드가 그룹별로 분류되는지에 대한 비율인 오탐율을 얻게 된다. 또한 전체 훈련데이터 56개, 정상데이터 32개를 통해 정상데이터가 악성코드로 오탐되는 비율인 오탐율을 구한다. 탐지율과 오탐율의 최적지점을 임계치로 설정하여 유사도 비교의 검색 효율성을 높이고자 한다. 악성코드 그룹화에는 수십만 개의 알려진 악성코드를 포함하여 새로운 악성코드를 비교하는데 시간이 소요될 수 있다. 이러한 시간을 줄이기 위해서 악성코드 그룹화에 대한 임계치 비교를 통하여 유사도가 가장 높은 임계치내의 악성코드만을 선택하여 유사도 비교를 하게 된다.

탐지율 계산을 위해서 훈련데이터의 악성코드를 그룹 분류 시스템으로 분류한 후, 자카드 계수를 사용하여 훈련데이터들과 비교한다. 전체 훈련데이터(*MD*)는 총56개로 구성되며 서로 다른 임의의 두 정수 *i〉0, j〉0*에 대하여 56개의 악성코드에서 순서를 생각하지 않고 2개를 선택하여 이루어진 각각의 집합을 56개에서 2개 선택하는 조합으로 구성하고 ${}_{56}C_2$ 로 나타내면 다음과 같다.

$$ {}_{56}C_2 = \frac{56!}{2! \times 54!} = 1540 $$

〔표 10〕에서 보인 가중치에 따르면, *MDI*1 의 속성이 *MP8, MP11* 일 때 유사도 값은 $N(MDI_{MP_n}) \times \sum MW = 20$이 된다. *MDII*1 의 속성이 *MP3, MP8, MP11* 일 때 유사도 값은 $N(MDII_{MP_n}) \times \sum MW = 26$이며, $N(MDI_{MP_n} \cap MDII_{MP_n})$의 속성이 *MP8, MP11*에서 일 때 공통 유사도 값은 $N(MDI_{MP_n} \cap MDII_{MP_n}) \times \sum MW = 20$이 된다. 가중치를 기반으로 한 유사도 계산을 하게 되면 속성값이 동일할 때, 0.7692의 유사도 값을 가진다. 단 *MP8, MP11*에 해당하는 속성값이 일치한다고 가정한다.

$$ \frac{(9+11)}{(9+11)+(6+9+11)-(9+11)} = 0.7692 $$

나머지 1540개의 유사도 값들에서 제일 큰 유사도 값을 가진 훈련데이터들에 대한 가장 근접한 유사도 값은 〔표 11〕과 같다.

〔표 11〕 매핑된 전체 훈련데이터 유사도

순서	훈련데이터I	훈련데이터II	유사도 값
1	*MDI1*	*MDII1*	0.7692
2	*MDI2*	*MDII2*	0.8302
3	*MDI3*	*MDII3*	0.2895
….	…..	…..	…..
26	*MDI26*	*MDII26*	0.8269
27	*MDI27*	*MDII27*	0
28	*MDI28*	*MDII28*	0.8723

〔표 12〕에서 변종 악성코드의 유사도를 0.05 임계치 간격으로 총 20구간으로 나누고 있다. 탐지율에 대한 임계치가 0일 때 가장 높은 유사도를 가지고 있으며 임계치가 1일 때, 가장 낮은 유사도를 가지고 있다. 제일 낮은 임계치를 가질 때 유사도 개수가 많아지며 악성코드의 탐지율이 높다.

〔표 12〕 임계치에 대한 훈련데이터 탐지율

임 계 치	유사도 개수	탐 지 율
1	2	3.5714
0.91 ~ 0.95	2	3.5714
0.86 ~ 0.9	2	3.5714
0.81 ~ 0.85	12	21.4285
0.76 ~ 0.8	22	39.286
0.71 ~ 0.75	28	50
0.66 ~ 0.7	30	53.5714
0.61 ~ 0.65	32	57.1428
0.56 ~ 0.6	34	60.7142
0.51 ~ 0.55	34	60.7143
0.46 ~ 0.50	38	67.8571
0.41 ~ 0.45	38	67.8571
0.36 ~ 0.40	40	71.4285
0.31 ~ 0.35	40	71.4285
0.26 ~ 0.30	42	75
0.21 ~ 0.25	50	89.2857
0.16 ~ 0.20	50	89.2857
0.11 ~ 0.15	50	89.2857
0.06 ~ 0.10	50	89.2857
0.01 ~ 0.05	50	89.2857
0	56	100

