情報保護學會論文誌
第 23 卷 第 1 號, 2013. 2

프라이버시 보장 k-비트 내적연산 기법*

이 상 훈,† 김 기 성, 정 익 래‡
고려대학교

Privacy-Preserving k-Bits Inner Product Protocol*

Sang Hoon Lee,† Kee Sung Kim, Ik Rae Jeong‡
Korea University

요 약

정보의 양이 많아짐에 따라 많은 양의 정보를 효과적으로 관리, 운용할 수 있는 데이터 마이닝 기법의 연구가 활발해졌다. 다양한 데이터 마이닝 기법들이 연구되었는데 그 중에는 프라이버시를 보호할 수 있는 프라이버시 보호 데이터 마이닝(Privacy Preserving Data Mining) 연구도 진행됐다. 프라이버시 보호 데이터 마이닝은 크게 연관규칙, 군집화, 분류 등의 알고리즘이 존재한다. 그 중 연관규칙 알고리즘은 데이터간의 연관규칙을 찾아내는 알고리즘으로 주로 마케팅에 주로 사용된다.

본 논문에서는 Shamir의 비밀 분배 기법을 이용하여 다자간 프라이버시 보호 데이터 마이닝 환경에서 단일 비트가 아닌 멀티 비트 정보를 공유할 수 있는 내적연산 기법을 제안한다.

ABSTRACT

The research on data mining that can manage a large amount of information efficiently has grown with the drastic increment of information. Privacy-preserving data mining can protect the privacy of data owners. There are several privacy-preserving association rule, clustering and classification protocols. A privacy-preserving association rule protocol is used to find association rules among data, which is often used for marketing.

In this paper, we propose a privacy-preserving k-bits inner product protocol based on Shamir's secret sharing.

Keywords: Data Mining, Association Rule, Inner Product, Secret Sharing

I. 서 론

무선 인터넷의 발달과 스마트폰, 태블릿 PC 등의 보급으로 인하여 시간과 공간에 제약받지 않고, 인터넷이나 소셜 미디어를 손쉽게 사용할 수 있게 되었다. 이로 인하여 개인이 만들어 내는 데이터의 양이 기하급수적으로 증가하게 되었으며, 이렇게 생산된 데이터들은 실세계의 데이터와 일치하는 데이터들이다. 데이터들은 일반적으로 데이터베이스에 저장되며, 데이터베이스로부터 질의 검색을 통해 얻는 정보들은 기본적이고 일반적인 정보들이다. 이러한 대용량의 데이터베이스에서 기본적이고 일반적인 데이터들 간의 연관성이나 함축적인 정보를 얻는 방법이 데이터 마이닝(Data Mining)이다. 데이터 마이닝은 단순한 질의 검색을 통해 정보를 얻어 내는 것이 아니라 복잡하고 다양한 마이닝 기법을 이용하여야만 정보들 간의 연관성이나 함축적인 정보를 얻어낼 수 있다. 데이터 마이닝으로 얻어진 정보들은 의사 결정에 적용될 경우, 보

접수일(2012년 10월 26일), 수정일(2013년 2월 4일), 게재확정일(2013년 2월 4일)
* 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (20120007037)
† 주저자, chclan@naver.com
‡ 교신저자, irjeong@korea.ac.kr

다 효율적인 결정을 하는데 도움이 된다. 데이터 마이닝 기법에는 연관규칙(Association Rules), 분류(Classification), 군집화(Clustering) 등 여러 알고리즘이 존재한다. 이러한 알고리즘들은 여러 분야에서 다양하고 지속적으로 연구되고 있다. 특히, 연관규칙 알고리즘은 트랜잭션 데이터들뿐만 아니라 XML 등의 문서들 간의 연관성이나 태그 및 속성들 간의 연관성을 찾는 데에도 사용된다.

이렇게 데이터 마이닝은 많은 실생활에 적용될 수 있는데, 이로 인하여 프라이버시 침해 문제가 나타나기도 한다. 예를 들어 병원에서는 환자의 치료를 위해 데이터를 분류하고 이를 통해 환자의 정확한 상태를 파악하여 상황에 맞는 치료를 하거나 약을 처방할 수 있다. 이 경우 환자의 데이터에 대해 데이터 마이닝 수행 시 환자의 진료 기록 데이터를 그대로 사용면 환자의 민감한 정보가 노출될 위험이 있다. 또한 대형마트에서 소비자들의 구매내역을 분석해 마케팅에 활용할 수 있는데, 만약 구매내역 데이터를 그대로 사용할 경우 소비자의 구매패턴이 공공연히 노출되는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위한 데이터 마이닝 기법이 프라이버시 보호 데이터 마이닝(Privacy Preserving Data Mining)이다. 프라이버시 보호 데이터 마이닝은 기존의 데이터 마이닝에서 사용하던 근원 데이터의 분포를 변화시키거나 트랜잭션의 익명화, 암호학적 기법을 이용한 데이터 전달을 통해 데이터에 포함된 민감한 정보를 보호하면서, 이러한 기법이 적용된 마이닝 결과가 근원 데이터의 마이닝 결과와 같도록 하는 방법이다.

