교차선별을 통한 하이브리드 클라우드의 로그 기반 보안 정책 생성 안정화 연구

김예진[1], 김태림[2], 김건민[3], 김경백[3]
[1]전남대학교 소프트웨어공학과 학부생
[2]전남대학교 인공지능학부 학부생
[3]전남대학교 인공지능융합학과
ye031010@jnu.ac.kr, ktr0706@jnu.ac.kr, geonminkim@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

Stablizing Security Policy Generation from Logs in Hybrid Clouds via Cross-Validation

Yejin Kim[1], Taerim Kim[2], Geonmin Kim[3], Kyungbaek Kim[3]
[1]Dept. of Software Engineering, Chonnam National University
[2]Dept. of Artificial Inelligence, Chonnam National University
[3]Dept. of AI Convergence, Chonnam National University

요 약

하이브리드 클라우드의 등장으로 네트워크 경계가 모호해지고 공격이 지능화되며 정적 규칙에 기반한 기존 보안 모델은 한계를 드러내며, 이에 대안으로 머신러닝 및 AI를 활용한 자동화된 동적 보안 모델이 주목 받고 있다. 본 연구는 하이브리드 클라우드 환경의 동적 보안 정책 생성 모델의 견고성을 높이고, 보안 정책의 안정화를 위하여 RandomForest와 L1-Logistic 모델을 결합한 이중 모델 교차선별 기법을 제안하며, 그 효과를 실험을 통해 실증한다. 이 접근법은 데이터 노이즈와 단일 모델로 인한 편향을 제거하여 모델의 불안정성을 효과적으로 제거하여 정책의 신뢰도와 재현성을 높여, 예측 가능하고 안정적인 보안 자동화 프레임워크를 구축하는데 기여한다.

1. 서론

빅데이터의 등장과 클라우드 컴퓨팅의 발전으로 인한 디지털 전환이 가속화됨에 따라 많은 기업들이 온프레미스(On-Premise) 인프라의 견고성과 클라우드의 유연성을 결합한 하이브리드 클라우드 환경을 데이터 저장 및 관리를 위한 핵심 IT 전략으로 채택하고 있다[1]. 그러나, 이질적인 온프레미스와 퍼블릭 및 프라이빗 클라우드가 결합된 구조는 네트워크 경계를 모호하게 만들고, 구조적 복잡성을 심화시켜 기존의 보안 체계로는 대응하기 어려운 새로운 위협 요인을 만들어 낸다[2-4]. 특히, 하이브리드 클라우드는 동적이고 분산된 환경으로, 정적 시그니처나 규칙에 기반한 전통적인 보안 모델은 약점을 보인다. 이에 대한 대안으로, 인공지능(AI) 및 머신러닝(ML) 기술을 활용하여 대규모 로그 데이터를 실시간으로 분석하고, 변화하는 공격 패턴을 학습하여 동적으로 보안 정책을 적용하는 연구가 활발히 진행되고 있다[5-7]. API 게이트웨이를 통해 트래픽을 중앙에서 제어하고, 수집된 로그를 머신러닝 모델로 분석하여 정책 엔진의 규칙으로 자동으로 업데이트하는 아키텍처는 지능형 위협에 신속하게 대응할 수 있는 효과적인 방법론으로 평가 받는다.

이러한 머신러닝 기반 정책 자동화는 새로운 과제를 제시한다. 특히, 정책 생성이 단일 머신러닝 모델의 판단에 의존할 경우, 해당 모델이 가진 고유의 편향이나 특정 노이즈에 과적합될 위험이 존재한다. 이로 인해 학습이 반복될 때 마다 생성되는 정책의 내용이 크게 변화할 수 있으며, 이는 잦은 정책 변경으로 인한 운영 오버헤드, 예측 불가능성, 그리고 오탐 증가로 이어져 시스템 전체의 안정성을 저해하는 요인이 된다. 본 연구에서는 이러한 정책 생성의 안정성 문제를 해결하기 위해 RandomForest와 L1-Logistic[8]을 결합한 이중 모델 교차선별 전략을 제안하고, 그 효과를 실증한다. 이를 통해 복잡한 하이브리드 클라우드 환경을 위한 신뢰도 높은 자동화 프레임워크 구축에 기여한다.