다음은 오탐율 계산을 위해서 훈련데이터를 악성코드 그룹으로 형성하고, 자카드 계수를 사용하여 정상데이터와 비교하여 유사도를 구한다. 전체 훈련데이터(*MD*)는 56개로 구성하며 정상데이터는 안티바이러스에서 악성코드로 오탐 될 수 있는 웹하드 다운로더와 업데이트 프로그램을 포함한 32개로 구성한다. 서로 다른 임의의 두 정수 i〉0, j〉0에 대하여 총 88개의 프로그램에서 순서를 생각하지 않고 2개를 선택하여 이루어진 각각의 집합을 88개에서 2개 선택하는 조합으로 구성하고 ${}_{88}C_2$ 로 나타내면 다음과 같다.

$$ {}_{88}C_2 = \frac{88!}{2! \times 86!} = 3828 $$

〔표 10〕에서 보인 가중치에 따르면, *MDI*1의 속성이 *MP1,MP2,MP11* 일 때 유사도 값은 $N(MD_{MP_n}) \times \sum MW = 43$이 되며, *AD*1 의 속성이 *MP1,MP2,MP3,MP4,MP5,MP6,MP7,MP8, MP10,MP11*일 때 전체 유사도 점수는 $N(AD_{MP_n}) \times \sum MW = 86$이며, $N(MDI_{MP_n} \cap MDII_{MP_n})$의 속성이 *MP1, MP2* 일 때 공통 유사도 점수는 $N(MD_{MP_n} \cap ADII_{MP_n}) \times \sum MW = 32$이 된다. 가중치를 기반으로 한 유사도 계산을 하게 되면 속성값이 동일할 때 0.3299의 유사도 값을 가진다. 단 *MP1,MP2*에 해당하는 속성값이 일치한다고 가정한다.

$$ \frac{(11+21)}{(11+21+11)+(11+21+6+15+7+9+6+11)-(11+21)} $$

$$ = 0.3299 $$

이와 같이 나머지 3828개의 유사도 값들에서 제일 큰 유사도 값을 가진 정상데이터와 전체 훈련데이터들에 대한 유사도 비교는 〔표 13〕과 같다.

〔표 13〕 매핑된 정상데이터와 훈련데이터 유사도

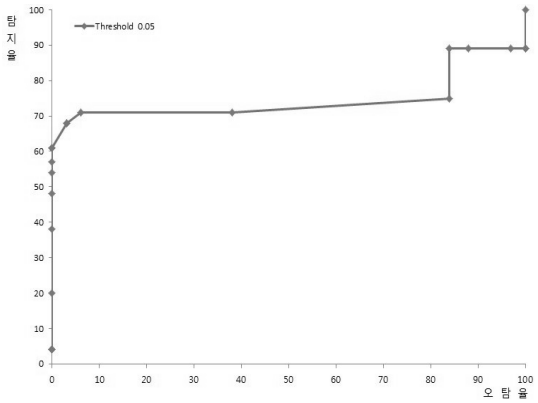
순서	정상데이터	훈련데이터	유사도 값
1	*AD1*	*MDI17*	0.3299
2	*AD2*	*MDI17*	0.3299
3	*AD3*	*MDI17*	0.3478
….	…..	…..	…..
30	*AD30*	*MDI17*	0.3721
31	*AD31*	*MDI17*	0.3299
32	*AD32*	*MDI17*	0.3556

정상코드의 유사도를 0.05 임계치 간격으로 총 20구간으로 나누어 〔표 14〕에 따라 구하였다. 오탐율에 대한 임계치가 0 일 때 가장 높은 유사도를 가지고 있으며 임계치가 1일 때 가장 낮은 유사도를 가지고 있다. 제일 낮은 임계치를 가질 때 유사도 개수가 많아지며 정상코드의 오탐율 결과가 좋았다.