기존의 프라이버시 보호 데이터 마이닝 기법의 알고리즘은 오직 하나의 비트 정보만을 얻을 수 있거나, 멀티 비트의 정보를 얻을 수 있다 하더라도 통신 복잡도가 높아지는 단점이 존재하였다. 또한 각 사용자들이 가지고 있는 비트의 정보가 노출되는 문제점도 발견되었다. 이러한 문제점을 보완하기 위해 본 논문에서는 Shamir의 비밀 분산 기법을 이용하여 다자간 프라이버시 보호 데이터 마이닝 환경에서 멀티 비트 정보를 공유할 수 있는 내적연산 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 마이닝과 프라이버시 보호 데이터 마이닝에 대한 관련 연구를 살펴본다. 3장에서는 Shamir 비밀 분산 기법과 연관규칙 마이닝 알고리즘을 설명한다. 4장에서는 다자간 프라이버시 보호 데이터 마이닝 환경에서 멀티 비트 정보를 공유할 수 있는 내적연산 기법을 제안하며, 마지막으로 5장에서 결론을 맺는다.

II. 관련 연구

본 장에서는 제안하는 기법의 환경이 되는 데이터 마이닝과 프라이버시 보호 데이터 마이닝에 대해 설명한다.

2.1 데이터 마이닝

데이터 마이닝은 대용량 데이터로부터 데이터를 추출하는 과정, 추출된 데이터를 적절히 가공하고 정제하는 과정, 가공된 데이터에 마이닝 알고리즘을 수행하는 과정, 알고리즘 결과를 재가공하여 기존의 지식과 통합하여 새로운 정보를 얻는 과정을 거치게 된다. 이러한 과정으로 얻은 정보는 예측이나 분류 모델을 만들거나 데이터 간의 연관성을 밝혀내거나 전체 데이터의 내용을 함축해 주기도 한다. 이런 정보들은 크게 분류, 연관규칙, 군집화, 이 3가지 알고리즘을 통해서 얻어지게 된다.

분류 알고리즘은 기존 데이터 사이의 규칙을 찾아내 분류 코드를 만들어 새로운 데이터가 들어오게 되면 만들어진 분류 코드에 따라 나눈다. 예를 들어, 은행이 고객의 나이, 성별, 연봉, 결혼유무 등의 데이터를 갖고 있고, 신용불량자인지를 나타내는 집합을 갖고 있다면, 이를 근거로 규칙을 찾아내어 새로운 고객에 대한 신뢰도를 예측할 수 있게 된다. 대표적인 분류 알고리즘으로는 의사결정나무분석(Decision Tree Analysis)[8]이 있다.

연관규칙 알고리즘은 데이터들 간의 보이지 않는 연관규칙을 찾는다. 이를 위해 전체 지지도, 신뢰도, 향상도의 3가지 측도를 사용하여 분석한다. 지지도는 전체 데이터에서 임의의 두 데이터가 동시에 나타나는 비율이고 신뢰도는 데이터 A가 포함된 레코드에서 데이터 A, B가 동시에 나타나는 비율을 의미한다. 향상도는 두 데이터가 서로 독립적인지를 판단하는 비율이다. 이 3가지 측도는 데이터에서 많은 패턴과 규칙들 중에서 가장 유용하고 흥미 있는 것들을 찾아내는 척도로 사용한다. 연관규칙 알고리즘에 대표적으로 사용되는 기법에는 내적연산 기법[7]이 있다.

군집화 알고리즘은 전체 데이터의 분포 상태를 찾아내는 데 유용하게 사용된다. 분류 알고리즘과 다른 점은 군집화 알고리즘은 미리 정의된 클래스나 분류 규칙을 학습할 예제 데이터가 없다는 것이다. 데이터가 가지는 자기 유사성에 의해서만 군집을 나누게 된다. 군집화의 결과는 전적으로 데이터 마이너(data

miner)에 의해 의미가 부여된다. 대표적인 군집화 알고리즘으로는 데이터들의 밀도를 기반으로 하는 DBSCAN[10]과 데이터간의 거리를 기반으로 하는 k-means 알고리즘[30]이 있다.

2.2 프라이버시 보호 데이터 마이닝[3]

인터넷이 보급되고 사람들이 수많은 인터넷 사이트에 가입함으로써 사용자들의 프라이버시는 제3자에 의해 관리가 되고 있다. 이런 제3자에 의한 프라이버시의 관리가 소홀하여 많은 프라이버시 유출 사고로 이어졌다. 이에 근래에는 프라이버시가 어느 시스템에서라도 안전해야할 중요한 요소로 자리 잡게 되었다. 이러한 이유로 데이터 마이닝에서도 의미 있는 정보를 유출하는 과정에서 민감한 프라이버시가 노출이 되지 않는 방법을 연구하게 되었다. 그렇기에 프라이버시 보호 데이터 마이닝의 목적은 대용량의 데이터에서 의미가 있는 정보를 추출하면서 동시에 민감한 정보를 보호하는 것이다.

프라이버시 보호 데이터 마이닝에서 주된 고려 사항은 두 가지이다. 첫 째는 이름, 주소 같은 개개인의 민감한 정보이다. 이런 정보는 데이터 수령인에게 조금의 정보도 노출이 되면 안 되기 때문에 데이터의 변조가 필요하다. 두 번째는 데이터 마이닝 알고리즘에 사용된 데이터로부터 얻어지는 민감한 정보이다. 이러한 정보는 잘 조합되면 원본 데이터의 프라이버시를 침해할 수 있기 때문에 노출되지 않도록 주의해야 한다.

