암호학 기반의 프라이버시 보존형 데이터 마이닝 기술에 관한 연구

유준석*, 홍도원*, 정교일*
*한국전자통신연구원 정보보호연구단
e-mail : jsyu92@etri.re.kr

A Study on Techniques for Cryptographic-based Privacy-Preserving Data Mining

Joon-Suk Yu*, Do-Won Hong*, Kyo-Il Chung*
*Information Security Research Division, ETRI

요 약

최근 들어서 데이터 마이닝은 마케팅, 시장 분석, 사업전략 및 도시계획 수립 등 다양한 분야에서 폭넓게 활용되고 있으며, 새로운 분야로 그 활용 영역을 넓혀가고 있다. 하지만 데이터 마이닝은 그 과정에서 데이터 소유자들의 프라이버시가 침해될 수 있는 문제를 내포하고 있으며, 최근에는 이러한 문제를 해결하고자 하는 노력들이 나타나고 있다. 본 논문에서는 데이터 마이닝에서 이러한 문제를 해결하기 위한 프라이버시 보호 기술들에 대해서 살펴보고 각 방법의 특징에 대해서 기술한다. 특히, 안전한 다자간 계산(Secure multiparty computation)에 기반한 암호학적 프라이버시 보호 기술과 그 활용 가능성에 대해서도 기술한다.

1. 서론

데이터 마이닝과 지식 탐구(Knowledge discovery) 분야는 대량의 데이터에서 이전에 알려지지 않은 패턴을 자동으로 추출해 내는 새로운 연구 분야이다. 지금까지 데이터 마이닝은 데이터를 하나의 중앙 사이트에 모으고 이들 데이터에 대해서 마이닝 알고리즘을 적용하는 데이터 웨어하우스 모델을 사용해 왔다. 하지만 이는 프라이버시 보호 차원에서는 안전하지 못한 방법이다. 예를 들어 보건복지부에서 특정 질병의 발생 패턴을 알아내기 위해서 데이터 마이닝 기술을 사용하기 원한다고 가정하자. 보험회사들은 본 용도에 유용한 방대한 자료를 가지고 있지만 고객들의 프라이버시와 관련된 사항이므로 자료제공을 원하지 않을 것이다. 이에 대한 대안으로 각 보험사들은 개인 환자들을 추적할 수는 없지만 질병의 패턴이나 경향 등을 확인할 수 있는 통계 정보를 제공할 수 있을 것이다.

프라이버시 보존형 데이터 마이닝(Privacy-preserving data mining)은 이러한 문제를 다루기 위한 목적으로 근래에 들어서 연구되기 시작하였으며, 최근에는 다양한 접근 방법을 통한 노력이 이루어지고 있다.

프라이버시 보존형 데이터 마이닝 기술은 크게 두 가지 측면에서 고려되어야 하는데 첫째로 식별자, 주소, 이름 같은 민감한 데이터는 원래 데이터베이스에 변경되거나 제거된 형태로 저장되어야 하고, 둘째로 데이터 마이닝 알고리즘을 통해 주출될 수 있는 민감한 지식 또한 제거되어야 한다. 프라이버시 보존형 데이터 마이닝은 이러한 측면들을 고려하여 마이닝 절차 후에도 비밀 데이터나 지식이 노출되지 않도록 원래 데이터를 변형하는 알고리즘 개발을 목표로 한다.

본 논문에서는 프라이버시 보존형 데이터 마이닝에 대한 개괄적인 내용과 각 방법의 특징들을 살펴본다. 특히, 최근에 주목 받고 있는 암호학적 방법에 기반한 프라이버시 보존형 데이터 마이닝 기술을 설명하고 그 발전 가능성에 대해서 기술한다.

2. 데이터 마이닝 분류

프라이버시 보존형 데이터 마이닝 기술은 크게 다음과 같은 기준으로 분류할 수 있으며[1], 이에 대해서는 본 장의 나머지 부분에서 기술하도록 한다.