〔표 14〕 임계치에 대한 정상 데이터 오탐율

임 계 치	유사도 개수	오 탐 율
1	0	0
0.91 ~ 0.95	0	0
0.86 ~ 0.9	0	0
0.81 ~ 0.85	0	0
0.76 ~ 0.8	0	0
0.71 ~ 0.75	0	0
0.66 ~ 0.7	0	0
0.61 ~ 0.65	0	0
0.56 ~ 0.6	0	0
0.51 ~ 0.55	0	0
0.46 ~ 0.50	1	3.125
0.41 ~ 0.45	1	3.125
0.36 ~ 0.40	2	6.25
0.31 ~ 0.35	12	37.5
0.26 ~ 0.30	27	84.375
0.21 ~ 0.25	27	84.375
0.16 ~ 0.20	28	87.5
0.11 ~ 0.15	31	96.875
0.06 ~ 0.10	31	96.875
0.01 ~ 0.05	32	100
0	0	100

〔표 15〕에서는 〔표 12〕의 임계치에 대한 훈련데이터 매핑율과 〔표 14〕의 임계치에 대한 정상데이터 매핑율을 정리한다. 즉, 전체 훈련데이터들의 유사도와 탐지율, 정상데이터와 훈련데이터의 유사도와 오탐율을 임계치 기준으로 정리 한다. 오탐율이 높은 정상데이터도 본 실험을 통해 악성코드 유사도 그룹 여부를 확인 할 수 있다. 탐지율과 오탐율이 절충되는 위치는 임계치로 0.4를 기준으로 〔그림 8〕의 ROC 그래프에서 보여주게 된다.



〔그림 8〕 ROC 그래프

〔표 15〕 유사도 비교결과

임 계 치	탐 지 율	오 탐 율
1	3.5714	0
0.95	3.5714	0
0.9	3.5714	0
0.85	21.4285	0
0.8	39.286	0
0.75	50	0
0.7	53.5714	0
0.65	57.1428	0
0.6	60.7142	0
0.55	60.7143	0
0.5	67.8571	3.125
0.45	67.8571	3.125
0.4	71.4285	6.25
0.35	71.4285	37.5
0.3	75	84.375
0.25	89.2857	84.375
0.2	89.2857	87.5
0.15	89.2857	96.875
0.1	89.2857	96.875
0.05	89.2857	100
0	100	100

훈련데이터 그룹들에서 임계치 0.4보다 크면 탐지율이 우수한 것이며, 임계치 0.4보다 작으면 속성값이 다른 악성코드로 분류한다. 훈련데이터와 정상데이터에서는 임계치 0.4보다 작으면 오탐율이 작으며, 임계치 0.4보다 크면 정상데이터를 악성코드로 볼 수 있는 오탐율이 높다. 악성코드 그룹은 최소한의 속성을 포함하고 있으며 전체를 포함 할 수 있다. 다양한 형태의 악성코드 속성을 통해 변종 악성코드가 어느 정도의 최소 속성을 만족하는지 알 수 있게 된다. 또한 유사도 비교를 위한 연산속도를 향상시키기 위해서 알려지지 않은 새로운 악성코드가 네트워크 속성만 가지고 있다면 그룹 분류 시스템에서 네트워크 속성만을 가지는 악성코드만을 비교하면 연산속도는 더욱 향상 될 수 있다. 공통된 악성코드 생성 그룹에서 네트워크 속성 이외의 기타 속성을 사용하는 악성코드는 비교 할 필요가 없게 된다. 악성코드 그룹 분류에서는 악성코드 생성기과 변종 악성코드를 n개 그룹으로 구성하고 속성 및 속성값을 정리한다. 변종 악성코드에 대한 유사도 값을 비교하고 임계치 조건을 만족하고 유사도 값이 높은 악성코드를 악성코드 그룹 분류에 포함시킨다. 하지만, 임계치 조건을 만족하지 못한 악성코드는 알려지지 않은 새로운 악성코드 또는 정상데이터로 분류하게 된다.

IV. 실험 및 결과 분석

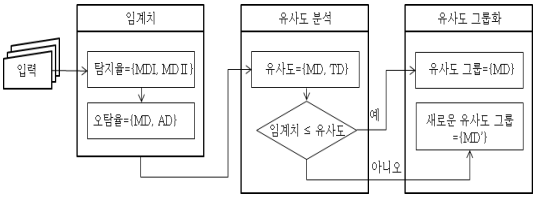
본 장에서는 분류율과 오탐율에서 구한 최적의 임계치로 악성코드 생성기에서 생성된 실험데이터와 전체 훈련데이터에 대한 유사도 비교 실험을 한다.

