사이버 ISR에서의 점진적 학습 방법과 일괄 학습 방법 비교

신경일*, 윤호상**, 신동일*, 신동규*
*세종대학교 컴퓨터공학과
**국방과학연구소
e-mail:sgi@gce.sejong.ac.kr

Comparison of incremental learning method and batch learning method in Cyber ISR

Gyeong-Il Shin*, Hosang Yooun**, DongIl Shin*, DongKyoo Shin**
*Dept of Computer Engineering, Se-jong University
**Agency for Defense Development

요 약

사이버 ISR을 통하여 정보를 획득하는 과정에서 데이터를 추출하고 이를 스스로 가공하여 의사결정에 도움을 줄 수 있는 에이전트를 연구하는 과정에서 폐쇄망에 침투했을 경우 이를 효과적으로 감시 정찰할 수 있는 방법을 논의한다. 폐쇄망으로 인하여 침투한 컴퓨터에 심어진 에이전트는 C&C서버와 원활한 교류가 불가능하게 되는데, 이때 스스로 살아남아 지속적으로 데이터를 수집하며, 분석을 하기 위해서는 한정된 자원과 시간을 활용하여야 발각되지 않고 계속하여 임무를 수행할 수 있다. 특히 분석하는 과정에서 많은 자원과 시간을 활용하는 때 이를 해결하기 위해 본인은 점진적 학습방법을 이용하는 것을 제안하며, 일괄학습 방법과 함께 비교하는 실험을 해보았다.

1. 서론

최근 컴퓨터 기술이 점차 발달되면서 종이로 쓰이던 문서들이 점차적으로 컴퓨터 텍스트 파일로 저장되고, 이러한 정보들이 네트워크를 통해 전세계 사람들과 정보를 공유할 수 있게 되었다. 기술이 점 차 발달되면서 이렇게 쌓인 데이터를 활용하여 많은 산업 분야를 걸쳐 활용되게 되고 있다. 그러나 이러한 정보를 악용하여 개인이나 기업, 국가 등을 타겟으로 하는 공격 또한 많이 발생하고 있다. 특히 이는 전쟁에도 이용이 되는데, 한국의 경우 아직 전쟁이 끝나지 않은 휴전국가이므로, 이러한 사이버전에 대한 피해를 무시할 수 없다. 북한군은 많은 비용을 소모하지 않고, 국가적 혼란 및 금전적 손실 등 효율적으로 타겟에게 피해를 줄 수 있으며, 기동전과 동시에 사이버공간에서 사이버공격을 통하여 군사적, 정치적 등 다방면 적으로 혼란을 주며, 전장에서 우위를 달성할 수 있기에 북한군은 사이버 전력에 많은 관심을 보이며, 약 6천여 명의 사이버전사를 보유하고 있다. 이로 인하여 한국군도 사이버전에 대한 많은 관심을 가지고 있는 상황이다.

특히나 정보는 전투 시 우위를 달성할 수 있는 매우 중요한 요소로 아군측의 전반적인 상황과 각 종 정보와 함께 상대방의 정보를 얼마나 보유하고 있느냐에 따라서 전투의 승패가 갈릴 수 있다. 특히 전쟁의 영역이 사이버 공간으로 확대된 현재 사이버영역에서의 정보는 매우 중요해졌으며, 그로 인해 사이버 ISR(Intelligence Surveillance Reconnaissance) 또한 매우 중요하다. 본 논문에서는 이러한 중요한 정보를 취득하기 위해 정보를 수집하는 단계와 마지막에 분석하는 단계에서 기계학습 및 딥러닝을 이용하여 의사결정에 도움을 주는 에지전트 모델을 제안했다.

적군 측 정보를 수집할 때 적군의 네트워크에 있는 에이전트가 C&C 서버를 통하여 아군측의 명령을 받고 작전을 수행할 수 있다. 이 때 적군의 네트워크가 폐쇄망일 경우는 C&C서버와 정기적인 교류가 불가능하게 되어 스스로 정보의 중요도를 평가하며, 필요없는 정보를 버리고, 필요한 정보들만 모아서 속성별로 분류하여 보관하고 있다가 외부망과의 통신이 가능할 때까지 앞의 과정을 반복하며 지속적인 업데이트를 하는 방법을 생각해보았다.