- 데이터 분산 형태
- 데이터 변형 방법
- 데이터 마이닝 알고리즘
- 프라이버시 보존 기술

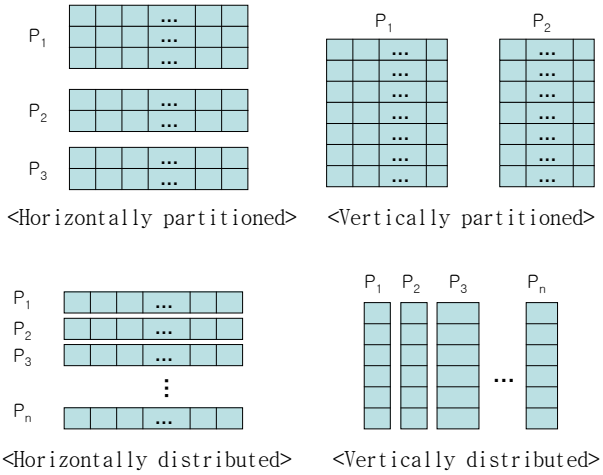
2.1 데이터 분산 형태

데이터 분산 형태에 따른 분류는 데이터 마이닝의 대상이 되는 원본 데이터가 어떠한 형태로 분산되어 있는가에 따라 이루어진다. 이는 모든 데이터가 한 사이트에 집중되어 있는지 혹은 여러 사이트에 분산되어 있는지에 따라 크게 집중형과 분산형으로 나눌 수 있다. 분산형은 다시 레코드 단위로 분산되었는지 혹은 특성값 단위로 분산되었는지에 따라 수평적 분산과 수직적 분산으로 나누어진다. 최근에는 이러한 것들을 혼합한 임의적 분산 데이터를 고려한 데이터 마이닝 기법이 소개되었다[2].

<표 1> 데이터 분산 형태에 따른 분류

구분		특성
집중형		마이닝의 대상이 되는 모든 데이터가 한 사이트에 집중
분산형	수평적 분산	데이터베이스의 레코드 단위로 분산
	수직적 분산	데이터베이스의 특성값 단위로 분산
	임의적 분산	데이터베이스가 레코드나 특성값에 상관없이 임의로 분산

특히, 분산형 데이터의 경우 분산된 정도에 따라서 파티션된 데이터(Partitioned data)와 완전 분산된 데이터(Fully distributed data)로 나누어질 수 있으며, 그림 1에서는 이러한 분산형 데이터에 대한 차이를 도식적으로 설명하고 있다.



(그림 1) 분산형 데이터의 종류

2.2 데이터 변형 방법

일반적으로 데이터 변형은 공개될 데이터베이스의 원래 값을 변형시킴으로써 프라이버시를 보호할 목적으로 사용되며, 어떠한 변형 기술을 사용할지는 데이터베이스를 소유한 조직의 프라이버시 정책에 따라서 결정된다. 이러한 데이터 변형 방법으로는 표 2 와 같은 기술들이 주로 사용되고, 이러한 기술들은 대개 데이터 마이닝 알고리즘을 수행하기 전에 수행되어 프라이버시를 보호하게 된다.

<표 2> 데이터 변형 방법에 따른 기술 분류

기술 구분	기술 내용
Perturbation	하나의 특성값을 새로운 값으로 변경
Blocking	특정 특성값을 "?"로 교체
Aggregation 혹은 merging	몇몇 특성값을 큰 범주로 통합
Swapping	개별 레코드의 특성값을 서로 교환
Sampling	샘플 데이터의 특성값만을 공개

2.3 데이터 마이닝 알고리즘

언급한 바와 같이 데이터 마이닝은 대량의 데이터로부터 특정 패턴을 추출하는 기술로써 비교적 근래에 들어 연구되기 시작한 분야이다. 데이터 마이닝은 추출하려는 지식이 어떤 것이냐에 따라서 다양한 알고리즘들이 존재한다. 그 중에서 A 가 발생하면 B 가 발생한다는 식의 상호 규칙을 밝혀내는 연관규칙 추출(Association rule mining), 특정 라벨을 가진 표본 데이터들로부터 데이터를 분류하는 기준을 추출하고 추출된 기준에 따라 새로운 데이터를 특정 라벨을 가진 부류로 분류해내는 분류화(Classification)와 데이터를 비슷한 특성을 가진 데이터들끼리 묶어내는 군집화(Clustering) 알고리즘 등이 가장 일반적이고 많이 사용되고 있다. 이 외에도 특성화(Characterization)나 요약화(Summarization) 등의 다양한 데이터 마이닝 알고리즘들이 존재한다.

