프라이버시 보호를 위한 지도학습 기반 Self-Destructing Scheme 연구

김영기O, 홍충선*
경희대학교 컴퓨터공학과
{qoo0144, cshong}@khu.ac.kr

A Study on Self-destructing Scheme Based on Supervised Learning for Privacy Protection

Dongkyu LeeO, ChoongSeon Hong*
Department of Computer Science and Engineering, Kyung Hee University

요 약

클라우드 컴퓨팅을 활용한 기술 및 서비스가 발전함에 따라 이를 사용하는 기업 또는 개인 사용자들이 점차 늘어나고 있다. 그러나 최근들어 클라우드 스토리지에 개인정보를 저장하고 활용하는 사용자들이 증가하면서 클라우드 환경에서의 프라이버시 보호 모델에 대한 연구가 더욱 중요시 되고 있다. 이를 위한 방법으로 분산 해시 테이블 네트워크를 이용하여 일정 시간이 지나면 암호화된 사용자의 데이터를 복호화할 수 없도록 하는 Self-Destructing Scheme이 제안되었다. 그러나 기존의 프라이버시 보호 모델에서는 데이터의 가용성과 보안성을 고려하여 임계값을 설정하는 방법에 대해서는 언급하고 있지 않다. 따라서 본 논문에서는 지도학습의 한 방법인 회귀분석을 적용하여 프라이버시 보호 모델의 데이터 가용성과 보안성을 모두 고려한 최적의 임계값 찾는 방법을 제안한다.

1. 서 론

클라우드 컴퓨팅을 활용한 기술 및 서비스가 발전함에 따라 클라우드 환경에서의 프라이버시 보호 모델에 대한 연구가 더욱 중요시 되고 있다.

이를 위한 방법으로 Shamir Secret Sharing[1]을 활용하여 암호화에 필요한 키를 여러 조각으로 나누고 이를 분산 해시 테이블 네트워크(Distributed Hash Table Network)를 이용하여 일정 시간이 지나면 암호화된 사용자의 데이터를 복호화할 수 없도록 하는 Self-Destructing Scheme이 제안되었다[2].

그러나 기존의 Self-Destructing Scheme에서는 암호화된 데이터에 대한 가용성과 보안성을 고려하는 임계값을 설정하기 위한 방법에 대해서는 언급하고 있지 않다. 따라서 본 논문에서는 지도학습의 한 방법인 회귀분석을 적용하여 Self-Destructing Scheme에서 데이터의 가용성과 보안성을 고려한 최적의 임계값을 찾는 방법을 제안한다.

2장에서는 관련연구로 기존의 Self-Destructing Scheme과 회귀분석에 대해 언급하고, 3장에서는 기존 연구의 문제점 및 제안사항에 대해 기술한다. 4장에서는 실험결과 및 성능평가를 분석하고, 마지막으로 5장에서는 본 논문의 결론 및 향후 연구계획에 대해 언급한다.

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (2015-0-00274,(ICBMS-2세부) ICBMS 플랫폼 간 정보 모델 연동 및 서비스 매쉬업을 위한 스마트 중재 기술 개발). *Dr. CS Hong is the corresponding author

2. 관련 연구

2.1 Self-Destructing Scheme

Self-Destructing Scheme은 2009년 Geambasu 등이 제안한 프라이버시를 보호하기 위한 모델이며 시스템의 구조는 그림 1과 같다.

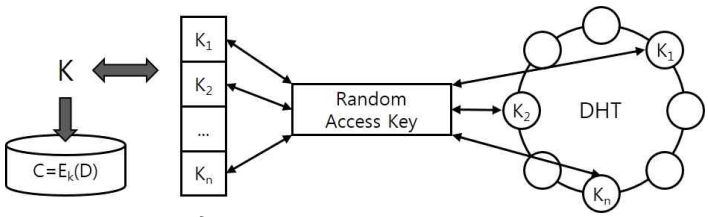

그림 1. Self-Destructing Scheme

Self-Destructing Scheme에서는 데이터를 암호화 하는데 사용되는 키를 여러 개의 조각으로 나누어 특정 기간이 지나면 데이터가 사라지는 분산 해시 테이블 네트워크에 분배하는 방식이다.