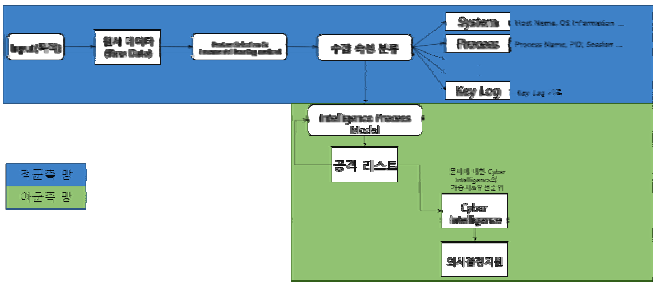

그림 1. 공격시 사이버 ISR 운용 프로세스

또한 폐쇄망에서 운영되는 에이전트는 스스로 학습하

고 분석할 수 있으며, 학습이 완료된 모델에 새로운 데이터가 들어오면 완료된 모델에 새로운 데이터를 학습하는 업데이트가 가능한 모델이 필요하다. 추가적으로 적이 에이전트가 침투한 사실을 알아차리지 못하게 제한된 리소스와 시간을 이용하여 분석하여야 한다.

본인은 이러한 조건을 만족시키기 위해 점진적 학습 방법을 이용하여 학습 시키는 것을 제안한다. 본 논문에서는 일괄처리 학습 방법과 점진적 학습 방법을 이용하여 학습하는 두 가지 방법을 여러 가지 측면에서 분석을 해보았으며 이를 비교해보았다.

2. 관련연구

2.1 일괄 학습 방법 vs 점진적 학습 방법

일반적으로 기계학습 이용 시 많이 사용되는 학습 방법으로는 일괄 학습 방법이 있겠다. 일괄 학습 방법이란 개념형성에 필요한 모든 예제들을 한 번에 모두 제공하며, 모든 예와 반례를 한 번에 처리하여 지식을 생성하는 방법이다. 일괄 학습 방법은 지속적으로 들어오는 데이터를 학습하는 경우와 제한된 리소스를 사용하여 학습을 시킬 때 적합하지 않다. 그 이유는 기존데이터의 처음 부분부터 새로 들어온 데이터의 마지막까지 처음부터 끝까지 다시 한번 학습과정을 거쳐야하기 때문이다. 결과적으로 소요되는 리소스와 시간이 증가할 수 밖에없다.

반면 점진적 학습 방법은 이용 가능한 자료들을 이용하며, 하나 이상의 개념 가설 형성을 하고 점차적으로 추가로 주어지는 예제들을 이용하여 가설을 개선한다. 현재 주어진 예와 반례로부터 지식을 생성하고 계속 새로운 예와 반례가 생길 때마다 점진적으로 현재의 지식을 수정하는 방향으로 진행되는 학습 방법이다. 이는 인간의 개념 학습 방법과 매우 유사하며, 여러 개의 개념습득에 유용한 방법이다.

