Gemini 모델의 문맥 추론을 이용한 K-익명성 데이터의 재식별 위험성 연구

조찬영, *이선영
순천향대학교

{ccy0531, *sunlee}@sch.ac.kr

Re-Identification of K-Anonymized Data Using the Gemini Model's Operational Line

Chan Yeong Cho, Sun-Young Lee*
Dept. of Information Security Engineering Soonchunhyang Univ

요 약

본 논문은 파이썬 라이브러리를 활용하여 가상의 데이터셋을 제작하고 LLM 의 문맥 추론 능력이 기존 K-익명성 비식별화 체계를 무력화할 수 있음을 실증하고 이에 대한 방어책을 제안하였다. Gemini 1.5 Pro 모델을 이용한 공격 시뮬레이션 결과, K=2 익명 데이터셋에서 2.8%의 재식별이 발생하여 통계적 방어의 한계를 확인하였다. 이에 민감 정보를 상위 개념으로 추상화하는 문맥 인식적 교란 기법을 제안하였으며, 이를 적용한 결과 재식별 성공률이 0%로 감소하여 그 실효성을 검증하였다. 결론적으로 본 연구는 AI 시대에 통계적 익명성을 넘어 의미론적 연결을 차단하는 새로운 프라이버시 보호 전략의 필요성을 입증하였다.

Ⅰ. 서 론

4 차 산업혁명의 가속화와 인공지능의 발전과 함께 데이터는 의료, 금융 등 다양한 산업 분야에서 다양하게 활용되고 있다[1]. 특히, 인공지능, 클라우드 컴퓨팅 등 디지털 기술의 비약적인 발전은 대규모 데이터의 수집과 처리를 가능하게 했다[1]. 국내의 경우 2020 년 데이터 3 법 개정을 통해 가명정보 개념이 도입되어 산업에서의 데이터 결합과 공유가 활발해지고 있다[2].

데이터 활용의 전제 조건인 프라이버시 보호를 위해 K-익명성과 같은 비식별화 모델이 발전되어 왔으며, 국내의 가명정보 처리 가이드라인 등을 포함한 대다수의 규제 컴플라이언스에서 K-익명성을 기반으로 식별자 삭제, 범주화, 마스킹 등의 조치를 취함으로써 프라이버시 안전성을 확보하고 있다[3].

그러나 최근 Gemini 와 같은 LLM 의 등장은 이러한 전통적 비식별화 체계에 근본적인 위험을 제기하고 있다. LLM 은 방대한 사전 학습 지식과 고도화된 문맥 추론 능력을 바탕으로 비정형 텍스트와 파편화된 정보를 논리적으로 결합하는 능력을 보여준다. LLM 은 단순한 키워드 매칭을 넘어, 텍스트에 내포된 의미론적 단서를 포착하여 통계적으로 익명화된 집단 내에서 특정 개인을 소거하거나 유일하게 특정할 수 있다[4].

본 논문에서는 이러한 기술적 환경에 주목하여, 임의로 제작한 가상의 데이터셋을 활용하여 LLM 을 프라이버시 위협의 주체이자 공격 도구로 정의하고, 현행 비식별화 기법의 취약성을 실증적으로 검증한다. 또한, LLM 의 공격에 대응할 수 있는 새로운 방어 기법인 문맥 인식 의미론적 교란 기법을 제안한다.

Ⅱ. 관련 연구

2.1 K-익명성

정보보호 분야에서 데이터 유용성과 프라이버시 간의 균형을 맞추기 위해 다양한 비식별화 모델이 제안되어 왔다. 그 중 K-익명성은 Sweeny 에 의해 2002 년 제안된 가장 대표적인 프라이버시 보호 모델로, 공개된 데이터셋 내에서 특정 개인을 식별할 수 없도록 하는 것을 목표로 한다. K-익명성은 동일한 준식별자를 가진 레코드가 최소 K 개 이상 존재하도록 하여 식별 가능성을 1/K 로 낮춘다, 그러나 이는 민감 정보의 동질성 공격이나 배경지식에 의한 공격에 취약하다는 한계가 지속적으로 지적되어 왔다[3].

2.2 LLM 기반의 추론 공격과 프라이버시 위협

LLM 의 등장은 데이터 프라이버시 환경에 새로운 차원의 위협을 제기하고 있다. 기존 프라이버시 공격이 주로 데이터베이스 간의 정확한 필드 매칭에 의존했던 연결 공격이었던 반면, LLM 기반의 공격은 문맥과 의미를 파악하는 추론 공격으로 진화하였다[4].

LLM 의 문맥 추론 능력은 기존 K-익명성이 전제하는 제한된 공격자 모델을 무력화시킨다. LLM 의 의미론적 연결은 기존의 일반화 기법으로는 차단하기 어려운 새로운 유형의 취약점으로 떠오르고 있다[4].

Ⅲ. 연구 방법

3.1 시나리오 구성

본 논문에서는 LLM 의 재식별 위협과 방어 실효성을 검증하기 위해 두 가지 시나리오를 설계하였다. 시나리오 A 는 K=2 익명 데이터에 대해 LLM 이 외부 지식과 문맥 추론을 통해 개인을 특정할 수 있는 지 분석하여 기존 기법의 취약점을 실증한다. 시나리오 B 는 민감정보를 일반화하는 의미론적 교란 기법 적용 시 재식별 성공률이 하락하는지 평가한다. 이를 통해 제안하는 기법이 데이터 유용성을 유지하면서 LLM 의 추론 연결 고리를 효과적으로 차단함을 정량적으로 입증한다.