2. 관련 연구

2.1 전통적 보안 모델의 하이브리드 클라우드 에서의 한계

전통적 보안 모델은 네트워크 경계를 기준으로 내부와 외부를 명확히 구분하고, 외부로부터의 위협을

차단하기 위한 방화벽, 접근 제어 목록 등의 기술을 구축하는 방식으로 설계되었다. 이는 온프레미스 환경에서는 유효하게 기능하였으나, 하이브리드 클라우드 환경은 퍼블릭 및 프라이빗 클라우드와 온프레미스 환경이 상호 연결된 구조로 네트워크 경계를 물리적으로 명확하게 설정하기 어렵다[2-4]. 이는 전통적인 방화벽이나 경계 기반의 보안 정책이 효과를 발휘하기 어려운 환경을 만든다. 특히, 클라우드 환경은 각각의 클라우드 서비스 제공 업체가 제공하는 보안 관리 도구와 API가 상이하여, 사용자는 하이브리드 클라우드 환경에서 보안 정책과 접근 제어의 일관성을 확보하기 어렵다.

2.2 머신러닝 기반 하이브리드 클라우드 보안 아키텍처

정적 규칙 기반 시스템의 한계를 극복하고 하이브리드 클라우드 환경의 보안을 강화하기 위해 인공지능(AI)와 머신러닝(ML) 기반의 기술을 활용하는 연구들이 활발히 제안되며 진행되고 있다[5-7]. 특히, 다양한 플랫폼에 분산된 서비스들의 트래픽을 단일 지점에서 통합 관리하는 API Gateway를 설치하여 모든 API 요청을 중앙에서 로깅하고 제어하여 OPA(Open Policy Agent)와 같은 별도의 정책 엔진에 위임하는 구조가 제안되었다. 이러한 구조는 각기 다른 클라우드 제공업체의 상이한 정책 문법이나 관리 도구로 인한 파편화 문제를 해결하고, 일관된 보안 정책을 전체 인프라에 적용할 수 있게 한다. 이러한 구조는 수집된 모든 로그의 텍스트 데이터를 벡터화하고, 머신러닝 모델을 통하여 정상 및 공격 요청을 판별한다. 공격으로 분류된 로그에서 발견된 새로운 악성 키워드나 잠재적 위험 패턴은 보안 정책으로 자동으로 갱신되며, 이러한 피드백 루프는 보안 자동화와 함께 제로데이 공격이나 다형성 공격과 같이 알려지지 않은 위협에 대응하는 시스템의 적응력을 크게 향상시킨다.

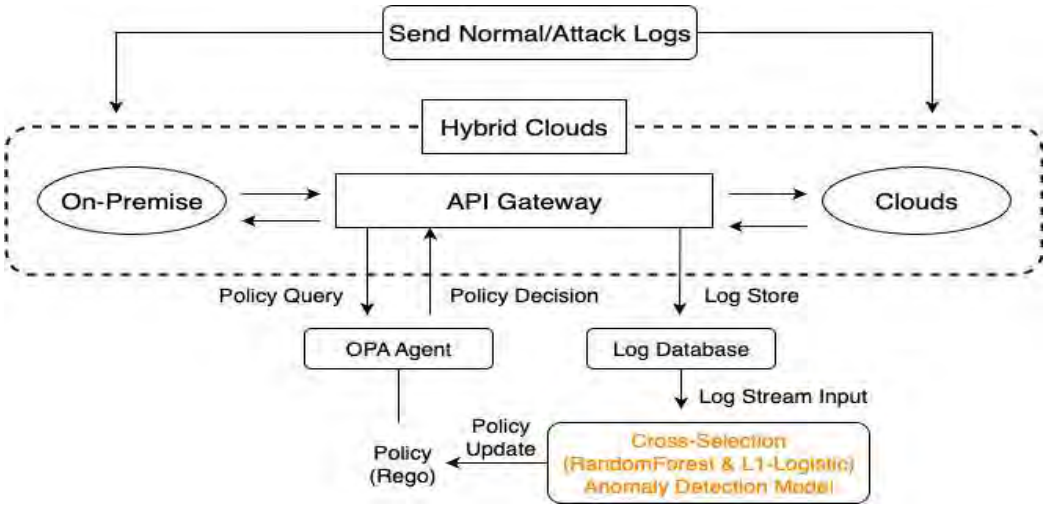
3. 제안 방법: 이중 모델 교차선별

머신러닝 모델, 특히 복잡한 비선형 모델은 학습 데이터의 작은 변화에도 예측 결과가 민감하게 반응하는 불안정성을 보일 수 있으며, 보안 정책 생성과 같이 높은 신뢰도가 요구되는 분야에서 이러한 불안정성은 큰 위험 요소로 작용한다. 본 연구는 기존의 단일 모델 기반 정책 생성 방식의 불안정성 문제를 해결과 모델 성능의 향상을 위하여 두 개의 상호보완적인 머신러닝 모델(RandomForest, L1-Logistic)을 결합한 이중 모델 교차선별 모델을 제안한다.