프라이버시 보호 데이터 마이닝이 민감한 정보를 완전히 보호할 수 있는 것은 아니다. 예를 들어 프라이버시 보호 데이터 마이닝에 사용되는 안전한 다자간 계산 기술(Secure Multiparty Computation, SMC)은 데이터의 민감한 정보를 노출하지 않는다. 하지만 데이터 마이닝의 결과는 모든 사용자들이 민감한 정보를 유추할 수 있게 한다. 이러한 결과가 SMC가 취약하다는 것이 아니라 단지 결과 그 자체가 프라이버시 보호를 침해한다는 것이다.

프라이버시 보호 데이터 마이닝에는 다양한 기법들이 존재한다. 대표적인 기법으로는 다음에 설명하는 신뢰하는 제3자를 이용하는 기법, 데이터 조작 기법, SMC를 이용하는 기법 등이 대표적이다.

2.2.1 신뢰하는 참여자를 이용하는 기법

분산되어 있는 민감한 데이터의 데이터 마이닝을 위한 고전적인 접근 방법이다. 모든 데이터를 한 곳에 저장하고 그 곳에서 데이터 마이닝을 수행하는 것이다. 이 기법은 모든 데이터를 저장하고 있는 참여자가 모든 데이터를 소유하고 있기 때문에 모든 참여자들의 프라이버시를 유지한다는 믿음을 전제로 한다. 만일 전적으로 신뢰할 수 있는 제3자가 있는 경우 모든 참여자들은 제3자에게 모든 데이터를 보내고, 제3자는 마이닝 결과 값을 다시 모든 참여자에게 배포함으로써 마이닝을 수행하게 된다.

하지만 실세계에서 이렇게 전적으로 신뢰할 수 있는 제3자를 찾는 일은 쉽지 않다. SMC 기법을 사용하여 신뢰하는 제3자 없이 데이터 마이닝을 수행할 수 있지만 그럴 경우 많은 계산 복잡도와 통신 복잡도가 요구된다.

2.2.2 데이터 조작 기법

데이터 조작 기법은 2000년에 R. Agrawal[3] 등에 의해 처음으로 제안되었다. 데이터 조작 기법에는 덧셈, 곱셈[17], 행렬 곱셈, k-익명화[26], 데이터 재분배[21] 등을 이용한 다양한 방법이 존재한다.

덧셈을 이용한 데이터 조작은 원본 데이터에 노이즈 값을 더한다. 이 때 노이즈 값은 각 원본 데이터를 알아볼 수 없을 정도로 충분히 큰 값이어야 한다. 그렇기 때문에 이 기법은 조작된 데이터로부터 특정분포를 얻도록 설계된다.

예를 들어 $X=\{x_1, x_2, \cdots, x_N\}$인 데이터 집합이 있다고 하자. 각 데이터에 확률 분포 $f_Y(y)$로부터 얻은 노이즈 값 $Y=\{y_1, y_2, \cdots, y_N\}$을 각각 더해준다. 그럼 변현된 새로운 데이터 $Z=\{z_1=x_1+y_1, \cdots z_N=x_N+y_N\}$를 얻게 된다. 데이터 소유자는 원본 데이터 집합 X를 Z로 바꾸어 데이터 마이닝을 수행한다. Z에는 데이터 소유자만이 아는 노이즈 값이 추가되어있기 때문에 원본 데이터에 대한 정보는 직접 노출이 되지 않지만 데이터 마이닝을 통해 원본 데이터와 동일한 분포를 갖는 결과를 얻어낼 수 있게 된다.

데이터 조작 기법은 효율성에서는 탁월하지만 노이즈 값이 고른 분포로 원본 데이터에 추가되지 않는다면 프라이버시 측면에서 취약점을 드러낸다.

2.2.3 SMC를 이용하는 기법

SMC는 1986년 A. C. Yao[32]에 의해 소개 되었다. SMC의 기본적인 의도는 계산이 끝나는 시점에

어느 참여자도 본인의 입력 값과 계산 결과를 제외하고는 얻을 수 있는 정보가 없도록 하는 것이다. 즉, 본인만의 비밀 데이터베이스를 갖고 있는 둘 이상의 참여자들은 불필요한 정보의 노출 없이 그들의 전체 데이터베이스를 데이터 마이닝을 원하게 된다. 예를 들어 분리되어 있는 의료 기관을 생각해보자. 이들은 각자 소유하고 있는 환자에 대한 프라이버시를 보호하면서 전체 환자에 대한 공동의 연구 결과를 얻도록 해주는 것이 SMC를 이용한 데이터 마이닝 기법이다. 이 기법은 최근에 연관규칙 알고리즘, 군집화 알고리즘 등에 사용되고 있다.

SMC는 어떠한 정보의 노출 없이 정확한 마이닝 결과를 얻을 수 있도록 하지만 많은 계산 복잡도와 통신 복잡도가 요구된다. 이는 특히 대용량의 데이터에서 매우 치명적으로 작용하게 된다.

III. 배경지식

3.1 Shamir 비밀 분배 기법[25]

Shamir가 제안한 비밀 분배 기법에서 분배자(dealer)는 수신자들에게 분할된 정보에 대한 비밀 보장이 기본 조건이다. 그리고 n명의 수신자들에게 분배된 분할 정보는 임의의 k명 이상이 정보를 재구성하였을 때 비밀 정보가 복원이 되어야 한다. 이 때 임의의 $k-1$명 이하가 모여 정보를 재구성하였을 때는 비밀 정보에 대한 어떠한 정보도 노출이 되지 않아야 한다. Shamir의 비밀 분배 기법은 다음과 같다.

