이 일 구 교수 지도
석사학위 청구논문

주성분 분석 기반의 안전한 데이터 공유 메커니즘

2023

성신여자대학교 대학원
미래융합기술공학과
신 나 연

주성분 분석 기반의 안전한 데이터 공유 메커니즘

이 일 구 교수 지도

이 논문을 석사학위논문으로 제출함

2023년 5월

성신여자대학교 대학원

미래융합기술공학과

신 나 연

인 준 서

신나연의 석사학위 논문으로 인준함

2023년 5월

심사위원장 임 연 섭 (서명 또는 인)

심 사 위 원 김 경 진 (서명 또는 인)

심 사 위 원 이 일 구 (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

모바일 디바이스의 사용량이 증가하면서 생성되는 데이터의 양이 폭발적으로 증가하고 있다. 최근에는 생성된 방대한 양의 데이터를 활용하여 다양한 가치를 창출하고 있다. 하지만 개인의 단말기에서 수집되는 데이터의 경우, 환경이나 디바이스, 이용자의 특성 등에 따라 수집되는 정보가 다르므로 Non-IID 문제가 발생할 수 있다. Non-IID 문제를 해결하는 대표적인 방안은 데이터 공유지만, 원본 데이터를 전송할 경우 사용자의 프라이버시 문제가 발생할 수 있으며, 중간자가 데이터를 탈취할 우려 또한 존재한다. 따라서 본 연구에서는 주성분 분석을 이용한 안전한 데이터 공유 메커니즘을 제안하여 종래의 문제를 해결하고자 시도한다. 각 노드들은 주성분 분석 모델을 생성한 뒤 차원을 축소하여 공유하며, 공유 이전 생성한 주성분 분석 모델을 사용해서 모델을 복구한다. 본 연구는 실험을 통하여 제안하는 메커니즘을 정상자와 공격자 관점에서 평가하였다. 이때 정상자 관점에서는 데이터 공유 전후의 정확도가 일정하게 유지되나 공격자 관점에서는 정확도가 하락함을 보였으며, 최적의 Components 값을 도출하였고, 최대 42배 메모리 효율적임을 입증하였다. 또한 최적의 Components 값일 때 가장 높은 프라이버시 척도를 보여 제안하는 메커니즘의 성능을 증명하였다.

목 차

표 차 례

그림 차례

Ⅰ. 서 론

1. 배경

인공지능(AI, Artificial Intelligence)과 ICBM(IoT, Cloud, Big data, Mobile)의 결합으로 인간의 고차원적인 정보 처리기술을 ICT(Information & Communications Technology)로 구현하는 지능 정보화 사회가 도래하였다[1]. 병원 진료기록이나 위치 정보, 거래 내역 등 다양한 종류의 데이터를 AI와 접목하고자 하는 요구가 증가하고 있으며[2, 3, 4], 다양한 단말 기기에서 실시간으로 데이터를 수집 및 분석하여 새로운 가치를 창출하고 있다[5, 6].

하지만 개인의 단말기에서 수집되는 데이터의 경우 환경이나 디바이스, 이용자의 특성 등에 따라 수집되는 정보가 다르므로 Non-IID 문제가 발생할 수 있다[7]. IID란 Independent and Identically Distribution의 약자로 각각의 단말 기기에서 수집된 데이터가 독립적이고 동일한 확률분포를 가지는 것을 의미한다. 반면에 Non-IID는 분산된 단말 기기에 종속적이며 불균형한 확률분포를 가지는 데이터가 존재함을 의미한다[8]. 즉 데이터가 균일하지 않고 소수인 클래스가 존재하는 클래스 불균형 문제(Class Imbalance Problem)가 발생한다. 임의의 노드 데이터에서 전체 모집단 분포를 표현하기 어렵고 이는 곧 학습 성량 저하로 이어진다. Non-iid 문제를 해결하기 위한 대표적인 방법은 데이터를 공유하는 방안이며[7], 각 장치들은 소유하고 있는 데이터를 서로 공유함으로써 Non-IID를 해결할 수 있다.

하지만 수집된 데이터는 장치 소유자의 생활 패턴이나 생활 정보 등과 같은 민감정보를 포함할 수 있으며, 데이터 공유 시 중간자가 데이터를 탈취할 수 있다[9, 10]. 따라서 데이터 공유를 위해서는 사용자의 프라이버시 침해

문제와 중간자 탈취 문제를 우선적으로 해결해야 한다[11].

한편, 주성분 분석(PCA, Principal Component Analysis)은 데이터의 차원을 축소하는 비지도학습 모델이며, 원본 데이터들의 특성을 최대한 유지하면서 데이터의 차원을 축소한다[12]. 축소할 데이터의 차원을 지정할 수 있고, 축소한 차원을 본래 차원으로 재구성할 수 있다. 재구성 시에는 차원축소 시 소실한 분산만큼의 차이가 생길 수 있으며, 이를 재구성 오차(Reconstruction Error)라고 한다. 주성분 분석을 통해 축소된 데이터를 복원할 경우 재구성 오차로 인한 손실로 차원 축소 전 데이터와 동일한 데이터를 얻을 수 없으나 비슷한 특성을 가진 재구성 데이터를 얻을 수 있다.

본 연구에서는 Non-IID를 해결하기 위해 주성분 분석 기반의 안전한 데이터 공유 메커니즘을 제안한다. 각 노드는 주성분 분석 모델을 생성한 뒤 차원을 축소하여 데이터를 공유하고, 공유받은 데이터는 자신의 주성분 분석 모델을 사용하여 재구성한다. 해당 메커니즘은 축소된 데이터를 공유할 경우 중간자가 탈취하더라도 문제가 되지 않음을 증명하며, 프라이버시 측면에서 평가하기 위한 척도를 제시한다.

본 논문의 기여점은 다음과 같다.

1) Non-IID 문제를 해소하기 위해 보다 안전한 데이터 공유 메커니즘을 제안한다.

2) 메커니즘을 혼돈행렬뿐만 아니라 메모리 사용량, 프라이버시 측면에서도 평가하였으며, 모든 부분에서 성능이 개선되었음을 입증하였다.

3) 데이터가 탈취되었을 상황을 가정하여 공격자와 정상자 관점을 나누어 평가하였다.

2. 논문 구성

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 제안하는 아키텍처의 주요 기술인 주성분 분석의 개념을 정리하고 종래의 안전한 데이터 공유를 위한 다양한 시도를 분석한다. 제 Ⅲ장에서는 제안하는 메커니즘에 대하여 정의하며, 제 Ⅳ장에서는 실험환경 및 내용에 대하여 서술하고 결과를 분석한다. 제 Ⅴ장에서는 결론을 작성하며 논문을 마친다.

Ⅱ. 선행 연구

1. 주성분 분석(PCA, Principal component analysis)

차원 축소란 고차원의 원본 데이터를 저차원의 데이터로 변환하는 방법이며, 특징 선택(Feature Selection)과 특징 추출(Feature Extraction)로 구분된다. 특징 선택은 원본 데이터의 특징 간 상관관계를 비교하고 가장 상관관계가 높은 특징의 부분집합을 선택한다. 특징 추출은 원본 데이터를 최대한 보존하는 방법으로 새로운 특성을 추출하여 축을 생성하는 방법이다[13].

주성분 분석은 특징 추출에 속하며, 원본 데이터의 분산을 최대한 보존하도록 새로운 축을 생성하는 차원 축소 도구이다[12]. 데이터를 투영하였을 때 분산을 최대한 보존하며 직교하는 주축을 찾는 것을 목표로 하며, 이는 행렬이 변화에 작용하는 주축의 방향을 의미하는 고유벡터(Eigen vector)에 특징을 사상시키는 것을 의미한다. 고유값(Eigenvalue)은 고유벡터 방향으로 늘어난 공간 벡터의 크기를 의미한다.

주성분 분석에서는 고유벡터를 구하기 위해 공분산 행렬에 대한 특이값 분해(SVD, Singular Value Decomposition)를 할 수 있다. 주성분 분석을 하는 과정은 다음과 같다.

