Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

컴퓨터 바이러스 탐지를 위한 퍼지 진단시스템

A Fuzzy Diagnosis System for Detecting Computer Viruses

이 현 숙[1]

[1] 서울시 구로구 동양공업전문대학 전산정보학부
E-mail: hsrhee@dongyang.ac.kr

요 약

본 논문에서는 컴퓨터 바이러스 정보 구축과 탐색에 학습기능을 도입함으로 새로 발생하는 바이러스를 찾아내어 대처할 수 있도록 설계된 퍼지 진단 시스템 FDS를 제안한다. FDS에서는 FCM 알고리즘을 사용하여 알려진 정보의 클러스터를 형성하고 이에 전문가의 지식을 포함하는 지식베이스를 구축한다. 진단을 위한 컴퓨터 파일에 대하여 그 파일의 결정 상태를 확인하고 이미 저장된 지식베이스를 바탕으로 바이러스 침입에 대한 정보를 보고하도록 설계되어있다. 이 시스템은 이미 알려진 테스트 데이터와 이전에 알려지지 않은 새로운 테스트 데이터를 실험데이터로 준비하여 그 성능을 테스트 한다. 제안된 시스템이 알려지지 않은 컴퓨터 바이러스의 경우도 효과적으로 진단할 수 있는 타당성을 보이고 있다.

Key Words : Fuzzy Diagnosis System, Knowledge Acquisition, Computer Viruses, Detection Status

1. 서 론

컴퓨터 바이러스 탐지를 위하여 널리 사용되어온 n-gram 분석방법은 알려진 바이러스 파일의 분석과 단순 매칭에 기반을 두고 있다[1]. 그러므로 알려지지 않은 바이러스를 탐지하지 못하고 그 바이러스에 의하여 시스템이 공격을 받고 난 후 바이러스 유형과 파일상태를 분석한 후에야 탐지 알고리즘에 반영될 수 있다. 그러는 사이 시스템은 손상되고 또 다른 형태의 바이러스가 만들어질 것이다. 이를 위하여 바이러스 전문가의 휴리스틱한 규칙이 적용되기도 했으나 간단하고 유연성이 없어 쉽게 노출되며 실세계에 적용하기 어려운 단점을 가지고 있다. 이에 바이러스정보구축과 탐지과정에 학습기능과 전문가의 경험지식을 체계적으로 통합하여 새로운 바이러스 파일도 탐지하고 대처할 수 있는 방법이 필요 하게 되었다.

본 논문에서는 학습기능과 경험지식을 통합할 수 있는 컴퓨터 바이러스 탐지를 위한 퍼지 진단 시스템(Fuzzy Diagnosis System, FDS)을 설계한다[2]. FDS에서는 주어진 파일로부터 바이러스 관련 정보를 구축하기 위한 사전정보(a priori information)로 사용하기 위하여 데이터 클러스터 분석기법을 사용한다. 널리 알려진 퍼지 클러스터 분석 알고리즘, fuzzy *c*-menas(FCM)을 적용하여[3] 얻어진 *c*개의 각 클러스터에 파일 수집에 참여한 바이러스 전문가의 지식을 부착하여 지식베이스를 구축한다. 다음은 FDS의 진단과정으로서 바이러스를 탐지하고자하는 주어진 파일데이터에 대하여 각 클러스터와의 퍼지 소속 값을 구한 후 그 값을 이용하여 결정상태(decision status)를 확인하고 지식베이스의 정보를 바탕으로 진단결과를 출력해 준다. 이미 알려진 테스트 데이터와 이전에 알려지지 않은 새로운 테스트 데이터를 실험데이터로 준비하여 그 성능을 테스트 하여 제안된 방법의 타당성을 검증해 보고자한다.

2. 퍼지 진단 시스템

본 논문에서 제안한 FDS는 컴퓨터 바이러스 탐지를 위한 특징들을 해석하여 판단하는 분류과정에 학습기능과 전문가시스템이 제공하는 지식구축 및 설명기능을 가지고 있다. 이러한 FDS는 지식획득모듈과 진단모듈로 나누어 설명할 수 있다.

지식획득 모듈은 이미 준비된 알려진 파일

들의 특징데이터를 시스템에 학습시켜 사전지식을 얻도록 해 준다. 이러한 학습을 위하여 기계학습이론, 신경망 분야의 많은 이론이 있으나 퍼지이론과 통계적인 접근방법을 기초로 학습하는 퍼지 클러스터 분석 알고리즘 FCM을 사용한다[3]. 클러스터분석 결과 학습되어 형성된 클러스터들은 그 대표정보를 가지고 있으며 파일을 준비하고 처리하는 전문가는 파일 유형이나 바이러스 유형 등의 관련정보를 부착시킨다. 이때 실세계에서 얻은 데이터를 클러스터 분석에 의하여 처리한 후 형성된 클러스터를 바탕으로 단순화 시킨 후 지식을 표현하므로 지식획득이 용이하다. 단지 정상 파일인지 바이러스 파일인지 만을 판정하려면 이를 지시하는 정보만 부착시킬 수도 있고 진단 후 대처할 수 있는 solution 함수를 호출 할 수도 있다.

진단 모듈에서는 진단을 원하는 파일의 추출된 특징데이터를 가지고 지식획득모듈에서 사전지식으로 학습하여 가지고 있는 클러스터 정보를 참조하여 퍼지 소속 값을 계산한다. 퍼지 소속 값은 각 데이터의 이미 저장되어 있는 클러스터 정보에 대한 일치도를 나타낸다. 이러한 일치도를 바탕으로 주어진 데이터 x_j 가 이미 학습된 사전지식을 가지고 분류할 수 있는지를 판정하게 된다. 이를 데이터의 결정상태라고 하며 퍼지 상태(fuzzy status)와 분명한 상태(crisp status)로 분류 하게 된다[2].