- 분배자는 비밀 정보 D를 상수항으로 하는 임의의 $k-1$차 다항식을 선택한다.

$$q(x) = a_0 + a_1 x + \cdots + a_{k-1} x^{k-1}, \text{ 단 } a_{k-1} \neq 0$$

- 비밀 정보인 D를 n개의 조각으로 나누어야 하는데 그 값은 각각 D_i로, $q(x)$에 0이 아닌 서로 다른 임의의 값 $x_1, \cdots, x_n$을 넣은 값으로 설정을 한다. 즉, $D_1 = q(x_1), D_2 = q(x_2), \cdots, D_n = q(x_n)$ 이다. 비밀 정보를 담고 있는 $q(x)$의 상수항을 구하기 위해서는 k개의 D_i 값만 있으면 된다. $q(x)$는 $k-1$차 다항식이기 때문에 k개의 D_i 값을 다항식 보간법(polynomial interpolation)을 이용하여 $q(x)$를 재구성할 수 있게 된다. 그러면 $q(0)$ 값인 비밀 정보 D를 재구성할 수 있게 된다.

3.2 연관규칙 마이닝 알고리즘

연관규칙 마이닝 알고리즘은 대용량의 데이터 집합에서 단위 트랜잭션에서 빈번하게 발생하는 사건의 유형을 찾아내는 방법이다. 이는 주로 시장 마케팅에 사용된다. 예를 들어 대형마트에는 매일 고객들이 구입한 품목에 대한 데이터가 수집이 된다. [표 1]처럼 간단한 구입 목록에서는 쉽게 맥주를 구입한 사람은 주로 스낵을 구입한다는 것을 알 수 있다. 이러한 규칙을 찾게 되면 마트에서는 맥주와 스낵을 함께 진열함으로써 판매 효과를 극대화할 수 있다. 하지만 이러한 구입 목록의 양이 매우 커진다면 연관규칙 마이닝 알고리즘을 사용하여 특정 규칙을 찾아낼 수 있게 된다.

본 논문에서 언급할 연관규칙 마이닝은 다음과 같이 정의한다.

아이템들의 집합 $I = \{i_1, i_2, \cdots, i_m\}$ 이 있다고 가정하자, T는 $T \subseteq I$인 아이템들의 집합이고, DB는 이러한 트랜잭션 T들의 집합이라 하자. 아이템의 집합 X는 필요충분조건으로 $X \subseteq T$를 만족하는 경우 트랜잭션 T가 아이템 집합 X를 포함한다고 한다. 연관규칙은 다음과 같이 나타낸다.

- 연관규칙 $X \Rightarrow Y$은 각각의 아이템 집합 X, Y가 $X \subseteq I, Y \subseteq I$ 이고 $X \cap Y = \varnothing$를 만족하는 경우에 표현된다.

- 연관규칙 $X \Rightarrow Y$은 DB에 $s\%$로 $X \cup Y$를 포함하는 트랜잭션이 존재하면 지지도(support) s를 갖는다고 한다. 이는 확률 $P(X \cup Y)$를 계산함으로써 얻을 수 있다. 이러한 지지도는 해당 규칙이 주어진 데이터 집합에 얼마나 자주 적용할 수 있는지를 결정한다.

- 연관규칙 $X \Rightarrow Y$은 DB에 X를 포함한 하고 Y 또한 포함한 거래 내역이 $c\%$ 존재하면 신뢰도(confidence) c를 갖는다고 한다. 이는 조건부 확률 $P(Y|X)$를 계산함으로써 얻을 수 있다. 이러한 신뢰도는 연관규칙에 의해 만들어지는 추론에 대한 확실성 정도를 나타낸다.

연관규칙 마이닝 알고리즘의 가장 중요한 부분은 사용자가 설정한 임계값 이상을 갖는 지지도와 신뢰도

〔표 1〕 구입 목록 예제

거래번호	구입 목록
1	빵, 맥주
2	빵, 맥주, 계란, 스낵
3	우유, 스낵, 맥주, 음료
4	빵, 우유, 맥주, 스낵
5	빵, 우유, 스낵, 콜라

를 가지는 모든 규칙을 찾는 것이다. 이러한 임계값을 모두 만족하는 규칙을 강한(strong) 규칙이라고 한다.

본 논문은 다자간 환경을 가정한다. 기존에 연구 되었던 다자간 환경은 대체로 단 두 명의 참가자만을 가정한다. 게다가 데이터의 프라이버시를 보호하기 위해 암호학적인 기법을 사용하였기 때문에 계산 복잡도는 높아지고 확장성이 떨어지는 문제점이 제기되었다.

본 논문은 불리안(Boolean) 연관규칙을 가정한다. 이는 모든 속성은 0 또는 1로 이루어지기 때문에 구매 내역은 0과 1로 이루어진 벡터로 처리한다.

데이터 마이닝 알고리즘 중 하나인 연관규칙 알고리즘은 주로 시장성 분석을 위해서 사용이 된다. 이는 엄청나게 많은 거래 사이에서 특정한 유형을 찾아내어 어느 구입상품의 존재가 또 다른 구입상품의 존재를 암시하는 특정 규칙을 찾아내기 위해 사용된다.

3.3 데이터 모델

다자간 환경의 데이터 마이닝에는 수직 분할 데이터와 수평 분할 데이터, 두 가지 타입의 데이터 모델이 존재〔31〕한다. 수평 분할 데이터는 정보 제공자들이 모두 같은 속성 데이터를 갖게 된다. 즉, 서로 다른 정보 제공자들은 동일한 속성으로 이루어진 데이터를 가지게 된다. 하지만 수직 분할 데이터는 정보 제공자들이 모두 동일한 구매에 대한 정보를 가지게 된다.