첫째, 행렬 X에 대하여 확률변수의 평균을 0으로 맞추는 센터링 작업을 한다. 수식 (1)과 같이 행렬 X가 $m \times n$ 형식의 데이터 행렬일 때 m은 샘플(표본)의 개수이고 n은 데이터의 특징(피처)을 의미하는 확률변수의 개수이다.

$$X = \begin{pmatrix} \vdots & \vdots & & \vdots \\ X_1 & X_2 & \cdots & X_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{m \times n} \qquad (1)$$

이때 행렬 X n열의 X_i는 확률변수가 가지는 모든 데이터를 의미한다. X의 열에 대한 평균을 수식 (2)와 같이 구할 수 있으며 확률변수 x_i의 평균은 $\overline{x_i}$로 표시할 수 있다. 이후 각 열별로 데이터에서 평균을 뺌으로써 수식 (3)과 같이 표현할 수 있고 이 센터링된 행렬을 $\widetilde{X}$라 한다.

$$\overline{X_1} = \frac{1}{m}(x_{11} + x_{21} + \cdots + x_{m1}),\ \ldots,\ \overline{X_n} = \frac{1}{m}(x_{1n} + x_{2n} + \cdots + x_{mn}) \qquad (2)$$

$$\begin{aligned} \widetilde{X} &= X - \overline{X} \\ &= \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} - \begin{pmatrix} \overline{X_1} & \overline{X_2} & \cdots & \overline{X_n} \\ \vdots & \vdots & & \vdots \\ \overline{X_1} & \overline{X_2} & \cdots & \overline{X_n} \end{pmatrix} \end{aligned} \qquad (3)$$

둘째. 센터링 된 데이터 행렬 $\widetilde{X}$에 대하여 특이값 분해를 한다. 특이값이란 $m \times n$ 행렬 A가 존재할 때 $A^T A$에 대한 고유값 λ의 제곱근 값을 의미한다. 특이값 분해란 행렬 A를 특이값을 가지는 행렬로 분해하는 것으로 수식 (4)와 같이 정의된다.

$$\widetilde{A} = U \Sigma V^T \qquad (4)$$

행렬 U와 V는 $m \times m$, $n \times n$ 형태의 직교 행렬이다. 행렬Σ는 대각성분에 특이값을 가지는 사각행렬이 된다. 행렬 V는 $A^T A$에 대한 고유벡터이며, 열벡터는 주축이 되며, US는 열벡터들이 원본 데이터를 주축에 정사영하여 얻은 주성분 점수 (PC, Principal Component Score)가 된다. 즉, $\widetilde{X}$에 대한 특이값 분해를 통해 해당 행렬의 데이터가 어떤 방향으로 분산되어있는지 알 수 있다.

셋째, 특이값 s_i를 이용하여 원본 데이터의 분포비율을 확인하고 차원을 축소할지 결정한다.

수식 (5)는 i번째 PC의 공분산을 의미하고 특이값이 큰 순서대로 고유벡터를 정렬하면 결과적으로 중요도를 기반으로 주성분을 구성할 수 있게 된다.

$$\frac{s_i^2}{tr(Cov(\widetilde{X}))} \qquad (5)$$

수식 (6)은 i번째 PC가 원본데이터의 분산을 보존하는 비율을 의미한다.

$$\frac{s_i^2}{s_1^2+\cdots+s_n^2} \qquad (6)$$

즉, 수식 (7)은 첫 번째 주성분부터 k번째 주성분까지의 누적분산비율이다. 위의 순서에 따라 사용할 주성분의 개수나 원하는 원본 데이터 분산비율로 차원을 축소할 수 있다.

$$\frac{\sum_{i=1}^{k} s_i^2}{s_1^2+\cdots+s_n^2} \qquad (7)$$

마지막으로 축소한 데이터를 다시 재구성할 수 있다. Fig. 1처럼 행렬 X에 원본 데이터 공분산의 고유 벡터를 곱하여 투영하여 차원을 축소했으므로 다시 고유벡터를 곱함으로써 이를 복원할 수 있다.

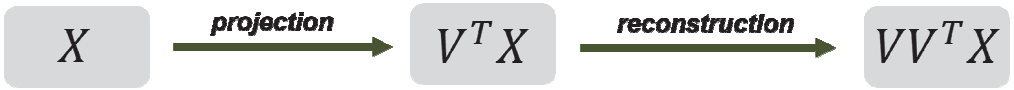


FIGURE 1. 주성분 분석의 투영 및 재구성

2. 선행 연구

본 장에서는 노드 간 안전한 데이터 공유를 목적으로 선행된 연구에 대해 살펴보고 해당 분야에서의 한계점에 대해 서술한다.

[14]에서는 금융분석 분야에서 동형암호를 통해 개인정보를 수집하고, 동형암호화된 데이터로 기계학습 모델을 생성하여 신용정보를 산출하는 방안에 대한 연구를 진행했다. 저자는 동형암호 기술을 통해 민감한 데이터를 안전하게 수집하고 처리하여, 데이터의 기밀성을 보호하고자 했지만 동형암호화된 데이터를 이용하여 연산할 경우 데이터 크기가 증가하고 이로 인해 연산속도가 느려진다는 한계점이 있었다.

[15]에서는 연합학습을 기반으로 VCPS(Vehicular Cyber-Physical Systems)의 데이터 유출 문제를 해결하고자 하였다. 차량 및 RSU(Road Side Unit)와 같은 노드들은 수집된 데이터를 로컬에서 학습하여 모델의 매개변수만 전송하기 때문에 데이터의 유출 위험을 제거할 수 있다. 또한 데이터의 유출을 감지하기 위한 연합학습을 배치하여 2차적으로 데이터 프라이버시에 대한 보안성을 높이고자 했다. 하지만 중앙 클라우드에 의존하므로 트래픽이 증가하면 병목현상이 발생할 수 있으며[17], 클래스 불균형이 발생하였을 경우 직접적인 데이터 공유가 불가능하다.

[16]에서는 모바일 클라우드 환경에서 데이터를 안전하게 공유하기 위해 대칭키(AES) 알고리즘 및 CP-ABE(Ciphertext-Policy Attribute-Based Encryption)를 기반으로 하는 이중암호화와 이중 프록시 모델을 기반으로 하는 시스템을 제안한다. 사용자는 메세지를 암호화 및 복호화하기 위해 대칭키 알고리즘과 CP-ABE를 사용하고, SignCryption 프록시에서는 서명 및 암호화를 진행하며 Decryption 프록시에서는 복호화를 진행한다. 이를 통해 데이터 공유 시 안전한 서명 암호화 및 함호 복호화를 가능하게 한다. 하지만

CP-ABE를 사용하기 때문에 사용자의 속성이 변경될 때마다 키를 관리해야 하며, 인증 기관을 거치는 등 보안을 고려한 만큼 상당히 복잡한 프로세스를 가지고 있다.

[17]에서는 분산 수집된 의료 이미지 데이터를 프라이버시 측면에서 안전하게 공유하고자 딥러닝과 블록체인을 결합한 새로운 데이터 공유 메커니즘을 제안한다. 블록체인에는 허가된 병원만 스마트 컨트렉트를 통해 참여할 수 있으며, 참여한 구성원들은 분산형 파일 시스템인 IPFS(InterPlanetary File System)를 통해 로컬에서 학습한 데이터의 가중치를 공유한다. 블록체인을 활용하여 분산된 정보를 동기화하고 글로벌 모델을 구성함으로써 모델의 분류 성능을 개선하였다. 하지만 해당 모델의 경우 블록체인 접근을 위해 스마트 컨트렉트를 이용할 때 다른 조직의 허가를 받는 과정이 복잡하고, 블록체인을 이용한 만큼 컴퓨팅 파워와 비용 측면에서 한계점이 존재한다.

[18]는 온라인 데이터 공유의 한계점을 극복하고 보다 효율적인 데이터 공유 솔루션을 제공하고자 제시하였다. 해당 연구가 제안하는 모델은 다양한 공격에서 제안 모델이 저항 가능함을 입증하였고 계산 및 비용측면에서 가볍고 효율적임을 입증하였다. 하지만 해당 모델의 경우 데이터 공유시에 발생할 수 있는 사이버 공격에만 초점을 맞추었으며 데이터 유출에 대해서는 고려하지 않았다는 한계점이 존재한다. 즉, 프라이버시를 고려하지 않았다.

[19]는 클라우드에서의 안전한 데이터 전송을 위하여 IPFS를 사용하는 새로운 블록체인 기반의 보안 분산 시스템을 제안하였다. 제안하는 방식에서 데이터의 소유자는 암호화된 파일을 IPFS에 업로드하고 이후 데이터 보안을 위해 N개의 비밀 섹션으로 분리한다. 데이터 소유자는 이 보안 데이터에 대한 접근 권한을 얻기 위해 접근 권한을 추가로 작성해야 하며 보안을 위해 2단계의 키 관리 시스템을 사용한다. 이를 통해 제안하는 방식은 기존의 중앙 집중식 시스템의 단일 지점 오류를 지우고 통신 및 계산과 관련된 오버헤드를

줄일 수 있다. 또한 보안 측면에서는 신뢰할 수 없는 클라우드 서버를 비롯하여 악의적인 사용자에 대해서도 효과적인 저항이 가능하다. 하지만 해당 모델의 경우 성능과 비용, 보안, 프라이버시 측면에서 충분히 안전한지에 대한 평가가 이루어지지 않았다는 한계점이 존재한다.