3.1 RandomForest와 L1-Logistic

제안된 방법론은 RandomForest와 L1-Logistic 모델로 구성된다. 결정 트리의 앙상블인 랜덤 포레스트는 특징 간의 비선형적 상호작용을 포착하고, 복잡한 결정 경계를 모델링하는데 탁월하다. 각 특징이 트리의 노드 분기에서 불순도를 얼마나 효과적으로 감소시키는지를 기준으로 중요도를 측정하여, 전반적인 예측 성능에 기여하는 특징들을 선별한다. L1-Logistic은 Logistic Regression에 L1 정규화를 적용한 모델로, 예측에 큰 영향을 미치지 않는 특징들의 가중치를 0으로 만들어 모델을 희소하게 만든다. 이를 통해 예측에 가장 결정적인 최소한의 특징 집합을 선택할 수 있으며, RandomForest가 놓칠 수

(그림 1) 제안 모델(Cross-Selection) 구성



있는 선형 관계를 명확히 잡아낸다. 또한, 각 계수의 부호(+/-)를 통해 해당 특징이 공격(+) 또는 정상(-) 판단에 기여하는 방향성을 명확히 해석할 수 있어, deny/allow 규칙 설계에 직관적이다. 이 두 모델은 상호 보완적이다. RandomForest가 높은 카디널리티를 가진 특징에 주요도를 편중시킬 수 있는 반면, L1-Logistic은 가장 핵심적인 소수의 특징만을 남겨 이러한 편향을 보완한다.

3.2 교차선별 동작 원리

교차선별 기반 정책 성생 파이프라인은 3단계로 진행된다. 첫째, 모델 학습 단계이다. RandomForest 모델이 특징 중요도를 계산한다. 수집된 API 로그 데이터를 전처리 및 벡터화한 후, RandomForest 분류기를 학습시킨다. 학습된 모델로부터 각 특징의 불순도 기반 중요도를 계산한다. 동시에, 동일한 API 로그 데이터에 대하여 L1-Logistic 모델 또한 학습을 진행한다. 그 후, 두 모델이 각각 제안한 상위 300개의 특징 집합 간의 교집합을 최종 특징 집합으로 채택한다. 이 과정을 통해 한 모델에서만 일시적으로 중요하게 평가된 노이즈성 특징을 효과적으로 필터링하고, 두 모델 모두가 동의하는 안정적인 핵심 패턴만을 남긴다. 최종 선택된 특징들을 기반으로 정규 표현식 규칙을 생성하고, 이를 OPA에 투입하기 위한 Rego 형식으로 변환한다. 생성된 정책은 동적 정책 관리 시스템을 통해 API 게이트웨이에 연동된 정책 엔진으로 실시간으로 배포된다.

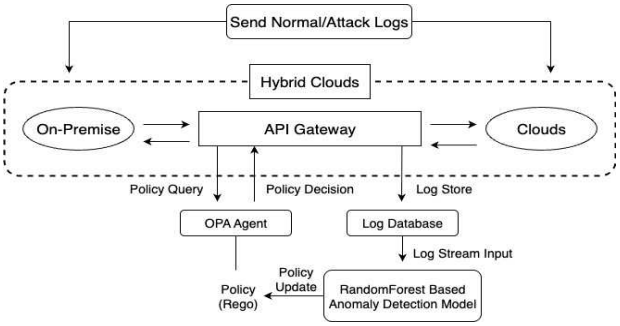
4. 보안 정책 적용 실험 설계

본 연구는 RandomForest 단일 모델을 사용하는 방식과, 제안된 모델의 성능을 정량적으로 비교 분석하는 것을 목표로 한다. 이를 위해 특징 선택 방식에 따라 <표 1>과 같이 두 개의 모델을 설정하고, 그 효과를 검증한다.

<표 1> 실험 비교 대조군

Model	설명
Random Forest	RandomForest 단일 모델을 사용하여 정책을 생성하는 방식
Cross -Selection	본 연구에서 제안하는 RF-L1 이중 모델 교차선별 방식

(그림 2) 실험 대조군 RandomForest 모델 구성



4.1 실험 절차 및 평가 지표

실험에는 실제 네트워크 트래픽을 기반으로 구성된 Bahviour-Centric Cybersecurity Center(BCCC) 데이터셋을 사용한다. 이 데이터셋은 정상 트래픽 로그와 함께 Malware, Phising, Spam 등 다양한 유형의 공격 로그를 포함하고 있어, 제안 모델의 탐지 성능과 일반화 능력을 평가하기에 적합하다. 특히, 모델의 안정성을 평가하기 위하여 각 모델 별로 3회의 반복 실험을 진행한다. 훈련 로그와 테스트 로그를 8:2로 분리하였으며, 평가 지표로는 모델의 공격 로그 차단 정확도를 측정하는 공격 로그 차단율(Attack Block Rate)과 더불어 모델 성능의 안정성을 측정하기 위하여 공격 로그 차단율에 대한 표준편차(Std.Dev)를 이용한다.