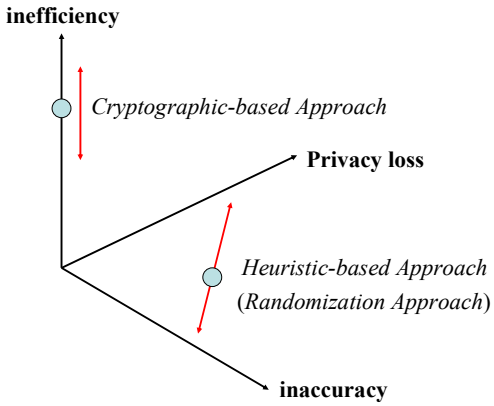
2.4 프라이버시 보존 기술

프라이버시 보존 기술은 데이터의 선택적 변형을 위해서 사용되는 기술로써 선택적 변형은 프라이버시는 침해하지 않으면서 변형된 데이터의 높은 효용성을 달성하기 위해서 필요하다. 이러한 기술은 크게 발견적 방법에 기반한 기술(Heuristic-based techniques)과 암호학적 방법에 기반한 기술(Cryptographic-based techniques)로 나눌 수 있다. 발견적 방법에 기반한 기술의 경우 모든 특성값이 아닌 효용성 손실과 프라이버시 누출을 최소화할 수 있는 선택된 특성값들에 대해서만 데이터 변형이 가해지며, 이때 데이터 변형은 2.2 절에서 설명한 방법들이 사용된다. 일반적으로 발견적 방법에 기반한 기술을 사용할 경우에는 원본 데이터에 어떤 형태로든 변화가 가해지거나 잡음이 포함된다. 따라서 데이터 마이닝을 통해 얻어지는 결과의 정확성은 필연적으로 떨어지게 되고 결과값의 효용성 또한 낮아진다. 그러나 원본 데이터에 변형이 많이 가해질수록 정보가 노출될 소지는 줄어들게 되며, 발견적 방법 기반의 기술을 사용할 때는 정확성과 프라이버시의 노출이라는 관점에서 적정한 수준의 변형

이 이루어져야 할 것이다[3].

반면 암호학적 방법에 기반한 기술은 안전한 다자간 계산(Secure multiparty computation)을 사용하여 입력 데이터가 노출되는 것을 방지하며, 데이터 마이닝의 대상이 되는 원본 데이터에는 어떠한 변형도 가해지지 않는다. 그러므로 데이터 마이닝 결과 얻어지는 정보는 항상 정확성을 보장하면서도 프라이버시 노출에 대한 문제도 가지지 않는다는 장점을 지니기 때문에 최근에 주목 받는 연구분야로 떠오르고 있다. 하지만 아직까지는 안전한 다자간 계산 기술이 프로토콜의 복잡성으로 인해 비효율적이라는 지적을 받고 있으며, 이는 안전한 다자간 계산을 범용 프로토콜이 아니라 특정 용도에 적합하도록 개발함으로써 효율성을 향상시킬 수 있을 것이다.

그림 2 는 발견적 방법 기반 기술과 암호학적 방법 기반 기술의 특징을 비교하여 나타내고 있다.



(그림 2) 발견적 방법 기반 기술과 암호학적 기반 기술의 특징 비교

3. 암호학적 방법에 기반한 프라이버시 보존 기술

본 장에서는 암호학적 방법 기반의 프라이버시 보존형 데이터 마이닝의 핵심 기술인 안전한 다자간 계산 기술에 대해서 기술하도록 한다.

3.1 안전한 다자간 계산의 개념

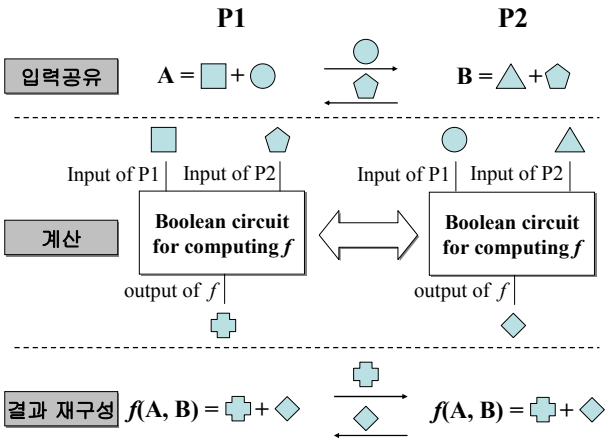
안전한 다자간 계산은 1982 년 Yao 에 의해 처음 제안된 이후로 Goldreich-Micali-Wigderson 에 의해 일반화되었고 지금까지도 많은 연구가 이루어져 오고 있다[4, 5]. 일반적으로 안전한 다자간 계산 개념은 계산 참여자들의 입력값으로부터 함수 f 를 계산하는 데에 있어서 계산과정 종료 후, 참여자들은 자신의 입력값과 계산된 결과, 그리고 이들로부터 유추할 수 있는 것 외에는 어떠한 정보도 얻을 수 없도록 하자는 것이다[6, 7].