[20]는 산업용 헬스케어 시스템에서의 안전한 데이터 공유를 위하여 블록체인 및 스마트 컨트랙트를 딥러닝 기술과 병합하여 안전하고 효율적인 데이터 공유 프레임워크인 PBDL(Permissioned Blockchain and Deep Learning)을 제안한다. 결과적으로 PBDL은 종래의 기술보다 성능이 우수함을 입증하였으나 복잡한 프레임워크는 메모리 및 컴퓨팅 파워 측면에서 효율적이지 못하다는 한계점이 존재한다.

종래의 연구들을 분석해 보았을 때 전반적으로 보안과 비용 및 사용성 관점에서 트레이드 오프를 해결하지 못하고 있다. 따라서 비용과 메모리, 프라이버시 관점에서의 트레이드 오프를 해결하기 위한 연구가 필요하다. 따라서 본 연구는 프라이버시와 메모리, 성능 측면에서 모델을 평가하여 종래 연구들의 한계점을 보완하고자 한다.

TABLE I. 노드 간 안전한 데이터 처리 연구

Ref.	Explanation	Limitation
[14]	• 동형암호를 통해 개인정보를 수집하고 기계학습을 통해 신용도를 산출함 • 동형암호를 사용하여 데이터의 유출이나 손실 없이 연산이 가능함	• 암호화 시 데이터의 크기가 증가함 • 암호화 시 연산 속도가 느려짐
[15]	• 연합학습을 기반으로 VCPS의 데이터 유출 문제를 해결하고자 함 • 차량 및 RSU와 같은 노드(객체)들은 수집된 데이터를 로컬에서 학습하고 가중치만 중앙서버로 보내어 글로벌 모델을 학습함	• 병목현상으로 인한 지연시간 발생 • 클래스 불균형이 발생하였을 경우 직접적인 데이터 공유가 불가함
[16]	• 대칭키 알고리즘 및 CP-ABE를 기반으로 하는 이중암호화를 사용 • 암호화 복호화 시 각각에 대한 프록시를 사용함	• CP-ABE를 사용하기 때문에 사용자의 속성이 변할 때마다 키를 새로 발급해야함 • 전반적인 프로세스가 복잡함

[17]	• 의료시스템의 폐암 예측성능을 개선하고자 딥러닝과 블록체인을 활용 • 허가된 집단만 스마트 컨트랙트에 참여할 수 있으며, 직접적인 데이터가 아닌 로컬 학습한 가중치만 공유하여 글로벌 모델을 생성함	• 스마트 컨트랙트 이용 시 다른 조직의 허가를 받아야하기 때문에 복잡함 • 블록체인을 이용한 만큼 컴퓨팅 파워와 비용 측면에서 한계점이 존재
[18]	• 신뢰할 수 있는 클라우드 서버에서 암호화 및 기계학습 기반의 상호 인증 프로토콜 제시	• 데이터 공유 시에 발생할 수 있는 사이버 공격에만 초점 • 프라이버시를 고려하지 않음
[19]	• 클라우드에서의 안전한 데이터 전송을 위하여 IPFS를 사용하는 새로운 블록체인 기반의 보안 분산 시스템을 제안	• 성능과 비용, 보안, 프라이버시 측면에서 충분히 안전한지에 대한 평가가 이루어지지 않음
[20]	• 블록체인 및 스마트 컨트랙트를 딥러닝 기술과 병합하여 안전하고 효율적인 데이터 공유 프레임워크인 PBDL을 제안	• 복잡한 프레임워크는 메모리 및 컴퓨팅 파워 측면에서 효율적이지 못함

Ⅲ. 주성분 분석 기반의 안전한 데이터 공유 메커니즘

본 장에서는 앞에서 언급한 종래 연구들의 한계점과 데이터 공유가 가지는 근본적인 문제점을 보완하기 위한 새로운 데이터 공유 메커니즘을 제안한다. 해당 메커니즘은 차원 축소 알고리즘인 주성분 분석에 기반하여 차원이 대폭 축소된 데이터를 공유함으로써 전송 도중에 공격자의 탈취 공격이 있어도 프라이버시 침해의 우려가 없는 데이터를 만드는 것에 초점을 맞춘 방식이다.

제안하는 메커니즘의 전반적인 구조도는 Fig. 2과 같다. 이때 Node 1은 클래스 불균형한 노드라고 가정하며, Node 2는 균형잡힌 데이터를 충분히 가지고 있다고 가정한다. 가정 사항에 대해서는 4장 1절에서 다시 한 번 언급한다.

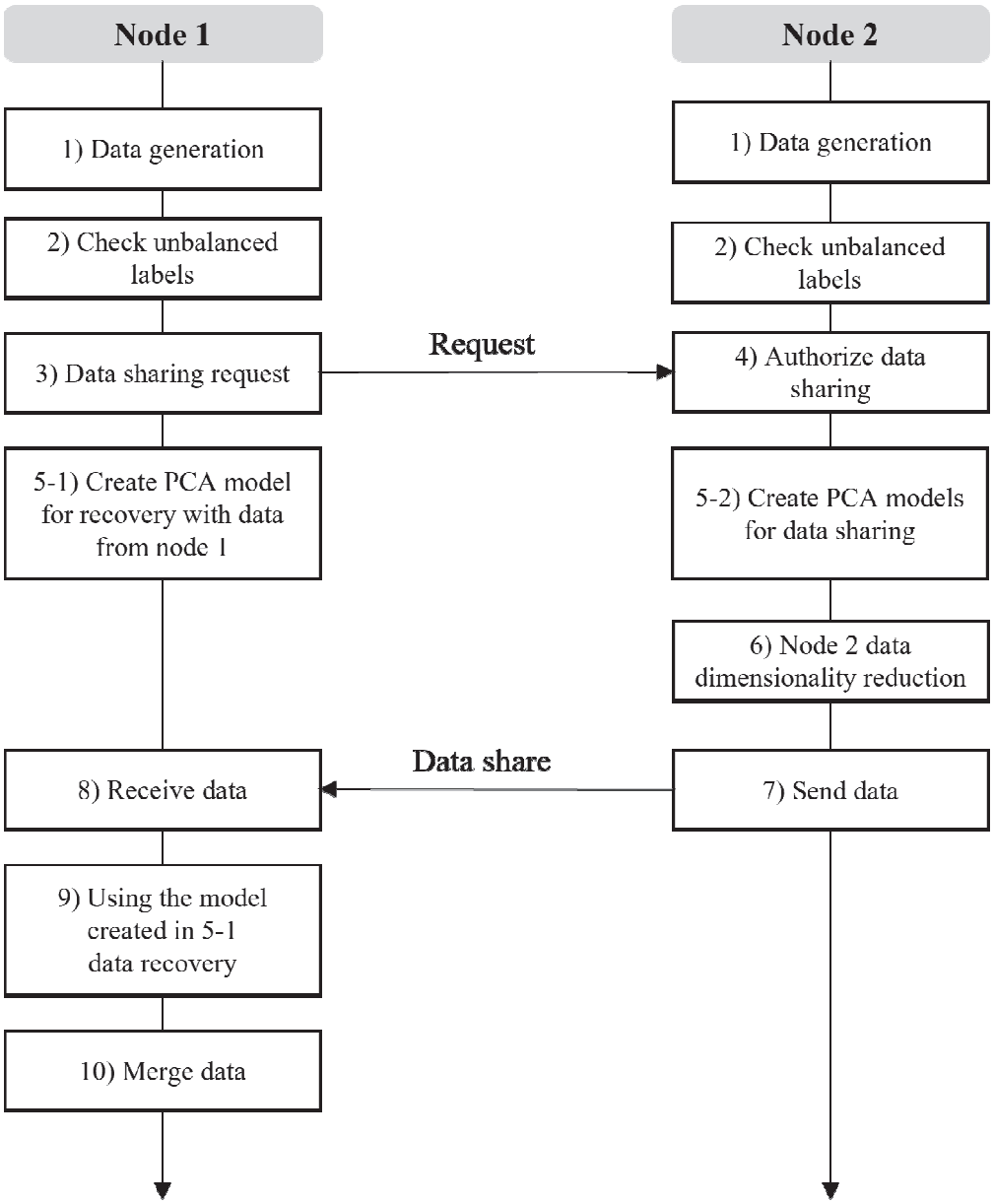


FIGURE 2. 주성분 분석 기반의 안전한 데이터 공유 메커니즘 흐름도

1)각 노드들은 데이터를 생성 수집하고 2)부족한 특징을 확인한다. 3)클래스가 불균형한 Node 1은 부족한 특징을 충분히 가지고 있는 Node 2에게 데이터 송신 요청을 보내고 4)Node 2는 요청을 허가한다. 데이터 송신요청을 허락한 Node 2는 5-2)자신이 소유한 데이터에 대하여 주성분 분석 모델을 생성하고, 6)7)차원을 축소하여 차원 축소된 데이터셋을 Node 1에게 전송한다. 이때 5-1)데이터 송신요청을 한 Node 1도 자신이 가지고 있는 데이터에 대하여 주성분 분석 모델을 생성하며, 8)Node 2가 보낸 차원축소 데이터를 수신한 뒤 9)5-1에서 생성해두었던 모델을 활용하여 데이터를 재구성한다. 10)차원을 복구한 데이터는 Node 1이 기존에 가지고 있던 데이터와 병합되어 학습에 이용된다.