2.2 회귀분석

회귀분석은 지도학습의 한 방법으로 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구하고 적합도를 측정하는 분석 방법이다. 본 논문에서는 Self-Destructing Scheme에서 키를 여러 조각으로 나누어 분배할 때, 전체 키 조각의 개수와 복호화 하는데 필요한 최소한의 키 조각의 개수의 관계에서 데이터의 가용성과 보안성을 고려하여 최적의 임계값을 찾는 방법을 제안하고자 한다.

3. 기존 연구의 문제점 및 제안사항

[2]에 따르면 나누어진 키 조각의 전체 개수와 복호화 하기 위해 필요한 최소한의 개수는 데이터의 가용성 및 보안성과 관련되어있다.

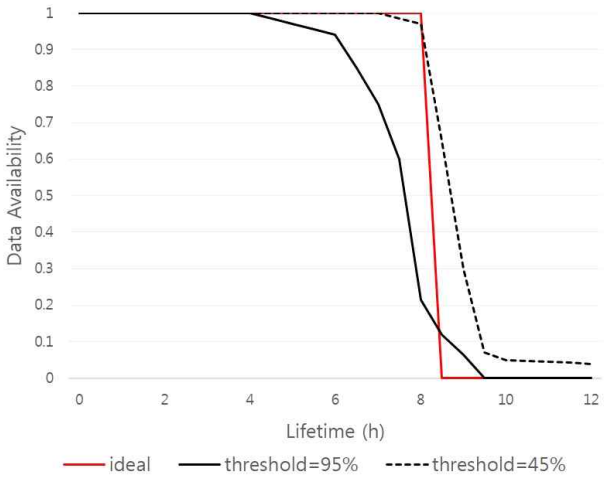

그림 2. Self-Destructing Scheme에서 데이터의 가용성

그림 2는 기존의 기법에서 N개의 같은 키 조각과 이를 복호화 하기 위해 필요한 임계값이 다른 두 그래프를 나타낸다. 임계값의 비율이 95%인 경우에는 데이터가 유지되어야 하는 시간을 충족하지 못했으며, 45%인 경우에는 일정 기간이 지난 뒤에도 키의 조각이 완전히 사라지지 않았으므로 보안상의 문제를 야기할 수 있다. 이러한 문제는 분산 해시 테이블 네트워크의 특수성 때문이며 이를 해결하기 위한 연구 또한 활발하게 진행되고 있다[3].

데이터의 가용성 및 보안성을 모두 고려하기 위해서는 이상적인 그래프와 가장 유사한 결과를 갖는 임계값을 찾아야 한다. 따라서 본 논문에서는 이를 해결하기 위해 그래프의 유사도와 임계값을 바탕으로 회귀분석으로 도식화하여 최적의 임계값을 찾는 방법을 제안한다.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i \qquad (1)$$

(1)은 임계값에 따른 데이터의 가용성을 측정하기 위한 비선형 회귀함수이다. Y_i는 이상적인 그래프와 실제 임계값에 따른 그래프의 유사도를 의미하며 X_i는 임계값을 나타낸다.

그래프 간의 유사도는 길이와 곡률, 그래프로 나누어진 면적을 이용하는 Gromov-Hausdorff 거리를 이용하여 측정한다[4]. 특정 임계값으로 그려진 그래프와 이상적인 그래프의 유사도가 높을수록 해당 데이터에 사용자가 원하는 접근 시간을 보장할 수 있고, 그 기간이 만료되면 사용자가 특별한 행동을 취하지 않아도 키를 복호화 할 수 없으므로 다른 사용자들이 데이터에 접근할 수 없게 된다.

최적의 임계값을 찾기 위한 비선형 회귀모델은 (1)에서 언급한 바와 같다. 훈련 집합과 그 결과를 바탕으로 최적의 임계값을 추론하기 위한 그래프를 만들기 위해서는 최소제곱법을 이용하여 잔차제곱합을 최소화하며 그 식은 (2)와 같다.