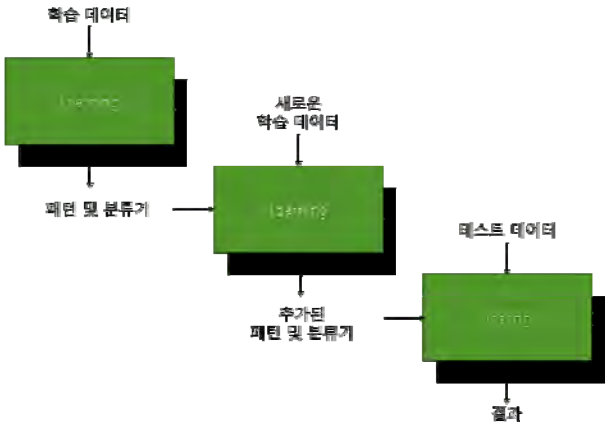

그림 2. 점진적 학습 방법

점진적 학습 방법은 일괄 학습 방법과 달리 새로 들어오는 데이터를 학습할 때 기존의 학습이 끝난 모델에 이어서 학습을 이어서 하는 모델로 이전에 학습한 데이터를 제외하고 새로 생성된 데이터에 관해서만 학습을 진행한다. 그로 인하여 학습 시 사용되는 시간도 단축할 수 있으며, 사용되는 리소스 또한 감소하게 된다.

2.2 KDD CUP 99 데이터셋

본 실험에서는 사이버 ISR시 사용할 적합한 데이터를 만들지 못하여, KDD CUP 99 데이터 셋을 이용하여 실험해보았다. KDD CUP은 ACM에서 매년 열리는 데이터 마이닝 대회로 KDD CUP 99 Dataset은 KDD CUP 99년도에 쓰인 데이터이다. 해당 데이터는 네트워크 패킷 데이터이로, 침입 탐지 시스템 관련 데이터 셋으로 자주 쓰이고 있다. 주요 공격은 DoS(Denial of Service), R2L(unauthorized access from a remote machine), U2R(unauthorized access to local superuser privileges), Probe 총 4가지로 구별되며, Normal 라벨까지 총 5가지의 라벨이 존재한다. 또한 KDD Cup training dataset에는 24개의 공격 유형이 포함되어 있고 test dataset에는 training dataset에 없는 14개의 유형이 추가로 포함된다.

3. 실험

본 실험에서는 적군 측 네트워크가 폐쇄망이라고 가정하에 실험하였으며, 적군 컴퓨터에 침입한 에이전트가 지속적으로 수집되는 데이터를 학습할 때 적군에게 발각되지 않도록 제한된 자원으로 정보 분석을 시도하는 상황에 적합한 학습법을 찾기 위해 일괄 학습 방법과 점진적 학습 방법을 비교하는 실험을 해보았다.

해당 실험은 CPU기반으로 학습하여 실험을 하였으며, 사용된 CPU는 Ryzen 7 1700X이며, MOA 라이브러리를 이용하여 실험하였다. 그리고 실험에 사용된 데이터 셋은 KDD CUP 99 Dataset의 train 데이터와 test 데이터를 이용하였으며, 두 개의 데이터 중 train 데이터만을 가지고 학습을 해보고 두 번째로는 train데이터와 test 데이터를 합쳐 학습하였다. 이때 학습 시 사용되는 메모리양과 학습경과시간에 대해 비교를 해보았다.

표 1. 데이터 셋 설명

데이터 셋	데이터 개수	데이터 크기
train	494,020개	51.1MB
test	311,029개	39.9MB
train + test	805,049개	83.9MB

먼저 일괄 학습 방법을 이용하는 알고리즘을 이용하여 실험을 해보았으며, 해당 실험에서 쓰인 알고리즘은 트리 알고리즘의 하나인 Decision Tree를 이용해보았다. 메모리 양의 제한을 약 14545MB일 때 train 데이터만을 학습할 경우 약 52.576초가 소요되었으며, 이때 총사용한 메모리의 양은 2884.5MB이다. train과 test데이터를 모두 학습한 경우는 앞의 시험과 동일하게 메모리양의 제한은 14545MB였으며, 총 143.694초의 시간이 소요되었고, 총 사용된 메모리의 양은 4995MB였다. train 데이터만 사용하였을 경우 초당 약 55MB를 사용하였고, train 데이터와 test데이터를 학습한 경우에는 초당 약 35MB를 사용하였다. 당연하게 데이터가 커지면 커질수록 사용하는 메모리와 시간이 증가한다. 그렇다면 메모리를 1000MB 제한했을 경우에는 어떠한 결과가 나오는지 실험을 해보았다. train 데이터만을 학습 시킬 때 937MB가 사용되었으며, 총 51.582초가 경과되었다. train데이터와 test데이터를 합쳐 학습하였을 경우에는 heap 사이즈가 부족하여 학습을 하지 못하였다. 타겟 컴퓨터에 에이전트가 침투했을 경우 일괄 학습 방법을 이용하여 분석을 할 때 메모리에 제한을 주지 않고 해당 학습을 진행하였다면 사용되는 메모리의 양이 너무 많아 적에게 발각될 수 있으며, 반대로 메모리를 제한하여 학습을 진행하면, 데이터가 너무 클 경우 학습 진행을 못하는 경우가 발생할 가능성이 높다. 이러한 이유로 제안한 사이버 ISR모델에서는 일괄 학습 방법이 적합하지 않다고 판단했다. 그리하여 일괄 학습 방법의 문제점을 해결하기 위해 점진적 학습 방법을 이용해보았다.