이러한 안전한 다자간 계산의 목적은 TTP 를 사용하면 쉽게 달성될 수 있다. 즉, 계산 참여자들은 자신의 입력값을 TTP 에게 비밀리에 전달하고 TTP 가 결과를 계산하여 각 참여자에게 비밀리에 전송하는 것이다. 하지만 TTP 없이 동일한 기능을 수행할 수 있다면 이는 안전성의 관점에서 훨씬 우수한 프로토콜이며, 이런 맥락에서 안전한 다자간 계산은 암호학의 중요한 부분으로 자리잡아 오고있다.

지금까지의 연구에 의하면 유한 도메인에서 정의된 모든 함수는 안전한 양자간 혹은 다자간 계산이 가능한 것으로 알려져 있다. 이는 함수에 대한 결과를 계산하는 이진회로 구성을 기초로 하여 불확실 전송(OT, Oblivious Transfer), 위탁(Commitment), 비밀분산(SS, Secret Sharing) 및 검증 가능한 비밀분산(VSS, Verifiable Secret Sharing), 영지식 증명(ZKIP, Zero-Knowledge Interactive Proof) 등의 기술을 복합적으로 이용함으로써 설명이 가능하다[6, 7]. 즉, 트랩도어 일방향 순열(Trapdoor one-way permutation)의 존재를 가정하여 불확실 전송을 달성할 수 있고 불확실 전송을 통해 안전한 양자간 계산 프로토콜을 구성할 수 있다. 또한 안전한 양자간 계산에 검증 가능한 비밀분산을 적용하여 약간의 수정만 가한다면 다자간 환경으로 확장할 수 있다.

3.2 안전한 다자간 계산의 구성 방법

본 절에서는 안전한 양자간 계산 프로토콜의 구성 방법에 대해서 기술하고, 이를 안전한 다자간 계산 프로토콜로 확장하는 방법에 대해서 설명한다.



(그림 3) 안전한 양자간 계산 프로토콜의 구성 예

간단한 예로 두 계산 참여자 P1 과 P2 가 각각 한 비트의 입력 A 와 B 를 가지며, 한 비트의 출력을 내는 함수 f 를 안전하게 계산하는 경우를 살펴보자. 각 사용자는 함수 f 를 알며, 이를 계산하는 이진 회로를 공유하고 있다고 가정한다. 이 때, 각 계산 참여자가 함수 f 를 안전하게 계산하기 위해서는 입력 공유, 계산, 결과 재구성의 세 단계로 구성된 프로토콜을 수행하게 된다. 먼저 입력 공유 단계에서 각 참여자는 입력값이 상대방에게 노출되는 것을 방지하기 위해서 이를 두 개의 랜덤한 분배값으로 나누고 이 중 하나를 상대방에게 전달한다. 계산 단계에서 각 참여자는 입력 공유 단계에서 분배된 분배값을 회로의 적당한 입력단에 설정하여 계산을 수행한다. 마지막으로 결과

재구성 단계에서 각 참여자는 이진 회로의 출력으로 나온 결과값을 상대 참여자에게 전달하고 각 참여자는 상대방으로부터 수신한 값과 자신의 결과값을 더하여 f (A, B)를 계산한다. 그림 3 은 이를 간단히 도식화하여 보여준다.

언급한 바와 같이 안전한 다자간 계산은 양자간 계산 프로토콜을 확장하여 달성할 수 있는데 기본적인 확장 방법은 각 참여자들이 입력 공유 단계에서 입력값을 계산 참여자 수만큼의 분배값으로 나누어 분배하는 것이다. 하지만 이러한 단순한 확장 방법으로는 참여자들이 프로토콜을 따르지 않는 악의적인 공격자 환경에서 프로토콜의 안전성을 보장하지 못하며, 추가적인 기술을 사용한 변경을 필요로 한다. 위탁 및 영지식 증명 기술을 사용하는 검증 가능한 비밀분산 기법은 본 문제를 해결하기 위한 핵심 기술로써 강제적으로 참여자들이 프로토콜을 따르도록 하여 악의적인 공격자 환경에서 프로토콜의 안전성을 유지할 수 있도록 하고 있다[6, 7].