마지막 단계로 진단 모듈은 주어진 데이터의 진단상태가 crisp status이면 가장 큰 소속 값을 가지는 클러스터 l 에 부착된 정보를 출력해 주며, fuzzy status이면 경고를 보내며 전문가에게 분석을 요청한다. 전문가에 의해 분석된 데이터는 다음 FDS 갱신을 위한 데이터로 수집될 수 있다.

보통의 컴퓨터 바이러스 진단 시스템의 경우도 파일 수집, 특징패턴 추출, 데이터 표현, 분류알고리즘 적용 등의 전 과정에 바이러스 전문가가 참여하고 단독시스템을 구축하는 것은 어려운 것으로 알려져 있다. FDS는 새로운 바이러스를 처리할 수 있는 학습기능과 퍼지이론, 전문가시스템의 설명기능을 도입하여 전문가시스템의 지식획득의 어려움도 해소하고 바이러스를 탐지한 후에 처리 하고 시스템을 갱신할 수 있는 방법도 제공하고 있다.

3. 실험 및 고찰

FDS의 입력을 준비하기 위하여 VX heaven[4]으로 부터 400개의 바이러스 파일과 윈도우시스템 실행파일로부터 200개의 정상파일을 수집하였다. 특징추출과정을 통하여[5] 만들어낸 600*26의 데이터는 Ldata라고 부르며 이를 가지고 FDS의 지식획득 모듈의 입력데이터로 사용되어 시스템의 사전정보를 구축 하게 된다. 또한 진단에 사용될 실험 데이터 300*26의 TdataA와 300*26의 TdataB를 마련한다. TdataA는 사전정보구축에 사용된 데이터로부터 선택되어 이미 알려진 파일을 진단하는 것이고 TdataB의 경우는 같은 소스로부터 수집되었으나 FDS에 알려지지 않은 임의의 파일을 진단하는데 사용된다.

지식획득 모듈의 입력을 위하여 준비된 Ldata를 만들어 낸 파일은 2종류의 바이러스 파일과 유사한 종류의 정상 실행파일로부터 수집하였으므로 클러스터의 수 c를 3으로 하였다. 그러므로 Ldata를 지식획득모듈의 입력으로 하고 3개의 클러스터를 형성하고 각각에 대하여 benign data, virus fileI, virus fileII와 같은 단순한 정보를 부착하여 지식베이스에 저장하였다. 실세계 응용의 경우 전문가는 바이러스 유형정보나 바이러스 대처 요령이나 solution 함수로의 호출 등의 실제적인 정보를 부착하여 그 바이러스가 탐지되면 적절한 상담을 해주거나 전문가를 돕는 시스템이 될 수 있을 것이다.

이렇게 구성된 지식베이스를 가지고 TdataA의 데이터를 진단모듈의 입력으로 사용하여 각 데이터의 세 개의 클러스터에 대한 퍼지 소속 값을 구하여 각 데이터의 진단상태를 구한다. 같은 범주의 데이터 셋에 대하여 100번 실험한 결과 평균 95.7%는 crisp상태로 4.3%는 fuzzy 상태로 판정되었다. crisp 상태로 판정된 데이터는 평균 96.4% 정확하게 분류되었으므로 사전정보 구축에 사용된 데이터를 진단하는 경우 약 92.3%의 정확도를 얻었다. 제안된 FDS가 미리 알려지지 않은 데이터에 대하여 어떻게 적응하는지 알아보기 위하여 TdataB의 데이터를 진단모듈의 입력으로 사용하여 각 데이터의 진단상태를 구한다. TdataA의 경우와 마찬가지로 100번의 실험을 한 평균을 구하면 87.2%가 crisp status로 판정되고 12.8%가 지식획득모듈에서 가지고 있는 사전지식으로는 판정하기 어려운 상태인 것으로 나타났다. 이 때 crisp status로 판정된 경우 93.5%가 정확하게 분류되었으므로 TdataB의 경우 약 81.5%의 분류정확도를 얻었다. 이러한 실험 결과로부터 제안된 시스템 FDS는 컴퓨터바이러스 진단을 위한 시스템으로 도입될 수 있는 가능성을 확인할 수 있다.

4. 결론

제안된 시스템 FDS는 지식획득 모듈에서

퍼지 클러스터 분석과정을 통해 실세계 데이터를 분석하여 클러스터를 형성하므로 손쉽게 지식베이스를 구축할 수 있으며 진단과정에서는 이미 구축된 사전정보를 활용할 수 있는 방법을 제공하며 결정 상태를 분류하여 처리하므로 효율적으로 처리할 수 있음을 확인하였다. 또한 앞으로 계속적인 연구를 통하여 구축된 지식베이스에서 진단할 수 없는 새로운 파일의 경우 전문가에 의하여 판정되고 시스템에 추가적으로 학습될 수 있는 점증적 학습기법(incremental training method)이 도입되어야 할 것이다.

참 고 문 헌

[1] Abou-Assaleh, Nick Cercone, Vlado Keselj, and Ray Sweidan, "Detection of New Malicious Code Using N-grams Signatures, Proceedings of the Second Annual Conference on Privacy, Security and Trust (PST'04), pp. 193-196, 2004.
[2] 이현숙, "컴퓨터 바이러스 분류를 위한 퍼지 클러스터기반 진단시스템", 한국정보처리학회논문지, 제14권-B 제 1호, 2007.
[3] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum press, New York, 1981.
[4] VX Heaven : http://vx.netlux.org
[5] Jianyong Dai, Joohan Lee and Morgan C. Wang, "Detecting Unknown Computer Virus Using Data Mining Techniques", Business Intelligent Symposium, poster presentation, April, 2006.