4.1 실험 데이터 구성

실험을 위해 TELUS Security Lab에서 분류한 악성코드생성기, 다형성 / 메타모픽 바이러스를 바탕으로 전체 훈련데이터(*MD*)56개, 실험데이터(*TD*) 28개로 구성한다.

4.1.1 실험 데이터 유형

실험대상 악성코드로 악성코드생성기에서 새로 생성한 악성코드와 전체 훈련데이터를 사용한다.



〔그림 9〕 악성코드 그룹 분류 구성도

4.1.2 실험 요약

속성, 속성값, 가중치를 가지고 자카드 계수로 유사도 비교를 하는데 있어 악성코드가 다형성과 메타모픽 바이러스로 잘 분류되어야 한다. 그룹 분류에서는 오탐 없이 분류에 대한 정확도가 높아야 한다. 본 실험에서는 악성코드 생성기, 변종 악성코드를 가지고 유사도 비교를 통하여 이에 대한 검증을 하게 된다. 본 실험에서 실험데이터에 대한 유사도 평가는 3장에서 결정한 임계치를 기반으로 한 유사도 그룹 분류로 결정한다.

4.2 실험 결과

4.2.1 새로 생성한 악성코드에 대한 유사도 계산

훈련 데이터를 악성코드 그룹으로 최초 형성한 후 자카드 계수를 사용하여 실험 데이터와 비교하여 유사도를 구한다. 훈련 데이터는 56개로 구성하며 실험 데이터는 28개로 구성한다. 서로 다른 임의의 두 정수 i〉0, j 〉0에 대하여 총 84개의 악성코드에서 순서를 생각하지 않고 2개를 선택하여 이루어진 각각의 집합을 84개에서 2개 선택하는 조합으로 구성하고 ${}_{84}C_2$ 로 나타내면 다음과 같다.

$$ {}_{84}C_2 = \frac{84!}{2! \times 82!} = 3486 $$

〔표 24〕 가중치에 따르면, *TD*의 속성이 *MP8, MP11* 일 때 전체 유사도 점수는 $N(TD_{MP_n}) \times \sum MW = 43$이며, *MD*의 속성이 *MP2, MP8, MP10, MP11*일 때 유사도 값은 $N(MD_{MP_n}) \times \sum MW = 86$이며, $N(TD_{MP_n} \cap MD_{MP_n})$의 속성이 *MP1, MP2* 일 때 공통 유사도 값은 $N(TD_{MP_n} \cap MD_{MP_n}) \times \sum MW = 32$이 된다. 가중치를 기반으로 한 유사도 계산을 하게 되면 속성값이 동일할 때 0.4255의 유사도 값을 가진다. 단 *MP8, MP11* 속성에 해당하는 속성값이 일치한다고 가정한다.

$$ \frac{(9+11)}{(9+11)+(21+9+6+11)-(9+11)} = 0.4255 $$

나머지 3486개의 유사도 값들에서 제일 큰 유사도 값을 가진 실험 데이터와 훈련 데이터들의 유사도 값도 동일하게 구한다. 새로운 악성코드의 유사도 비교는 최대 유사도와 임계치 0.4를 기준으로 한다. 임계치 0.4보다 큰 새로운 악성코드는 훈련 데이터와 공통 그룹으로 판단하며 악성코드 그룹 분류에 편입시킨다.

4.2.2 계산결과

새로운 악성코드의 유사도 실험을 통하여 가중치를 기준으로 만족하는 유사도 비교결과를 가져왔으며 MD 유사도 그룹 분류에 새로운 악성코드를 편입 시켰다. 악성코드 그룹 분류에 속해있는 훈련 데이터들과 실험 데이터의 최대 유사도가 임계치 0.4 보다 큰 경우 새로운 악성코드는 총 28개 중 20개가 공통악성코드 분류 그룹으로 되었다.

4.2.3 결과분석

관련연구에서는 악성코드에 대한 유사도 기준 또는

속성만을 사용하여 악성코드와 정상코드를 판별하는 데 중점을 두는 반면, 본 실험에서는 악성코드 속성 및 가중치를 사용하여 악성코드 유형 판별을 포함한 악성코드 그룹을 탐지율과 오탐율의 최적화된 임계치로 분류 할 수가 있었다.