주성분 분석으로 압축한 데이터는 표준화된 특정 차원의 좌푯값이므로 해당 데이터셋을 공유하였을 때 원본 데이터셋의 프라이버시 침해 우려가 줄어든다. 또한 공격자가 데이터를 탈취하더라도 데이터 복구가 어렵기때문에 원본 데이터가 의미하는 바를 알 수 없으며, 차원이 압축된 데이터를 공유하기 때문에 원본 데이터셋을 공유하는 것에 비해 송수신하는 에너지와 메모리가 적게 소모될 것을 기대할 수 있다.

하지만 해당 메커니즘은 증명되어야 할 두 가지 요건이 있다. 첫 번째는 송신자의 주성분 분석 모델로 차원 축소된 데이터를 수신자의 주성분 분석 모델로 복구했을 때 성능이 유지되는지에 대한 검증이며, 두 번째는 최적의 Components 값을 찾는 것이다. 4장에서는 이 두 가지에 대하여 실험을 진행하고 성능 평가를 진행한다.

Ⅳ. 실험 및 결과 분석

본 장에서는 앞서 언급한 메커니즘을 실제로 구현하여 데이터 공유 전과 후의 성능을 평가한다. 프라이버시 관점에서의 평가를 위해 공격자 노드를 생성하고 중간자 공격(MITM, Man-In-The-Middle attack)을 통한 데이터 탈취를 가정한 실험을 추가로 진행한다. 3 장에서 언급한 두 가지 실험 진행을 위하여 Fig. 3와 같이 실험을 구성하였다.

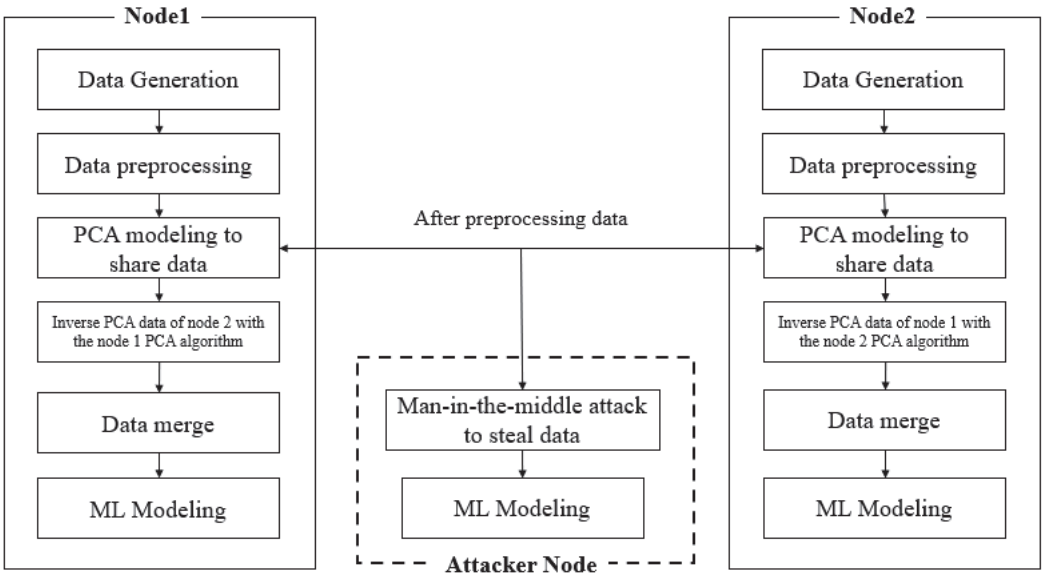


FIGURE 3. 실험 구조도

우선 데이터를 학습시키기에 적합한 데이터로 전처리를 진행한다. 이후 각 노드 별로 주성분 분석 모델을 생성하여 데이터의 차원을 축소하고 축소된 데이터를 공유한다. 이때 공격자가 데이터를 탈취하여 사용함을 가정하므로 공격자는 주성분 분석을 적용한 이후의 좌표 데이터를 사용하게 된다. 각 노드들은 본인들이 가진 주성분 분석 모델로 공유받은 데이터를 복구하며 본인의 데이터와 공유받은 데이터를 병합하여 기계학습을 통한 성능 평가를 진행한다.

1. 전제 조건

본 실험의 가정 사항은 다음과 같이 정의한다.

첫째, 정상자들은 서로 동일한 특징의 데이터를 수집한다.

둘째, 정상자들은 본인이 소유한 주성분 분석 모델을 이용하여 데이터를 축소하고 재구성하며, 모델은 공유하지 않는다.

셋째, 주성분 분석을 이용한 차원 축소 시, 수신 노드와 송신 노드는 Components 값을 고정하여 사용한다.

2. 실험 환경

실험은 Windows 10 Pro 운영체제에서 32GB RAM과 Intel i9 프로세서를 사용하고 Python 3.10버전으로 코딩하였다. 사용한 하드웨어 실험 환경은 TABLE 2에서 확인 가능하며 사용한 모듈 및 소프트웨어 환경에 대한 정보는 TABLE 3에서 확인할 수 있다.

TABLE II. 하드웨어 실험 환경

OS	Windows 10 Pro
CPU	Intel(R) Core(TM) i9-10850K CPU @ 3.60GHz
RAM	32.0 GB
DISK	4 TB

TABLE III. 소프트웨어 실험 환경

	name	Version
Tool	Anaconda	23.1.0
Python	Jupyter notebook	6.4.12
module	Python	3.10.11
	pandas	1.5.3
	numpy	1.22.4
	matplotlib	3.7.1
	tensorflow	2.12.0
	seaborn	0.12.2
	scikit-learn	1.2.2
	keras	2.12.0

사용하는 데이터셋은 UNSW-NB15 데이터셋으로 네트워크 침입 탐지를 목적으로 생성된 데이터셋이다[20, 21, 22, 23, 24]. 해당 데이터셋은 네트워크의 트래픽 정보를 담고 있으며, Analysis, Backdoor, DoS, Exploit, Fuzzers, Generic, Reconnaissance, Shellcode, Worm까지 총 9개의 공격 유형을 가지고 있고 45개의 특징으로 구성되어 있다.

주성분 분석 모델을 생성하기 전, 학습에 적합한 데이터를 구성하고자 데이터 전처리 과정을 진행한다. 본 연구에서의 전처리는 불필요한 특징 삭제와 라벨 인코딩, 스케일링, 그리고 데이터 재구성으로 이루어진다. 우선 불필요한 특징을 삭제하는 과정을 거친다. id 정보는 트래픽 유형 판단에 도움을 주지 않으므로 삭제하고, attack_cat 특징을 라벨로 사용할 것이기 때문에 공격인지 정상인지에 대한 정보를 담은 label 특징을 삭제한다. 라벨 인코딩은 사이킷런의 LabelEncoder 모듈을 이용하며, 수치 데이터가 아닌 proto, service, state 세 가지 특징을 학습에 용이하도록 인코딩한다. 또한 attack_cat의 Normal을 0, Generic을 1, Exploit을 2로 치환하여 라벨 인코딩을 진행하였다. 이후 각 특징 간의 값 조율을 위하여 사이킷런의 데이터셋 표준화 라이브러리인 StandardScaler를 사용해 데이터의 값을 조율한다.

다음으로 데이터를 목적에 맞게 재구성한다. 본 연구에서는 9개의 공격 유형 중 두 가지로 Exploit과 Generic을 선정하여 사용하며 정상 트래픽인 Normal을 포함한 세 가지 라벨을 사용하였다.