3.2 데이터셋 구성

실제 개인정보 유출 위험을 배제하기 위해 Python Faker 라이브러리를 활용하여 한국인 특성을 반영한 5,000 명의 가상의 의료 데이터셋을 생성하였다[5]. 해당 데이터셋은 성명, 나이, 성별, 주소, 직업, 질병 필드로 구성되며, 연구 시나리오에 맞춰 K=2 수준의 비식별화를 적용하였다.

3.3 위협 모델 및 외부 지식 생성

전체 데이터의 20%에 해당하는 외부 보조 지식을 생성하였다. 외부 지식은 SNS, 전문직 프로필, 커뮤니티 게시글 등의 유형의 자연어 텍스트로 구성되며, LLM 의 독해 능력을 검증하기 위해 질병 정보를 혈당 관리(당뇨), 수면유도제(불면증) 등과 같이 문맥적으로 변형하여 주입하였다.

3.4 공격 모델 설계

공격 도구로는 긴 문맥 처리에 강점이 있는 Google 의 Gemini 1.5 Pro 를 선정하였다[4]. 공격 프롬프트는 비식별 레코드와 외부 지식을 입력받아 준식별자의 일치 여부를 필터링하고, 민감 정보와 외부 텍스트 간의 의미적 유사성을 분석하여 유일한 후보를 특정하도록 설계 하였다.

Ⅳ. 실험 및 결과

4.1 실험 환경

총 1,000 개의 무작위 타겟을 선정하여 공격 시뮬레이션을 수행하였다. 공격 성공의 기준은 모델이 제시한 최우선 후보가 실제 원본 데이터의 주인과 일치하며, 동시에 해당 후보가 유일하게 특정된 경우로 정의하였다.

4.2 공격 결과 분석

실험 결과, K=2 익명성이 적용되었음에도 불구하고 1,000 회 중 28 회, 즉 2.8%의 성공률로 재식별이 발생하였다. 성공한 케이스를 분석한 결과 Gemini 는 단순한 키워드 매칭을 넘어 복합적인 문맥을 비식별 데이터의 속성과 정확히 연결하였다. 이는 K-익명성이 보장하는 통계적 방어막이 의미론적 추론에 의해 무력화되었음을 시사한다.

구분	결과
총 공격 시도 횟수	1,000 회
재식별 성공	28 건
재식별 실패	972 경
공격 성공률	2.80%

<표 1> 공격 수행 결과

Ⅴ. 문맥 인식 의미론적 교란

본 논문에서는 LLM 공격의 핵심인 구체적인 민감 정보의 연결성을 차단하기 위해 민감 정보를 상위 개념으로 추상화하거나, 문맥적 모호성을 주입하는 의미론적 교란 기법을 제안한다. 예를 들어, '초기 당뇨'라는 질병을 '대사 관리군'으로 변환하는 방식으로 데이터의 사실성을 유지하면서 LLM 의 추론 확신도를 낮추는 적대적 방어 전략이다.

제안하는 기법 적용 후 데이터셋에 동일한 공격을 수행한 결과, 재식별 성공률은 2.8%에서 0%로 감소하였음을 확인하였다. 이는 문맥 인식 의미론적 교란이 LLM 의 추론 능력을 효과적으로 방어했음을 입증한다.

Ⅵ. 결론

본 논문에서는 LLM 기술의 발전이 데이터 프라이버시에 미치는 영향을 실증적으로 분석하였다. 실험을 통해 K 익명성과 같은 전통적 비식별화 기법이 LLM 의 고도화된 추론 능력 앞에서는 취약할 수 있음을 확인하였으며, 이에 대한 대안으로 문맥 인식 의미론적 교란 기법을 제안하였다.

제안하는 기법은 민감 정보의 구체적인 의미를 상위 개념으로 추상화하거나 문맥적 모호성을 주입하여, LLM 이 외부 지식과 비식별 데이터를 논리적으로 연결하지 못하도록 차단하는 것을 목표로 한다. 시뮬레이션 결과, 해당 기법을 적용했을 때 LLM 기반 재식별 공격 성공률은 기준 2.8%에서 0%로 완전히 제거되었음을 확인하였다. 이를 통해 본 연구에서는 AI 시대의 데이터 활용에 있어, 단순히 통계적 익명성을 달성하는 것을 넘어 데이터 간의 의미론적 연결 가능성을 제어하는 새로운 차원의 프라이버시 보호 전략이 필수적임을 입증하였다..

ACKNOWLEDGMENT

본 연구는 정부(과학기술통신부)의 재원으로 한국연구재단의 지원을 받아 수행한 연구임(NO. RS-2024-00346749).

참 고 문 헌

[1]세계경제포럼. (2016). The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution. World Economic Forum.

[2]한국인터넷진흥원. (2020). 데이터 3 법 개정에 따른 개인정보보호 제도 개선 안내서. 한국인터넷진흥원.

[3]Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.

[4]Konzelmann, J., & Gworek, W. (2024, February 14). Gemini 1.5: Our next-generation model, now available for private preview in Google AI Studio. Google Developers Blog. https://developers.googleblog.com/en/gemini-15-our-next-generation-model-now-available-for-privatepreview-in-google-ai-studio/

[5]Curella, G., et al. (2025). Faker (Version 37.12.0) [Computer software]. Python Package Index. https://pypi.org/project/Faker/