본 논문에서는 수직 분할된 데이터에 적용하는 연관규칙 마이닝 알고리즘을 가정한다.

IV. 제안하는 기법

4.1 X. Ge 등의 기법〔11〕

데이터로부터 연관규칙을 얻기 위해서는 가능성 있는 아이템 집합의 신뢰도와 지지도를 계산해야만 한다.〔11〕은 특정 아이템 집합의 발생 빈도를 계산하기 위해 아이템의 속성 값을 1 또는 0이 되도록 구성한다. 특정 아이템 집합의 모든 속성이 1이 되는 트랜잭션의 수를 카운트(c.count)한다. c.count가 사용자가 지정한 임계값 이상이면 그 아이템 집합은 높은 빈도를 갖는 아이템 집합으로 지정한다.

다자간 환경에서 서로의 근원 데이터를 노출하지 않고 합동하여 c.count를 계산하는 것은 쉬운 일이 아니다. 〔11〕에서는 이 문제를 shamir의 비밀 공유 기법을 기반으로 하여 개인적으로 c.count를 계산하는 알고리즘을 제안한다. 다음은 그 알고리즘을 간략히 설명한다.

- n명의 P_i들은 Shamir 기법에 사용되는 다항식의 차수 k를 정한다. 모든 참여자가 알 수 있는 공개 값 $X=(x_1, x_2, \cdots, x_n)$을 정한다. 이 값은 각 참여자 P_i에게 pub_i 값이 된다. 그리고 참여자들의 데이터는 A_i로 표현한다. A_i는 벡터 형식으로 P_i의 데이터를 담고 있는데 이 속성들은 0 또는 1로 이루어진 비트 형식이다. 즉, $A_i=(1,1,0,1,\cdots,0)$ 형식으로 표현된다. 개인 다항식 $q_i(x)$의 상수항에 A_{ij} 값을 설정하고 나머지 계수에는 임의의 값을 설정한다. 사용자 P_i는 다른 사용자 P_j에게 $q_i(pub_j)$값을 전송한다. $q_1(pub_j), q_2(pub_j), \cdots, q_n(pub_j)$를 모두 받은 P_j는 $S(pub_j)$ 값을 모두와 공유하여 다항식 $S(x)$를 재구성한다. 이 때 재구성 된 $S(x)$의 상수항이 내적연산을 하기 위해 삽입된 비트 정보의 합이 된다.

〔표 2〕 제안하는 기법에서 사용하는 표기법과 의미

표기법	의 미
n	참여자 수
k	연산할 정보의 비트 수
P_i	참여자
P_1	다항식 생성자
P_2	랜덤 생성자
P_3	결과 배포자
A_i	참여자 P_i가 가지고 있는 데이터
A_{ij}	참여자 P_i가 갖고 있는 A_i의 j번째 트랜잭션
$q_i(x)$	참여자 P_i의 개인 다항식
t	$q_i(x)$의 차수, (단, $n-1 \le t$)
$S(x)$	모든 참여자의 개인 다항식을 더한 다항식
pub_i	참여자 P_i의 공개 값

〔11〕에서 제안한 기법은 다른 암호학적 기법을 사용하지 않았기에 연산 속도가 빠르지만 한 번 연산 시

에 단일 비트의 내적연산 결과만을 알 수 있었다. 게다가 마지막에 재구성된 $S(x)$에서 몇 명의 사용자가 비트 정보를 1로 삽입을 하였는가 알 수 있기 때문에 몇 명의 사용자가 공모하거나 또는 최악의 경우 공모를 하지 않아도 각 사용자들의 비트 정보가 드러나는 문제점이 존재한다.

다음 절에서는 이러한 문제점을 보완하여 한 번 연산 시에 멀티 비트의 내적연산 결과를 알 수 있고, 연산 결과 값에서 어떠한 프라이버시의 노출이 없는 새로운 기법을 제안한다.

4.2 제안하는 기법

제안하는 기법은 다자간 환경에서 프라이버시 보호 데이터 마이닝을 Shamir의 비밀 분배 기법을 사용하여 각 참여자들이 연산하고자 하는 비트 정보를 개인 다항식의 계수에 넣음으로써 내적연산의 결과를 안전하게 얻을 수 있는 기법이다. 제안하는 기법 역시 n명의 참여자 P_i들은 최소 $n-1$차인 다항식 $q_i(x)$와 pub_i, A_i를 갖는다. 역시 A_i는 벡터 형식으로 P_i의 정보를 담고 있다.

참여자들 중 다항식 생성자, 랜덤 생성자, 결과 배포자 참여자가 존재하는 데 각 참여자들은 다음과 같은 역할을 한다.

- 다항식 생성자 P_1는 참여자 P_i가 갖고 있는 $S(pub_i)$를 모두 받아서 내적연산 결과를 내포하고 있는 다항식 $S(x)=s_0+s_1x+\cdots+s_tx^t$를 재구성한다. 그리고 $S(x)$의 계수들을 임의의 순서로 섞어서 결과 배포자에게 전달한다.

- 랜덤 생성자 P_2는 $t+1$개의 임의의 값 $(r_0, r_1, \cdots, r_t)$를 처음 설정했던 다항식 $q_2(x)=a_0+a_1x+a_2x^2+\cdots+a_tx^t$의 계수에 더하여 새로운 다항식 $q_2'(x)=(a_0+r_0)+\cdots+(a_t+r_t)x^t$를 본인의 개인 다항식으로 재설정한다. 그 후 선택했던 임의의 값 $(r_0, r_1, \cdots, r_t)$의 순서도 임의의로 바꾼 후에 결과 배포자에게 전달한다.