3.3 데이터 마이닝을 위한 핵심 기술

데이터 마이닝 과정 중에 암호학적 방법을 통하여 프라이버시를 보호하기 위해서는 아래에 나열한 작업들을 안전하게 계산해 내는 것이 핵심 기술이며, 이를 위해서 안전한 다자간 계산 기술이 사용된다[8]. 또한 프라이버시를 보존하는 데이터 마이닝 과정에서 아래의 작업들은 복합적으로 사용될 수도 있다.

- 각 사이트에 산재해 있는 값의 합을 안전하게 계산
- 각 사이트가 가지고 있는 원소를 공개하지 않고 이들의 합집합을 계산
- 각 사이트가 가지고 있는 때, 이들 교집합의 원소의 수를 계산
- 두 벡터의 내적(scalar product)을 안전하게 계산

3.4 기타 응용분야 및 활용 가능성

프라이버시 보존을 위해 사용될 수 있는 대표적인 암호기술인 안전한 다자간 계산 기술은 데이터 마이닝 외에도 데이터베이스 질의, 통계 분석, 기하학 계산, 과학 계산 등 다양한 분야에서 폭넓게 활용될 수 있다[9]. 이러한 가능성과 폭넓은 효용성에도 불구하고 지금까지의 다자간 계산에 대한 연구는 대부분 이론적 수준에 머물러 있었고, 성능적인 측면을 고려할 때 실용적이라기 보다는 연구만을 위한 테마로 인식되어 온 것이 사실이다. 하지만 최근 들어 다자간 계산 기술을 실용적인 응용에 접목하려는 노력이 이루어지면서 향후 성능적인 측면에서의 장애는 해소될 것으로 예상되며, 이에 따라 그 활용 분야도 더욱 넓어질 것이다.

4. 결론

데이터 마이닝은 기업의 마케팅 및 사업 전략 수립, 국가의 정책 수립 등에 있어서 유용하게 활용되고 있으며, 꾸준히 응용 영역을 넓혀가고 있다. 하지만 데이터 소유자의 프라이버시 침해 가능성에 따라 프라이버시 보존 기술에 대한 연구가 활발히 이루어지고 있다. 데이터 마이닝에서 프라이버시 보존 기술은 크게 발견적 방법과 암호학적 방법으로 나눌 수 있으며, 각 방법은 마이닝 결과의 정확성, 프라이버시의 노출 정도, 성능 관점에서 각각 장단점을 가진다. 특히, 암호학적 방법을 이용한 프라이버시 보존 데이터 마이닝은 안전한 다자간 계산을 통하여 프라이버시를 보호함으로써 마이닝 결과의 정확성과 프라이버시 보호를 보장하는 장점을 지닌다. 아직까지 암호학적 방법에 기반한 데이터 마이닝 기술이 성능적인 면에서 문제를 가지지만 데이터의 활용 목적을 생각할 때 결과의 정확성에 더욱 무게를 둘 수 있을 것이다. 또한 향후 특정 용도에 적합한 암호학적 기술 개발이 이루어진다면 성능적인 측면에서도 실용적인 수준에 이를 수 있을 것으로 기대된다.

참고문헌

[1] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, “State-of-the-art in Privacy Preserving Data Mining”, Proc. of SIGMOD, 2004

[2] G. Jagannathan, and R. N. Wright, “Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data”, Proc. of 11th ACM SIGKDD, 2005

[3] B. Pinkas, “Privacy-Preserving Data Mining: A Cryptographic Approach”, IBM Almaden Institute on Privacy in Data, 2003

[4] A. C. Yao, "Protocols for Secure Computation", Proc. of IEEE FOCS '82, pp. 160-164, 1982

[5] O. Goldreich, S. Micali and A. Wigderson, "How to Play Any Mental Game or Completeness Theorem for Protocols with Honest Majority", Proc. of ACM STOC '87, pp. 218-229, 1987

[6] O. Goldreich, "Secure Multi-Party Computation (Final (incomplete) Draft, Version 1.4)", available at http://www.wisdom.weizmann.ac.il/~oded/PS/prot.ps

[7] R. Cramer, "Introduction to Secure Communication", Lecture Note of Aarhus Summer school in Cryptography and Data Security, 2000

[8] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, “Tools for Privacy Preserving Distributed Data Mining”, Proc. of ACM SIGKDD, 2003

[9] W. Du, and M. J. Atallah, “Secure Multi-Party Computation Problems and Applications: A Review and Open Problems”, Proc. of New Security Paradigms Workshop, 2001