또한 데이터가 불균형한 상황을 가정하고 있으므로 Node 1부터 Node 8까지 데이터 구성의 비율(소수 클래스 : 다수 클래스)에 차등을 두었다. Node 1 Node2의 경우 데이터 비율은 1 : 20이며, Node 3 Node 4의 경우는 1 : 10으로 설정하였다, Node 5 Node 6는 1 : 5이며 Node 7 Node 8은 1 : 2의 비율로 설정하였다. 홀수 Node의 경우 Generic 라벨 데이터를, 짝수 Node의 경우 Exploit 라벨을 소수 클래스로 가정하였다. 자세한 Node 별 데이터 구성 및 비율은 TABLE 4, TABLE 5에서 확인 가능하다. 이때 학습 데이터의 경우 다수 클래스 데이터를 5000개에 맞추었고 소수 클래스 데이터를 비율에 맞게 맞추었다. 테스트 데이터셋의 경우는 각 라벨당 1500개씩 추출하여 각 노드가 동일한 개수의 데이터로 테스트를 진행한다.

TABLE IV. 훈련 데이터셋 구성

Node	label: 0	label: 1	label: 2	total
Node_1	5000	250	5000	10250
Node_2	5000	5000	250	10250
Node_3	5000	500	5000	10500
Node_4	5000	5000	500	10500
Node_5	5000	1000	5000	11000
Node_6	5000	5000	1000	11000
Node_7	5000	2500	5000	12500
Node_8	5000	5000	2500	12500

TABLE V. 테스트 데이터셋 구성

	label: 0	label: 1	label: 2
Test dataset	1500	1500	1500

이후 주성분 분석을 진행한다. 주성분 분석에는 사이킷런의 decomposition 라이브러리에 포함된 주성분 분석 모듈을 사용한다. 불균형 분포 별로 각각의 노드는 서로에게 부족한 데이터셋을 주성분 분석으로 차원 축소한 뒤 공유하고 복구 및 병합을 진행한다. 이때 Components를 비율로 지정하게 될 경우 결괏값으로 출력되는 특징 개수에 차등이 있으므로 Components 수는 전체 특징 수인 43보다 작은 정수값으로 지정한다. 차원 축소에 사용된 components 수가 같으면 다른 노드의 주성분 분석 모델로 학습한 차원 축소 데이터의 복구가 가능해진다. 다만 이렇게 다른 주성분 분석 모델로 차원을 축소한 데이터를 복구하였을 때 성능이 유지되는지에 대해서는 검증이 필요하다. 이에 대한 검증은 4장 3절에서 진행한다.

3. 실험 및 결과 분석

제안하는 방법은 압축할 때와 복구할 때 사용하는 주성분 분석 모델이 동일하지 않아도 데이터 공유 후 학습 시 성능이 유지되는지에 대한 선행 증명이 필요하다.

서로 다른 주성분 분석 모델로 차원 축소와 복구를 진행하기 위해서는 각 모델이 동일한 형태의 데이터셋을 사용해야 하며, Components 수가 고정되어야 한다. 이때 동일한 형태의 데이터셋이란 특징의 개수와 순서, 라벨이 같은 데이터셋을 의미한다. Components는 비율과 고정값으로 지정할 수 있는데 비율로 차원 축소를 진행하게 될 경우 데이터의 값에 따라 특징의 개수가 달라지므로 복구가 어렵다. 하지만 Components 수를 전체 특징 수보다 낮은 정수값으로 고정하게 될 경우 해당 Components 값으로 특징의 수가 고정되기 때문에 복구가 가능해진다.

본 절에서는 해당 메커니즘을 적용할 경우에 정확도가 유지되는지에 대해 증명한다. 앞서 언급한 환경과 동일한 환경에서 차원 축소 전의 불균형 노드 비율별 노드들과 각 노드에서 차원 축소를 진행한 뒤에 축소된 데이터를 서로 공유하고 복구한 경우에 대해 학습을 진행하고 그 결과를 확인한다.

TABLE 6은 데이터 공유 전과 후에 대해서 다중 분류 학습을 진행한 결과에 대한 표이다. 전반적으로 데이터 공유를 진행한 뒤에도 성능이 유지되는 것을 알 수 있으며, 이는 특정 조건이 만족될 경우 차원을 축소한 모델과 복구하는 모델이 같지 않더라도 데이터 복구가 가능함을 의미할 뿐만 아니라 데이터의 공유로 인해 정확도가 소폭 향상하였음을 알 수 있다.

TABLE VI. 데이터 공유 메커니즘 파일럿 테스트 결과

Accuracy (Unit: %)		Node 1	Node 2	Inverse Node 1	Inverse Node 2
	1:20	98.04	93.86	98.06	93.82
		Node 3	Node 4	Inverse Node 3	Inverse Node 4
	1:10	98.48	96.97	98.53	97.02
		Node 5	Node 6	Inverse Node 5	Inverse Node 6
	1:5	98.11	97.28	98.15	97.41
		Node 7	Node 8	Inverse Node 7	Inverse Node 8
	1:2	98.06	97.82	98.04	97.82

이후에는 Components에 따른 공격자와 정상자의 성능을 평가한다. 실험은 3장에 서술된 대로 진행하였으며 데이터 병합을 마친 상태에서의 정상자 노드와 차원 축소된 데이터를 탈취한 공격자 노드를 나누어 평가한다. 평가지표는 성능을 평가하기 위한 정확도(accuracy), 정밀도(precision), 재현율(recall)로 두었고, 데이터의 송수신 메모리 사용량 또한 평가 요소로 두었다.

마지막으로 연구의 목적인 데이터의 프라이버시 보호 측면에서 평가하기 위하여 프라이버시 척도를 정의하고 해당 관점에서 결과를 평가한다.

TABLE 7는 각 비율별 정확도를 의미한다. 45개의 특징 중에서 두 개의 특징을 삭제하였으므로 최대 Components 값은 43이다. 단 Components 값을 43부터 1까지 줄여가며 기계학습 모델 성능의 변화를 확인하였을 때 20까지는 성능이 유지되는 모습을 볼 수 있었으므로 가시성을 위해 Components 값을 20, 15, 10, 5, 3, 2, 1일 때의 성능을 TABLE 7과 Fig. 4부터 Fig. 7까지로 정리하였다.

TABLE VII. Components 값에 따른 비율별 정확도

Accuracy(%)	Components						
Node Name	20	15	10	5	3	2	1
Normal node 1	98.02	98.02	98.04	98.08	98.06	98.15	98.06
Normal node 2	94.08	94.15	94.33	93.82	94.11	94.11	94.266
Inverse Node 1	95.31	94.43	94.76	93.85	89.43	84.87	74.47
Inverse Node 2	97.82	98.27	97.85	97.98	96.91	94.76	90.00
Normal node 3	98.46	98.51	98.51	98.51	98.57	98.53	98.51
Normal node 4	97.26	97.44	97.40	97.15	97.31	97.00	97.13
Inverse Node 3	96.38	95.30	94.06	91.93	89.46	82.38	72.44
Inverse Node 4	97.6	97.30	96.63	96.22	95.33	92.66	85.71
Normal node 5	98.06	98.02	98.08	98.17	97.86	97.95	98.11
Normal node 6	97.44	97.57	97.40	97.60	97.44	97.53	97.40
Inverse Node 5	95.27	95.03	94.60	92.30	89.06	95.51	69.45
Inverse Node 6	96.72	96.39	96.39	94.27	93.12	90.30	82.24
Normal node 7	97.97	97.97	97.84	97.91	97.80	97.62	97.97
Normal node 8	97.26	97.44	97.40	97.15	97.31	97.00	97.13
Inverse Node 7	96.38	95.30	94.06	91.93	89.46	82.38	72.44
Inverse Node 8	97.65	97.30	96.63	96.22	95.33	92.66	85.71