- 결과 배포자 P_3는 다항식 생성자와 랜덤 생성자로부터 받은 두 개의 데이터 집합의 교집합의 원소 개수를 모든 배포자에게 배포한다.

모든 참여자들은 다음 〔알고리즘 1〕을 수행하여 개인 다항식 $q_i(x)$를 생성하고 $S(pub_i)$를 얻는다.

〔표 3〕 알고리즘 1

1: for each party P_i $(i=1,\cdots,n)$ do
2: Select a random polynomial
$q_i(x)=a_ix^t+B_{it}x^{t-1}+\cdots+B_{i2}x+B_{i1}$
3: Compute the share of each party P_t
4: for $t=1$ to n do
5: Send $q_i(pub_t)$ to party P_t
6: Receive the $q_i(pub_t)$ from every party P_t
7: Compute
$S(pub_i)=q_1(pub_i)+q_2(pub_i)+\cdots+q_n(pub_i)$
8: Send $S(pub_i)$ to party P_1

〔알고리즘 1〕의 2번째 단계에서 P_2는 0이 아닌 임의의 값 $(r_0, \cdots, r_t)$을 $q_2(x)$의 계수에 더하여 $q_2'(x)$를 만든다.

이 알고리즘을 수행한 후에는 P_1은 모든 참여자들에게서 받은 $S(pub_i)$를 이용하여 $S(x)$를 재구성한다. 그리고 $S(x)$의 계수의 순서를 섞어서 P_3에게 보낸다. P_2는 $(r_0, \cdots, r_t)$의 순서를 임의로 섞은 다음 P_3에게 보낸다. P_3는 P_1과 P_2로부터 받은 값들 간에 일치하는 값을 카운트하여 그 결과를 모든 참여자에게 배포한다. 배포된 결과가 내적연산의 결과가 된다.

〔알고리즘 1〕의 2번째 단계에서 다항식을 생성할 때 각 계수에 들어가는 B_{ij} 값은 다음과 같이 정의 한다.

$$\begin{cases} B_{ij}=0, & where\ A_{ij}=1 \\ B_{ij}=random, & where\ A_{ij}=0 \end{cases}$$

예를 들어 참여자 P_1, P_2, P_3, P_4가 있고, 각 참여자는 $A_1=(1,1,1)$, $A_2=(0,0,1)$, $A_3=(1,0,1)$, $A_4=(0,1,1)$을 갖고 있다고 가정하자. 먼저 개인 다항식의 차수를 $t=3$으로 정하고 각 참여자들의 $pub_1=2$, $pub_2=3$, $pub_3=5$, $pub_4=1$ 값을 공개한다. 각 참여자는 다음과 같은 개인 다항식을 만든다.

$q_1(x) = 2x^3$
$q_2(x) = x^3 + 7x^2 + 3x$
$q_3(x) = 4x^3 + 5x$
$q_4(x) = 2x^3 + 3x^2$

P_2는 임의의 값 $(1,1,5,3)$을 더하여 다음과 같은 $q_2'(x)$를 만든다.

$q_2'(x) = 2x^3 + 8x^2 + 8x + 3$

P_1은 다음 값을 계산하여 해당 pub_i를 갖는 P_i에서 보낸다.

$q_1(pub_1) = q_1(2) = 16$
$q_1(pub_2) = q_1(3) = 54$
$q_1(pub_3) = q_1(5) = 250$
$q_1(pub_4) = q_1(1) = 2$

이와 비슷하게 나머지 P_2, P_3, P_4도 다음 값을 다른 참여자들과 공유한다.

$q_2'(2) = 67,\ q_2'(3) = 153,\ q_2'(5) = 493,\ q_2'(1) = 21$
$q_3(2) = 42,\ q_3(3) = 123,\ q_3(5) = 525,\ q_3(1) = 9$
$q_4(2) = 28,\ q_4(3) = 81,\ q_4(5) = 325,\ q_4(1) = 5$

P_i는 $S(pub_i) = q_1(pub_i) + \cdots + q_4(pub_i)$ 값을 P_1에게 보낸다. P_1은
$S(x) = q_1(x) + q_2'(x) + q_3(x) + q_4(x) = s_3x^3 + s_2x^2 + s_1x + s_0$
에 $(2,3,5,1)$를 대입한 값을 알고 있기에 다음과 같은 연립방정식을 세울 수 있다.

$$\begin{cases} 8s_3 + 4s_2 + 2s_1 + s_0 = 153 \\ 27s_3 + 9s_2 + 3s_1 + s_0 = 411 \\ 125s_3 + 25s_2 + 5s_1 + s_0 = 1593 \\ s_3 + s_2 + s_1 + s_0 = 37 \end{cases}$$

이 연립방정식을 풀어냄으로써
$S(x) = 10x^3 + 18x^2 + 16x + 3$을 재구성해낼 수 있게 된다. P_1은 $S(x)$의 계수 집합 $(10, 18, 16, 3)$의 순서를 임의로 섞어 $(3, 10, 16, 18)$을 만들어 P_3에게 보내준다. P_2 역시 $q_2(x)$에 더했던 임의의 값 $(1,1,5,3)$의 순서를 임의로 $(1,3,5,1)$로 바꾸어 P_3에게 보내준다. P_3은 P_1과 P_2로부터 받은 두 개의 값의 공통된 값을 카운트한다. 여기서 P_1에게서 받은 $(3,10,16,18)$과 P_2에게서 받은 $(1,3,5,1)$의 공통된 값은 3 하나이다. 이는 P_1, P_2, P_3, P_4가 갖고 있는 데이터의 내적연산 결과로서 하나의 속성에서만 모든 값이 1을 갖고 있다는 의미가 된다. 실제로 A_1, A_2, A_3, A_4를 확인해 보았을 때 3번째 값이 1로 일치함을 확인할 수 있다. 마지막으로 P_3가 카운트 된 값 1을 P_1, P_2, P_4에게 보냄으로써 마이닝 결과를 공유하게 된다.