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (2)$$

또한 그래프의 형태를 결정하는 계수 β는 다음 식을 만족한다.

$$\frac{\delta}{\delta\beta_0}[\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2] = 0 \qquad (3)$$

$$\frac{\delta}{\delta\beta_1}[\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2] = 0 \qquad (4)$$

(3)과 (4)는 선형 회귀함수에서 β를 구하는 수식이다. 이를 이용하여 비선형 회귀함수의 각 항을 개별적인 독립변수로 취급하고 (2)의 최소제곱법을 이용하여 계산한다. 제안하는 환경에서의 회귀분석은 모든 임계값에 대한 결과를 분석하여 최적의 값을 찾는 것이 불가능하기 때문에 중요한 의미를 갖는다. 특히 분산 해시 테이블 네트워크의 특성상 노드들이 유동적으로 추가/삭제되므로 보호하고자 하는 데이터의 가용성 및 보안성을 결정하는 최적의 임계값의 그때마다 다르다.

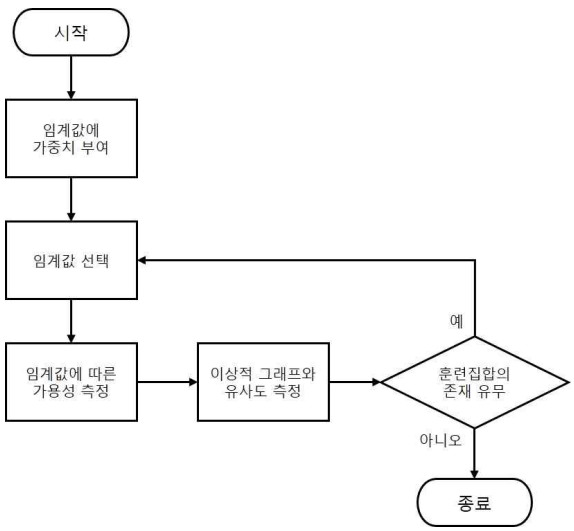

그림 3. 회귀분석을 위한 학습과정의 순서도

그림 3은 회귀분석을 위한 학습과정의 순서도이다. 학습과정이 시작되면 이상적인 그래프와 유사할 확률이 높은 임계값에 가중치를 부여하고 임계값을 선택한다. 그리고 그림 2와 같이 임계값에 따른 그래프가 관측되면 이상적인 그래프와의 유사도를 측정하고 추가 훈련 집합의 존재유무를 검사한다.

4. 성능평가

표 1. 실험 환경

DHT Network Type	Vuze(Azureus)
Num. of Key Shares	100
Num. of Training Set	1000

표 1은 본 논문에서 성능 평가를 위해 구축한 실험 환경이다. P2P 애플리케이션으로 널리 사용되고 있는 Vuze 분산 해시 테이블 네트워크 환경에서 지도학습 과정을 위한 100개의 키 조각들과 1000개의 학습 집단을 구성하였다. 이 때 하나의 학습 집단은 암호화를 위한 키 K와 100개의 키 조각들 $K_1, K_2, \ldots, K_{100}$으로 구성된다.

본 논문에서의 성능 평가를 위한 시나리오는 다음과 같다. 먼저 데이터를 암호화하기 위한 키를 생성하고 Shamir Secret Sharing을 활용하여 100개의 키 조각들을 생성한다. 초기 학습 집단에는 키 조각들로부터 온전한 키를 얻기 위한 임계값이 설정되어있지 않다. 따라서 측정한 가용성 그래프가 이상적인 그래프와 유사할 것으로 예상되는 임계값에 우선순위를 부여하고 임의로 선택하여 결과를 측정한다.