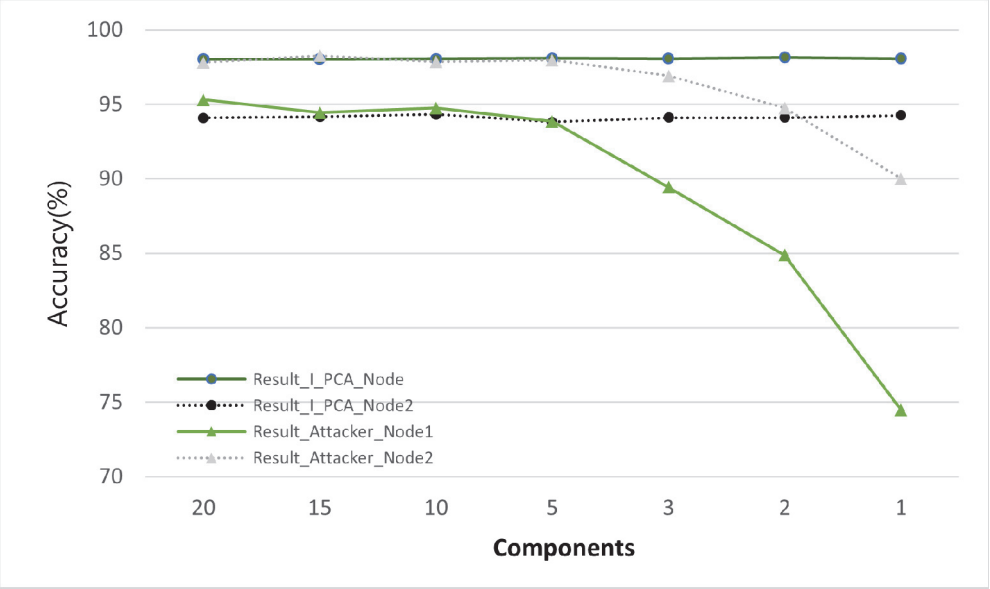


FIGURE 4. 1 : 20 비율에서 Components 값에 따른 노드별 정확도

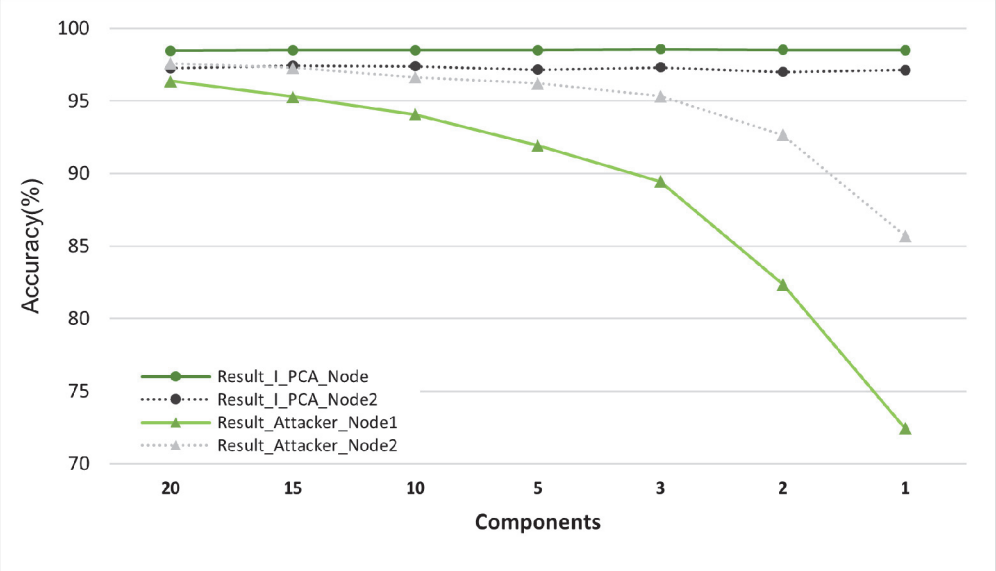


FIGURE 5. 1 : 10 비율에서 Components 값에 따른 노드별 정확도

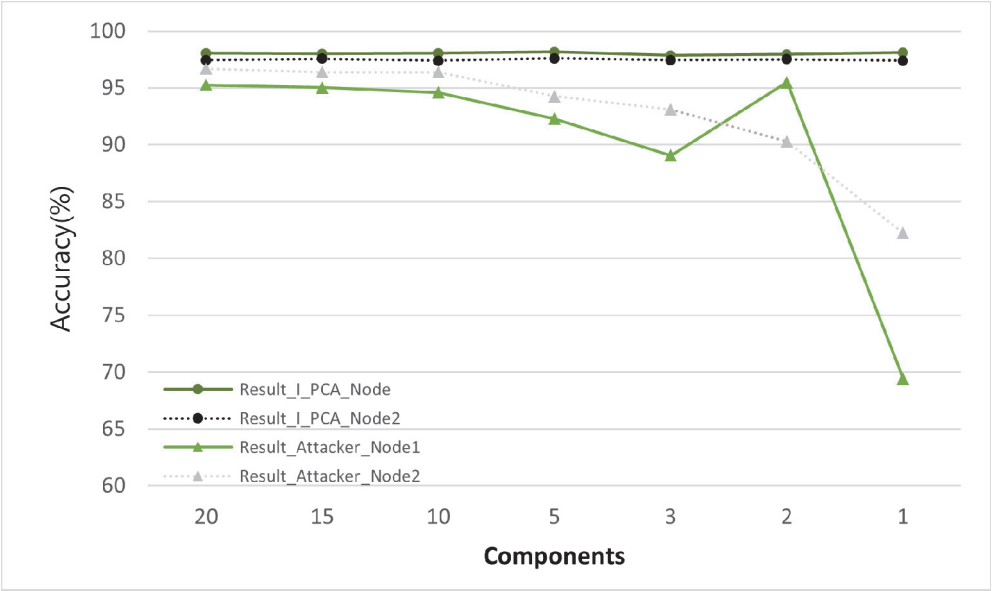


FIGURE 6. 1 : 5 비율에서 Components 값에 따른 노드별 정확도

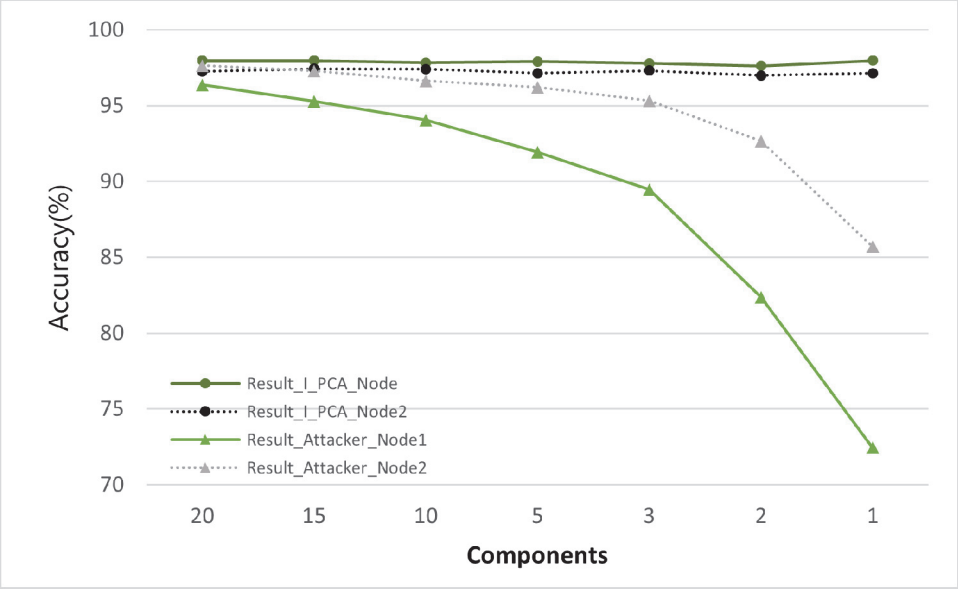


FIGURE 7. 1 : 2 비율에서 Components 값에 따른 노드별 정확도

정확도, 정밀도와 재현율, F1 점수를 구하기 위해 사용된 공식은 수식 (8), (9), (10), (11)과 같고 수식에서 사용된 약어에 대한 풀이는 TABLE. 8에서 확인 가능하다.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F1-score = 2\times\frac{\mathrm{Pr}ecision\times Recall}{\mathrm{Pr}ecision+Recall} \quad (11)$$

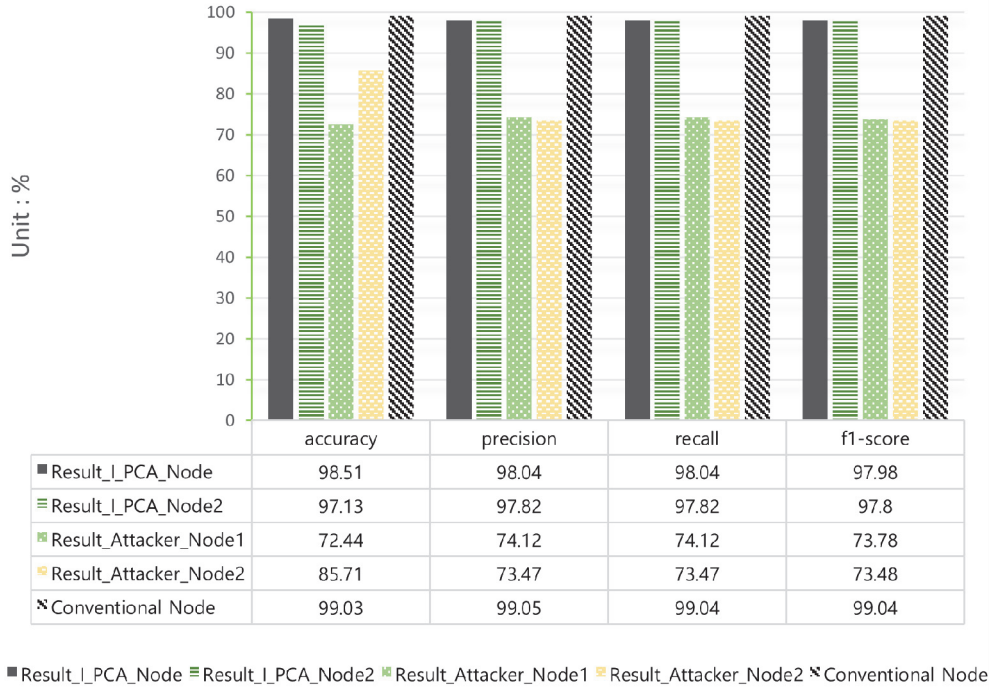
TABLE VIII. 혼돈 행렬 및 공식에 대한 약어 정리

예측\실제	음성	양성
음성	TN(True Negative, 진음성)	FN(False Negative, 위음성)
양성	FP(False Positive, 위양성)	TP(True Positive, 진양성)