V. 결 론

본 논문에서는 악성코드 그룹 분류를 위해 속성, 가중치, 유사도 계수 등을 이용하여, 악성코드생성기를 바탕으로 한 전체 훈련데이터(*MD*), 실험데이터(*TD*)에 대한 실험결과를 확인하였다. 악성코드생성기로 생성된 악성코드 간의 유사도를 서로 비교하여 근접한 악성코드 그룹의 악성코드생성기까지 찾아내는데 큰 의미가 있으며 해킹에 사용된 악성코드유형을 예측 할 수 있는 정보로서 활용하게 된다. 악성코드 그룹으로 분류된 속성 정보를 통해 개발자, 지리적 위치, 사용된 컴파일러 등의 악성코드 정보를 확보할 수 있어 향후 발생할 수 있는 악성코드의 개체수를 파악하고 개발자를 추적 할 수 있는 사이버수사 증거자료로 활용하게 된다. 본 연구를 통하여, 새로운 악성코드에 대한 불필요한 분석시간을 감소하고 감염된 경유지, 배포목적 등을 파악하는 수사정보로서 활용하는 사이버수사에 도움이 될 것으로 기대한다. 예측된 실험을 위해 악성코드 중에서 악성코드생성기로 생성된 변종 악성코드로 제한하였다. 향후 다양한 악성코드들에서도 유사도 분석 연구를 지속적으로 진행함으로써 국내외 악성코드 근원지 및 발생 국가 등을 추적하는데 많은 도움을 주리라 기대한다.

참고문헌

〔1〕 Paul Mockapetris, "The History, Present, and Future of Evolution in the DNS," Nominum, pp. 8-19, Oct. 2011.

〔2〕 Kris Kendall, "Practical Malware Analysis," MANDIANT Intelligent Information Security, 2007.

〔3〕 서희석, 최중섭, 주필환, "윈도우 악성코드 분류시스템에 관한 연구," 한국시뮬레이션 논문지, 18(1), pp. 63-69, 2009년 3월.

〔4〕 P.Vinod, V.Laxmi, and M.Gaur, "Survey on malware detection methods," Peedings of the 3rd Hackers' Workshop on Computer and Internet Security, 2009.

〔5〕 한경수, 김인경, 임을규, "API 순차적 특징을 이용한 악성코드 변종 분류 기법," 보안공학연구논문지, 8(2), pp. 319-335, 2011년 4월.

〔6〕 Andrew W, Michael V, Matthew H, Christopher T, and Arun L, "Exploiting similarity between variants to defeat malware," white paper for BlackHat DC, 2007.

〔7〕 Felix Leder, Bastian Steinbock, Peter Martini, "Classification and Detection of Metamorphic Malware using Value Set Analysis," Malware 4th international conference, 2009.

〔8〕 G. Wagener, R. State, and A. Dulaunoy, "Malware behaviour analysis," Journal in Computer Virology, Nov. 2007.

〔9〕 Konard R, Thorsten H, Carsten W, Patrick D, and Pavel L, "Learning and classification," Springer Verlag Berlin Heidelberg, 2008.

〔10〕 정용욱, 노봉남,"공격용 툴킷 및 변형코드의 유사성 기준 선정," 보안공학연구논문지, 9(1), pp. 31-44, 2012년 2월.

〔11〕 S. Momina Tabish, M. Zubair Shafiq, Muddassar Farooq, "Malware Detection using Statistical Analysis of Byte-Level File Content," CSI-KDD'09, 2009.

〔12〕 Peter Szor, Peter Ferrie, "Hunting for metamorphic," Virus bulletin conference, pp. 123-142, Sep. 2001.

〈저자소개〉

정 용 욱 (Yong-Wook Chung) 정회원
1995년 2월: 금오공과대학교 토목공학과 졸업(학사)
2000년 10월: University of London, Information Security, MSc
2006년 8월~2008년 8월: 전남대학교 정보보호협동과정 박사수료
2006년~현재: 경찰청 수사국 소속, 사이버테러대응센터 디지털포렌식팀 연구관
〔관심분야〕 악성코드·네트워크 포렌식, 침해사고 대응, 정보보호

노 봉 남 (Bong-Nam Noh) 종신회원
1978년 2월: 전남대학교 수학교육과 졸업(이학사)
1982년 2월: KAIST 대학원 전산학과 졸업(이학석사)
1994년 2월: 전북대학교 전산과 (이학박사)
1983년~현재: 전남대학교 전자컴퓨터공학부 교수
2000년~현재: 시스템보안연구센터 소장
〔관심분야〕 디지털포렌식, 시스템 및 네트워크 보안, 정보사회와 사이버 윤리