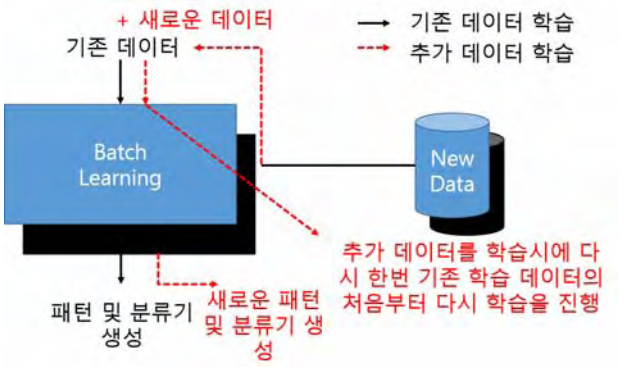

그림 3. 일괄 학습 과정

먼저 일괄 학습 방법을 적용한 트리 알고리즘의 하나인 Hoeffding Tree를 사용하여 실험해보았다. 앞서 Decision Tree를 이용하여 실험한 것과 동일하게 실험을 진행해보았다. 먼저 일괄 학습 방법으로 메모리양의 제한을 14545MB를 주었을 때 학습 경과시간은 18.541초가 소요되었으며 사용한 메모리의 양은 5295MB이다. train과 test 데이터를 합쳐 실험하였을 경우 학습 경과 시간은 33.192초 이며, 사용한 메모리양은 4088MB이다.

두 번째로 점진적 학습 방법을 적용하여 Hoeffding Tree를 학습하는 실험해보았다. 메모리양의 제한을 14545MB로 주었을 경우 train만을 학습하였을 때 경과 시간은 18.828초이며, 사용된 메모리양은 787MB이다. train과 test 합쳐 학습한 경우 경과 시간은 34.644초였으며 사용된 메모리양은 806MB이다.