결과를 보면 모든 비율에 대하여 정상자는 정확도를 유지하였고, 공격자는 components 수가 감소함에 따라 정확도도 하락하는 추세를 보였다. 이에 따라 최적의 Components 값을 1로 고정하며 성능에 큰 차이가 없으므로 데이터 편향 비율을 1:10 비율로 고정하도록 한다.

Fig. 5에서는 1 : 10 비율에서의 Components 값에 따른 정확도이다. 종래 모델과 비교하였을 때 정확도가 Components 값이 떨어짐에 따라 전반적인

성능이 하락할 것이라 예측할 수 있음에도 정상 노드의 성능은 높은 상태를 유지하고, 공격자의 성능은 Components 값이 떨어짐에 따라 30% 가량 하락하는 것을 알 수 있다. 이에 따라 Fig 5만 보았을 때, 가장 최적화된 Components 값은 1임을 알 수 있다.



	accuracy	precision	recall	f1-score
Result_I_PCA_Node	98.51	98.04	98.04	97.98
Result_I_PCA_Node2	97.13	97.82	97.82	97.8
Result_Attacker_Node1	72.44	74.12	74.12	73.78
Result_Attacker_Node2	85.71	73.47	73.47	73.48
Conventional Node	99.03	99.05	99.04	99.04

FIGURE 8. 1 : 10 비율에서의 Components 값이 1일때의 정상자 및 공격자 노드의 정확도, 정밀도, 재현율, f1 점수

1 : 10 비율에서 Components 1일 때의 공격자 및 정상자 노드의 정확도, 정밀도, 재현율, f1 점수에 따른 결과는 Fig. 8과 같다. 전반적으로 네 가지 지표에 대해서 정상자가 공격자보다 좋은 성능을 보이고 있음을 알 수 있다.

Fig. 9는 1 : 10 비율에서 Components 값에 따른 종래 모델과 제안하는 모델의 메모리 사용량에 대한 비교 그래프이다. 종래에는 방대한 데이터를 그대로 전송하기 때문에 메모리 사용량 측면에서 한계가 있으나 제안하는 모델은 데이터셋의 차원을 1로 축소함으로써 데이터 송수신 시 사용되는 메모리가 줄어들 것을 기대할 수 있다. 앞선 결과에 의하면 가장 최적의 Components 값은 1이므로 Components 값이 1일 때 제안하는 모델은 종래 모델 대비 42배 메모리 효율적인 모델임을 증명하였다.

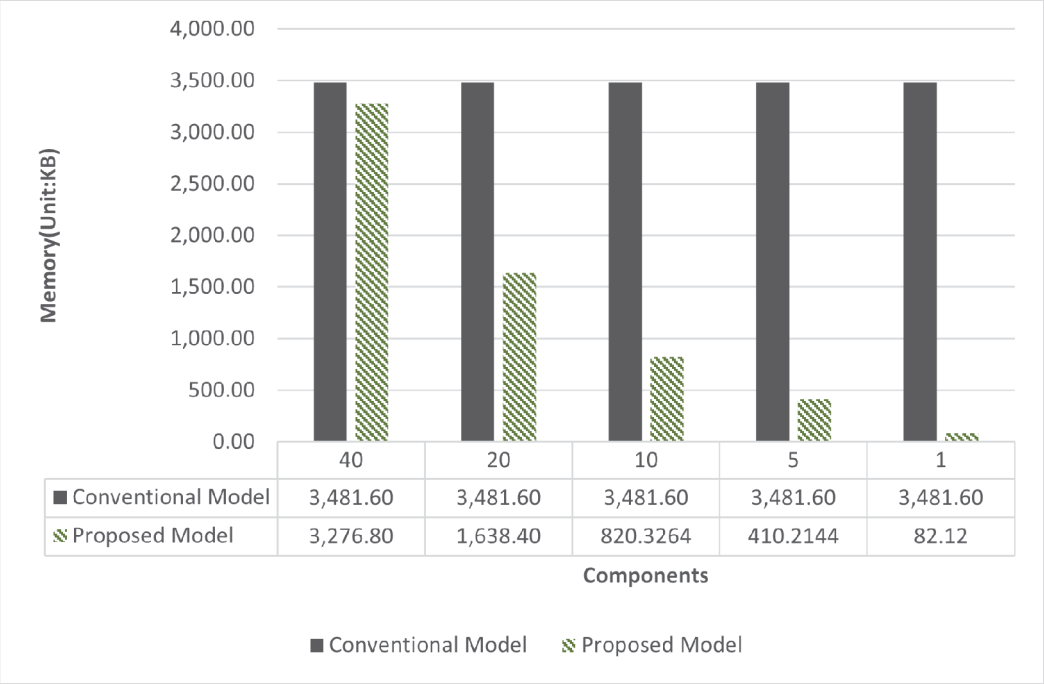


FIGURE 9. 1 : 10 비율에서의 Components 값에 따른 메모리 사용량

마지막으로 1 : 10 비율에서 데이터가 프라이버시 측면에서 안전하게 전처리 되었는지 확인하기 위한 방법으로 프라이버시 척도를 정의하고 평가한다. 프라이버시 척도는 수식 9와 같이 정의하였고 이는 정상 노드의 정확도와 공격자 노드의 정확도의 차이를 의미한다. 정상자의 정확도가 유지되고 공격자의 정확도가 떨어진다면 이는 프라이버시 측면에서 보호가 되었음을 의미한다. 따라서 프라이버시 척도가 높을수록 안전한 데이터 공유 환경임을 의미한다. 이에 따른 Components 별 정확도 추이는 Fig. 10과 같다. 그림에서 알 수 있듯 Components가 줄어들수록 프라이버시 척도는 향상됨을 알 수 있다.

$$Privacy\ Capability = (Normal\ Accuracy - Attacker\ Accuracy) \quad (9)$$

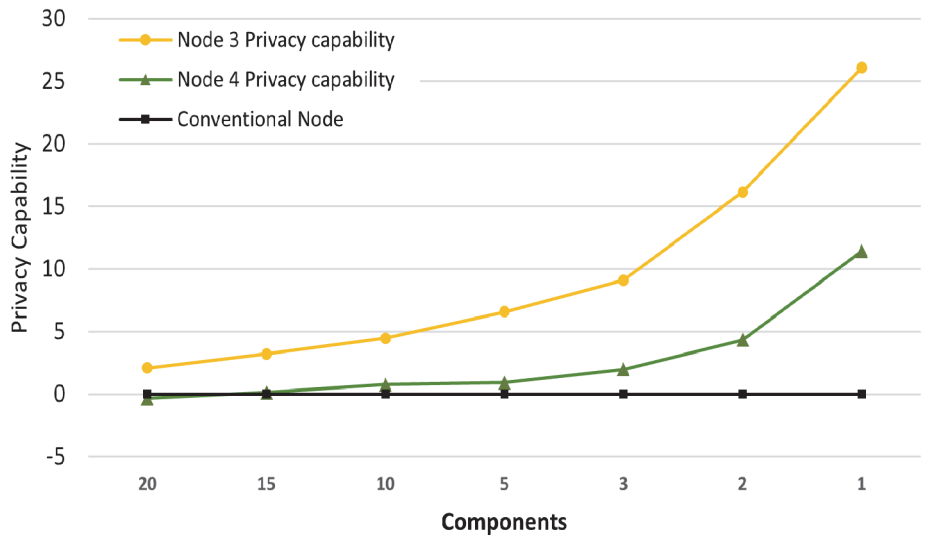


FIGURE 10. 1 : 10 비율에서의 Components 값에 따른 프라이버시 척도

따라서 제안하는 데이터 공유 메커니즘은 Components 값에 관계없이 정상자의 정확도는 유지되면서 공격자의 정확도는 떨어뜨릴 수 있으며, Components 값이 가장 낮은 값인 1일 때 가장 프라이버시 척도가 높고 비용 효율적임을 증명하였다.