참여자들의 개인 다항식을 구성할 때 A_{ij} 값이 1이면 계수에 0을 넣었다. 그렇기 때문에 참여자 모두가 j번째에 1이라는 공통된 비트를 입력하였으면 P_1이 재구성한 $S(x)$의 해당 계수에는 P_2의 임의의 값이 들어가 있게 된다. P_3가 결과를 생성할 때 그 결과가 j번째의 내적연산 결과 값이라는 것을 숨기기 위해 P_1과 P_2는 보내는 값의 순서를 임의로 섞는다.

각 참여자들의 다항식 $q_i(x)$은 $n-1$차 이상의 다항식이어야 한다. 만일 다항식이 $n-1$차 다항식보다 차수가 낮으면 n명보다 적은 참여자의 공모로 $S(x)$가 재구성될 수 있으며, 따라서 공모하지 않은 참여자 P_i의 A_i가 노출되는 문제가 발생하게 된다. 그렇기에 개인 다항식은 항상 $n-1$차 이상의 차수를 유지하는 다항식을 선택해야 한다. $q_i(x)$가 더 많은 비트를 한 번에 연산하기 위해서 $n-1$보다 높은 차수 일 때에는 몇 명의 참여자 P_i가 한 개 이상의 pub_i를 가짐으로써 P_1이 $S(x)$를 재구성하기에 충분한 값을 가질 수 있도록 한다. P_1이 $S(x)$를 재구성하기에 충분하려면 $t+1$개의 pub_i 값을 n명의 참여자가 나누어 가져야 한다.

V. 분 석

본 논문에서 제안하는 기법은 [11]이 제안하는 기법에서 사용했던 Shamir의 비밀 분배 기법을 사용한다. [11]은 참여자들의 A_i 중에서 j번째 비트를 $q_i(x)$의 상수항에 그대로 사용한다. 그러나 본 논문에서 제안하는 기법은 [11]이 제안했던 기법을 그대로 따르지만 참여자들의 A_i의 모든 정보를 한 번에 $q_i(x)$의 모든 계수에 담는다. 내적연산을 하는 과정에서 [11]에서 나타났던 프라이버시 침해를 방지하기 위해 P_2가 자신의 개인 다항식에 임의의 값을 넣음으로써 기존의 문제점을 보안하였다.

기존에 제안된 다자간 프라이버시 보호 데이터 마이닝 기법들과 제안하는 기법을 [표 4]와 같이 비교하였다.

[11]과 [27]는 Shamir 비밀 분배 기법을 사용하

〔표 4〕 비교 분석

	통신 복잡도	연산 비트	공모 안전성
〔11〕	$O(nk)$	1	X
〔27〕	$O(n2^{k-1}+nk2^{k-2})$	k	X
제안하는 기법	$O(nk)$	k	△

n : 참여자 수
k : 비트 수

여 다자간 환경에서 적용 가능한 프라이버시 보호 데이터 마이닝 기법들을 제안하였다. 하지만 〔11〕은 한 번 마이닝을 할 때 단일 비트만 가능하였고, 〔27〕는 멀티 비트 연산이 가능하였지만 높은 계산 복잡도가 요구되었다. 이렇듯 기존의 기법은 참여자가 많아지거나 연산하려는 정보의 크기가 커짐에 따라 효율성이 많이 떨어지는 한계점이 존재하였다. 또한 참여자들 간의 공모에 안전하지 못한 단점을 갖고 있었다. 〔27〕은 〔29〕를 확장한 기법이지만 〔14〕에서 〔29〕의 기법이 안전성에 문제가 있음을 제기하였다. 기존 논문들이 다자간의 환경에서 프라이버시 보호 데이터 마이닝을 수행한다는 점에서 이러한 공모 안전성이 낮은 것은 치명적인 문제이다. 특히 〔11〕은 참여자들이 공모하지 않아도 다른 참여자의 데이터 정보가 노출될 수 있는 문제점이 나타났다.

본 논문에서 제안하는 기법은 〔11〕과 같은 Shamir의 비밀 분배 기법을 사용하여 외부의 공격자들로부터 안전함이 증명된다. 외부 공격자들은 참여자들 간에 주고받는 값들은 알 수 있지만, 각 참여자의 공개값을 대입한 개인 다항식 값은 모르기 때문에 그 정보들만으로는 내적 연산의 결과 값은 알 수 있겠지만, 유용한 정보인 참여자들 각자의 개인 다항식과 결과 다항식을 재구성할 수 없다. 또한 멀티 비트 연산이 가능하고 암호학적인 기법이나 SMC을 사용하지 않았기에 효율적인 계산을 할 수 있다. 하지만 본 논문의 기법은 공모 안전성에 대해서는 완전한 안전성을 제공하지 못한다. P_1, P_2, P_3가 서로 공모하지 않아야 한다는 가정을 기반으로 해야 하기 때문이다. 하지만 이러한 가정을 기반으로 한다면 P_1, P_2, P_3 중 한 명과 나머지 참가자들의 공모에는 안전함이 증명된다.