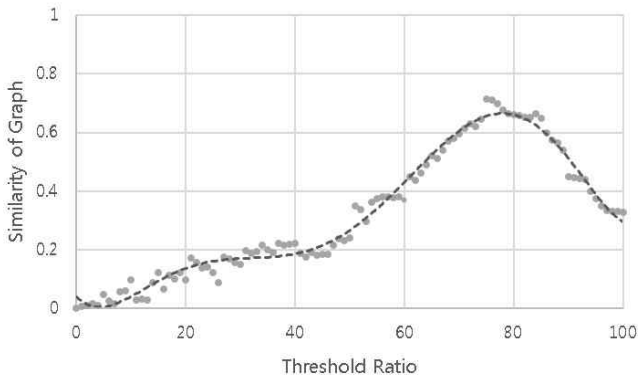

그림 4. 임계값에 따른 가용성 그래프의 유사도 분석

그림 4는 임계값에 따른 데이터 가용성 그래프와 그림 2에서 언급한 이상적인 그래프의 유사도를 분석한 결과이다. 입력 값(임계값의 비율)에 대한 결과 값(그래프의 유사도)을 바탕으로 비선형 회귀함수를 사용하여 최적의 임계값을 판단하기 위한 함수와 그래프를 도출한다. 그리고 그 결과를 바탕으로 그래프의 유사도가 가장 높은 지점에서의 임계값을 계산하여 실제 Self-Destructing Scheme에 적용한다.

회귀분석을 통한 임계값의 예측은 학습이 진행되면서 그래프의 모양이 지속적으로 변한다는 것과 데이터의 크기 및 키 조각의 개수에 따라 최적의 임계값이 다르기 때문에 중요한 의미를 갖는다.

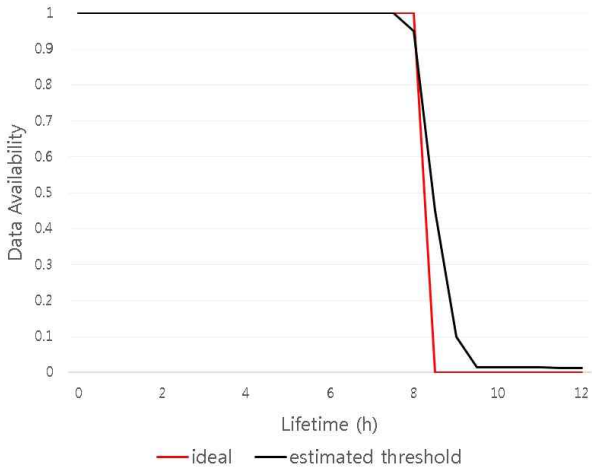

그림 5. 회귀분석으로 추정된 임계값의 가용성 그래프

그림 5는 회귀분석을 통해 추정한 임계값을 바탕으로 데이터의 가용성 그래프를 이상적인 그래프와 비교한 결과이다. 앞서 언급한 그림 2에서의 다른 임계값들의 결과와 비교했을 때, 예측한 임계값으로 키를 분배한 경우 그 조각들이 8시간동안 분산 해시 테이블 환경에서 유지가 되었으며 그 이후에는 거의 모든 키 조각들이 사라진 것을 확인할 수 있다.

5. 결론 및 향후 연구계획

본 논문에서는 Self-Destructing Scheme에서 회귀분석을 적용하여 데이터의 가용성과 보안성을 모두 고려한 최적의 임계값을 찾는 방법을 제안했다. 또한 추정된 임계값의 그래프를 비교하여 성능을 검증했다. 그러나 본 논문에서는 키 조각의 개수가 100개로 제한되었다는 점에서 그 한계를 갖고 있다. 따라서 추후에는 키 조각의 개수와 실행시간 및 데이터의 가용성과 보안성의 trade-off를 고려한 연구가 진행될 것으로 예상한다.

참 고 문 헌

[1] B. Poettering, "Shamir's Secret Sharing," [Online], Available: http://point-at-infinity.org/ssss/, 2006.

[2] Roxana Geambasu, Tadayoshi Kohno, Amit A. Levy, Henry M. Levy, "Vanish: Increasing Data Privacy with Self-Destructing Data," USENIX Security Symposium, pp.299-316, June 2009.

[3] Sean Rhea, Dennis Geels, Timothy Roscoe, John Kubiatowicz, "Handling Churn in a DHT," USENIX Annual Technical Conference, December 2003.

[4] Alexander M. Bronstein, Michael M. Bronstein, Ron Kimmel, Mona Mahmoudi, Guillermo Sapiro, "A Gromov-Hausdorff Framework with Diffusion Geometry for Topologically-Robust Non-rigid Shape Matching," International Journal of Computer Vision, Vol. 89, No. 2, pp.266-286, September 2010.