V. 결 론

클래스 불균형한 환경에서 데이터 샘플링과 안전한 데이터 공유를 위한 종래 연구들에 있어 한계점을 보완하기 위해 안전한 데이터 공유 환경 구축의 필요성이 대두되고 있다. 또한 인공지능 기술을 사용하는 기기는 자원이 한정되어 있으므로 데이터 송수신 시 메모리 및 에너지 효율 또한 중요한 척도로써 평가되어야 한다. 이에 따라 본 연구는 보다 프라이버시 측면에서 안전하고 메모리 효율적인 새로운 데이터 공유 메커니즘을 제안한다.

제안하는 메커니즘은 주성분 분석의 차원 축소와 복구 기능을 활용한 방법이며 데이터를 공유할 때 각 노드에서 주성분 분석 모델을 생성한 뒤 주성분 분석으로 압축한 데이터를 공유하고 수신한 차원 축소 데이터를 본인의 주성분 분석 모델로 복구하는 방식이다. 이때 공유되는 데이터는 차원을 축소한 데이터이며, 중간자가 탈취하더라도 프라이버시 측면에서 안전해야 한다. 따라서 공격자가 데이터를 탈취한 상황을 가정하여 제안하는 메커니즘이 안전한지 평가하였다. 해당 메커니즘을 평가하기 위하여 정확도, 정밀도, 재현율 이외에도 데이터 송수신 시 메모리 사용량 및 프라이버시 척도 등의 평가지표로 평가를 진행하였다. 실험 결과를 통하여 제안하는 메커니즘이 정상자 관점에서는 성능이 유지되었으나 공격자의 관점에서는 성능이 하락함을 보였다. 또한 최적의 Components 값에서 제안하는 메커니즘은 일반적인 공유 방식보다 42배 메모리 효율적이고 프라이버시 척도가 가장 높았다.

해당 연구는 각각의 노드가 일회성으로 데이터를 공유한 결과에 대해 다

루었기 때문에 향후 연구로 데이터 공유 횟수에 따른 성능의 변화를 확인할 수 있으며 다양한 데이터셋에 대하여 적용해볼 예정이다.

참 고 문 헌

[1] Kankanhalli, A., Charalabidis, Y., & Mellouli, S., “IoT and AI for smart government: A research agenda.” Government Information Quarterly, 36(2), 304-309. 2019.

[2] Allam, Z., & Dhunny, Z. A., “On big data, artificial intelligence and smart cities.” Cities, 89, 80-91. 2019.

[3] Ahmad, T., Zhang, D., Huang, C., Zhang, H., Dai, N., Song, Y., & Chen, H., “Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities.” Journal of Cleaner Production, 289, 125834. 2021.

[4] Dlamini, Z., Francies, F. Z., Hull, R., & Marima, R., “Artificial intelligence (AI) and big data in cancer and precision oncology.” Computational and structural biotechnology journal, 18, 2300-2311. 2020.

[5] Betty Jane, J., & Ganesh, E. N., “Big data and internet of things for smart data analytics using machine learning techniques.” In Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI-2019) Springer International Publishing. pp. 213-223. 2020.

[6] Kankanhalli, A., Charalabidis, Y., & Mellouli, S., “IoT and AI for smart government: A research agenda. Government Information Quarterly,” 36(2), 304-309. 2019.

[7] Ma, X., Zhu, J., Lin, Z., Chen, S., & Qin, Y., “A state-of-the-art survey on solving non-IID data in Federated Learning.” Future Generation Computer Systems, 135, 244-258. 2022.

[8] Criado, M. F., Casado, F. E., Iglesias, R., Regueiro, C. V., & Barro, S., “Non-IID data and Continual Learning processes in Federated Learning: A long road ahead..” Information Fusion, 88, 263-280. 2022.

[9] Al-Turjman, F., Zahmatkesh, H., & Shahroze, R., “An overview of security and privacy in smart cities' IoT communications.” Transactions on Emerging Telecommunications Technologies, 33(3), e3677. 2022.

[10] Ioannidou, I., & Sklavos, N., “On general data protection regulation vulnerabilities and privacy issues, for wearable devices and fitness tracking applications.” Cryptography, 5(4), 29. 2021.

[11] Zheng, X., & Cai, Z., “Privacy-preserved data sharing towards multiple parties in industrial IoTs.” IEEE Journal on Selected Areas in Communications, 38(5), 968-979. 2020.

[12] Maćkiewicz, A., & Ratajczak, W.,. “Principal components analysis (PCA).” Computers & Geosciences, 19(3), 303-342. 1993.

[13] Khalid, S., Khalil, T., & Nasreen, S., “A survey of feature selection and feature extraction techniques in machine learning.” In 2014 science and information conference. IEEE pp. 372-378. 2014,

[14] Jung Hee Cheon, Yunhee Euh, & Jae-yun Kim. “Privacy-Preserving Finance Data Analysis Based on Homomorphic Encryption.” Review of Financial Information Studies, 7(1), 33-60. 2018.

[15] Lu, Y., Huang, X., Dai, Y., Maharjan, S., & Zhang, Y., “Federated learning for data privacy preservation in vehicular cyber-physical systems.” IEEE Network, 34(3), 50-56. 2020.

[16] Fugkeaw, S., A secure and efficient data sharing scheme with outsourced signcryption and decryption in mobile cloud computing. In 2021 IEEE International Conference on Joint Cloud Computing (JCC) pp. 72-79. 2021.

[17] Singh, A. K., & Saxena, D., “A cryptography and machine learning based authentication for secure data-sharing in federated cloud services environment.” Journal of Applied Security Research, 17(3), 385-412. 2022.

[18] Athanere, S., & Thakur, R., “Blockchain based hierarchical semi-decentralized approach using IPFS for secure and efficient data sharing.” Journal of King Saud University-Computer and Information Sciences, 34(4), 1523-1534. 2022.

[19] Kumar, R., Kumar, P., Tripathi, R., Gupta, G. P., Islam, A. N., & Shorfuzzaman, M., “Permissioned blockchain and deep learning for secure and efficient data sharing in industrial healthcare systems.” IEEE Transactions on Industrial Informatics, 18(11), 8065-8073. 2022.

[20] Moustafa, N., & Slay, J., “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set).” In 2015 military communications and information systems conference (MilCIS) pp. 1-6. 2015.

[21] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[22] Moustafa, N., & Slay, J., "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset." Information Security Journal: A Global Perspective 1-14. 2016.

[23] Moustafa, N., Slay, J., & Creech, G., “Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks.” IEEE Transactions on Big Data, 5(4), 481-494. 2017.

[24] Moustafa, Nour, et al. "Big data analytics for intrusion detection system: statistical decision-making using finite dirichlet mixture models." Data Analytics and Decision Support for Cybersecurity. Springer, Cham, 127-156. 2017.

ABSTRACT

Secure Data Sharing Mechanism based in Principal Component Analysis

Na-Yeon Shin

Department of Future Convergence Technology Engineering
Graduate School of Sungshin University

As the usage of mobile devices increases, the amount of data generated is exploding. Recently, various values have been created by utilizing the vast amount of data generated. However, in the case of data collected from an individual's terminal, the information collected varies depending on the environment, device, and user characteristics, which can lead to a non-IID problem. Data sharing is a typical solution to non-IID problems, but sending original data can cause privacy issues for users and could lead to third parties stealing the data. Therefore, this study aims to solve problems by proposing a secure data-sharing mechanism using principal component analysis. Each node creates a principal component analysis model, then reduces and shares the data, and recovers the model using the principal component analysis model created before sharing. This study evaluated the proposed mechanism through experiments from the

perspective of normal nodes and attackers. At this time, normal nodes maintained constant accuracy before and after data sharing, but the accuracy of the attacker node decreased. In addition, optimal Components values were derived, and up to 42 times memory efficiency was proved. We also demonstrate the performance of the proposed mechanism by showing the largest privacy scale at optimal Components values.

ACKNOWLEDGEMENTS

본 논문을 지도해주신 이일구 교수님과 연구에 기여해 준 이연지 학생에게 감사드립니다. 또한 논문 검토에 참여해 준 길예슬 김소연 박나은 오예솔 학생에게도 감사를 표합니다.