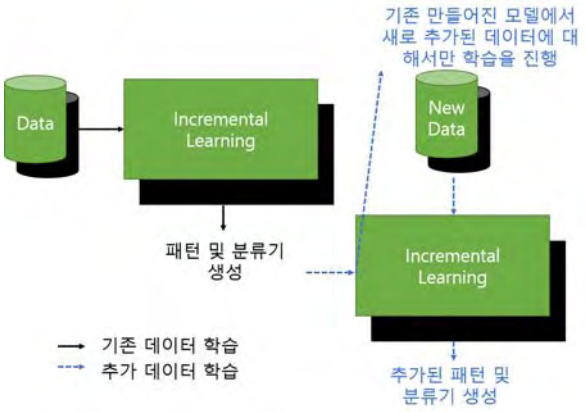

그림 4. 점진적 학습 과정

표2. 모델별 학습 경과시간 및 사용된 메모리양 비교

구분	모델	학습 데이터	제한된 메모리양	사용한 메모리양	경과시간
일괄 학습	DecisionTree	train	14545MB	2884.5MB	52.576초
	DecisionTree	all	14545MB	4995MB	143.694초
	DecisionTree	train	1000MB	937MB	51.582초
	DecisionTree_	all	1000MB	heap space 부족으로 인하여 측정불가	
	Hoeffding Tree	train	14545MB	5295MB	18.541초
	Hoeffding Tree	all	14545MB	4088MB	33.192초
	Hoeffding Tree	train	1000MB	781.5MB	22.987초
	Hoeffding Tree	all	1000MB	928MB	38.642초
점진적 학습	Hoeffding Tree	train	14545MB	787MB	18.828초
	Hoeffding Tree	all	14545MB	806MB	34.644초
	Hoeffding Tree	train	1000MB	761.5MB	18.951초
	Hoeffding Tree	all	1000MB	778MB	29.554초

표 2를 보면 점진적 학습법을 사용한 Hoeffding Tree모델이 가장 사용한 메모리의 양이 적었다. 또한 점진적 학습 방법을 사용한 경우 일괄학습을 사용한 경우보다 경과 시간이 단축되었으며 이는 점진적 학습 방법이 더 적은 메모리를 사용하였는데도 학습 시간이 줄어든 것을 볼 수 있었다.

4. 결론

본 실험에서는 폐쇄망에서 스스로 생존하면서 스스로 학습하는 에이전트 모델에 적용하기 적합한 학습 방법으로 일괄 학습 방법과 점진적 학습 방법 두 가지를 비교하여보았다. 실험 결과 점진적 학습 방법은 일괄 학습 방법에 비해 적은 메모리를 사용하여 더 적은 학습 시간이 소요된다는 것을 알 수 있었다. 또 한 본 실험에는 포함이 되어있지 않지만 일괄 학습 방법은 메모리 제한을 200MB 주면 힙 공간 부족으로 인하여 학습이 실행되지 않지만 점진적 학습 방법을 이용하면 5MB에서도 학습이 실행되는 것을 알 수 있었으며, 메모리 제한이 20MB일 때 점진적 학습 방법을 통해 학습을 하였을 경우 약 22초의 학습 경과 시간이 걸리게 되는데, 이는 일괄 학습을 781.5MB 메모리를 사용한 경우보다 약간 더 우수하다. 또한 점진적 학습 방법은 표 1에서 볼 수 있듯이 데이터의 크기가 크면 메모리의 크기와 단축된 학습 경과 시간이 일괄 학습에 비해 확연하게 좋다는 것을 비교할 수 있었다. 이러한 실험 결과를 통해 제안한 사이버 ISR모델에서 점진적 학습 방법을 사용하는 것이 월등하게 우수하다는 것을 알 수 있었다. 향후 사이버전에 적합한 데이터 셋이 만들어진다면 해당 방법을 이용하여 시나리오를 바탕으로 실험을 해볼 계획이며, 향후 계획으로는 해당 방법을 이용해 생성된 분류기의 성능 또한 평가해 볼 계획이다.

Acknowledgement

본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다(UD160066BD).

참고문헌

[1] Baek, Hey-Jung, and Young-Tack Park. "The Study on Improvement of Cohesion of Clustering in Incremental Concept Learning." The KIPS Transactions: PartB 10.3 (2003): 297-304.

[2] Fuangkhon, Piyabute, and Thitipong Tanprasert. "An incremental learning algorithm for supervised neural network with contour preserving classification." Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th International Conference on. Vol. 2. IEEE, 2009.

[3] Bifet, Albert, et al. "Moa: Massive online analysis." Journal of Machine Learning Research 11.May (2010): 1601-1604.

[4] 신경일, 신동일, 신동규, 윤호상. "IDS 알고리즘에 대한 탐지율 연구 비교." 한국정보처리학회 2017년 춘계학술발표대회 Vol.24 No.001 (2017): 0223~0226.

[5] 지현정, 신동규, 신동일, 김용현, 김동화. "인공신경망을 통한 KDD CUP 99와 NSL-KDD 데이터 셋 비교." 한국정보처리학회 2017년 춘계학술발표대회 Vol.24 No.01 (2017): 0211~0213.