<표 2> RandomForest 단일 적용 모델과 Cross-Selection 모델의 실험 결과 비교

Model	Round	Total Num of Attack Logs	Blocked Num of Attack Logs	Attack Block Rate(%)	Std. Dev
RandomForest	1	1736	1431	82.43	
	2	1758	1520	86.46	
	3	1751	1544	88.18	
	Avg	1748	1498	**85.69**	**2.41**
Cross-Selection	1	1721	1520	88.32	
	2	1747	1542	88.27	
	3	1746	1551	88.83	
	Avg	1738	1538	**88.47**	**0.25**

5. 실험 결과

<표 2>에 제시된 바와 같이, Cross-Selection 모델은 RandomForest 모델 대비 2.78%의 정확도 향상을 보였다. 특히, 3회의 실험의 공격 차단율에 대하여 RandomForest 모델은 2.41의 표준편차를 보였으나, Cross-Selection 모델은 0.25의 표준편차를 보이며 모델의 안정성이 대폭 향상되었음을 확인하였다. 이는 단일 모델 평향과 데이터 노이즈의 영향을 최소화하였음을 보여주는 것으로, 보안 자동화는 안정적인 성능을 보이는 것이 중요함을 고려하였을 때, 이는 유의미한 성능 개선으로 평가된다.

6. 결론 및 향후 연구

본 연구에서는 하이브리드 클라우드 환경에서 머신러닝 기반 동적 보안 정책 생성 시 발생할 수 있는 안정성 문제를 해결하고, 모델의 견고성을 강화하기 위해 RandomForest와 L1-Logistic을 결합한 이중 모델 교차선별 기법을 제안하여, 서로 다른 강점을 가진 두 모델의 합의를 통해 2.78%의 공격 차단율 향상과 함께, 성능 표준 편차를 2.41에서 0.25로 감소시키며 모델 안정성 향상을 도출하였다. 향후 연구에서는 교차선별 기법을 LightGBM, Transformer 기반 모델 등의 모델 조합으로 확장하여 성능을 비교 분석할 수 있다. 또한, 실시간 위협 수준이나 비즈니스 영향도를 고려하여 동적으로 정책을 선별하는 강화학습 도입 등을 고려할 수 있다. 이를 통해 더욱 안정적인 보안 자동화 시스템을 구축할 수 있으며, 급변하는 위협 환경에 대응해야하는 실질적인 운영 환경에서 정책 관리의 복잡성을 낮추고 관리 용이성을 증대시킬 수 있다.

Acknowledgement

본 연구는 2025년도 과학기술정보통신부 및 정보통신기획평가원의 소프트웨어중심대학사업의 연구결과로 수행되었습니다.(2021-0-01409)(34%)본 연구는 한국인터넷진흥원(KISA)-정보보안 특성화대학 지원사업의 지원을 받아 수행된 연구임(33%)본 연구성과는 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2025-25398164)(33%)

참고문헌

[1] M. Gaianu, "On Premise Data Center vs CLOUD," 2023 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2023, pp. 1068-1071

[2] S. B. Mallisetty, G. A. Tripuramallu, K. Kamada, P. Devineni, S. Kavitha and A. V. P. Krishna, "A Review on Cloud Security and Its Challenges," 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2023, pp. 798-804

[3] L. B. Bhajantri and T. Mujawar, "A Survey of Cloud Computing Security Challenges, Issues and their Countermeasures," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 376-380

[4] G. Raktate, K. Shelar, P. Parjane, S. Pangavhane, S. More and S. R. Deshmukh, "A Survey on Security Issues and Challenges in Cloud Computing," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-5

[5] C. Anjani, R. M. Balajee, G. Divya, Y. S. Sree, K. Padmanabham and S. S. Srithar, "Evolving Threats and AI Solutions for Modern Hybrid Cloud Architectures," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 478-484

[6] D. K. Seth, K. K. Ratra and A. P. Sundareswaran, "AI and Generative AI-Driven Automation for Multi-Cloud and Hybrid Cloud Architectures: Enhancing Security, Performance, and Operational Efficiency," 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, pp. 00784-00793

[7] Y. Mansouri, V. Prokhorenko, and M. A. Babar, “An automated implementation of hybrid cloud for performance evaluation of distributed databases,” Journal of Network and Computer Applications, Volume 167, 2020, ISSN 1084-8045

[8] Muchlinski D, Siroky D, He J, Kocher M. Comparing Random Forest and Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analsysis. 2016:24(1):87-103. doi:10.1093/pan/mpv024