VI. 결 론

본 논문에서는 다자간 프라이버시 보호 데이터 마이닝 기법에 적합한 내적연산 기법을 제안하였다. 제안하는 기법을 통해 참여자들은 멀티 비트 내적연산이 가능하게 되었다. 이는 데이터가 점점 대용량화 되어가는 현재에 한 번의 연산으로 많은 정보를 처리할 수 있게 하여 많은 비용 절감을 가져온다. 많은 계산을 필요로 하는 암호학적인 기법을 사용하지 않고 단순 다항식의 연립방정식으로 원하는 내적연산 결과를 얻을 수 있기에 효율성이 더 높아졌다. 또한, 기존 연구에서 드러났던 참여자들의 공모에 의해 데이터의 프라이버시가 노출되는 문제점을 해결하였다. 하지만 완전한 문제 해결이 된 것은 아니기에 앞으로 이 부분에 대한 연구가 조금 더 진행되어야 할 것이다.

참고문헌

〔1〕 D. Agrawal, and C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," The 20th ACM Symposium on Principles of Database Systems,Santa Barbara, CA, pp. 247-255, May, 2001.

〔2〕 R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 207-216, May, 1993.

〔3〕 R. Agrawal, and R. Srikant, "Privacy-preserving data mining," The 19th ACM SIGMOD Conference on Management of Data, pp. 439-450, May, 2000.

〔4〕 C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," Proceeding of the National Science Foundation Workshop on Next Generation Data Mining, pp. 126-133, Nov, 2002.

〔5〕 D. W. Cheung, J. Han, V. T. Ng, A. W. Fu, and Y. Fu, "A fast distributed algorithm for mining association rules," In Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems, Dec, 1996.

〔6〕 E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino, "Hiding

association rules by using confidence and support," In Proceedings of the 4th Information Hiding Workshop, pp. 369-383, Apr, 2001.

〔7〕 W. Du, and M. Atallah, "Privacy-preserving cooperative statistical analysis," In Proceeding of the 17th Annual Computer Security Applications Conference, pp. 102-110, Dec, 2001.

〔8〕 W. Du, and Z. Zhan, "Building decision tree classifier on private data," IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, vol. 14, pp. 1-8, Dec, 2002.

〔9〕 F. Emekci, O. D. Sahin, D. Agrawal and A. El Abbadi, "Privacy preserving decision tree learning over multiple parties," Data and Knowledge Engineering, vol. 63, pp. 348-361, Oct, 2007.

〔10〕 M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, pp.226-231, Aug, 1996.

〔11〕 X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," Software Engineering and Data Mining(SEDM), 2010 2nd International Conference, pp. 345-350, Jun, 2010.

〔12〕 X. Ge, and J. Zhu, New Fundamental Technologies in Data Mining, InTech, Jun, 2011.

〔13〕 B. Gilburd, A. Schuster, and R. Wolff, "k-TTP: A new privacy model for largescale distributed environments," Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 563-568, Aug, 2004.

〔14〕 B. Geothals, S. Laur, H. Lipmaa, and T. Mielikäinen, "On private scalar product computation for privacy-preserving data mining," ICISC 2004, LNCS 3506, pp. 104-120, Dec, 2005.

〔15〕 J. Han, and M. Kamber, "Data Mining: concepts and techniques," Morgan Kaufmann Publishers, 2001.

〔16〕 A. Inan, S. V. Kaya, Y. Saygin, E. Savas, A. A. Hintoglu, and A. Levi, "Privacy preserving clustering on horizontally partitioned data," The Data and Knowledge Engineering (DKE), vol. 63, no. 3, pp. 646-666, Dec, 2007.

〔17〕 J. Kim, and W. Winkler, "Multiplicative noise for masking continuous data," Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C, Apr, 2003.

〔18〕 X. B. Li, and S. Sarkar, "A tree-based data perturbation approach for privacy-preserving data mining," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 9, pp. 1278-1283, Sep, 2006.

〔19〕 K. Liu, C. Giannella, and H. Kargupta, "An attacker's view of distance preserving maps for privacy-preserving data mining," PKDD, pp. 297-308, Sep, 2006.

〔20〕 Y. Lindell, and B. Pinkas, "Secure multiparty computation for privacy-preserving data mining," Journal of Privacy and Confidentiality, vol. 1(1), pp. 59-98, Apr, 2009.

〔21〕 K. Muralidhar, and R. Sarathy, "Data shuffling a new masking approach for numerical data," Management Science, vol. 52, no. 5, pp. 658-670, May, 2006.

〔22〕 G. Nayak, and S. Devi, "A survey on privacy preserving data mining: approaches and techniques," International Journal of Engineering Science and Technology (IJEST), vol. 3(3), pp. 2127-2133, Mar, 2011.

〔23〕 S. T. M. Oliveira, and O. R. Zaiane,

"Privacy preserving frequent itemset mining," In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, pp. 43-54, Dec, 2002.

〔24〕 M. O. Rabin, "How to exchange secrets by oblivious transfer," Technical Report TR-81, Aiken Computation Laboratory, May, 1981.

〔25〕 A. Shamir, "How to share a secret," Communications of the ACM, vol.22(11), pp. 612-613, Nov, 1979.

〔26〕 L. Sweeney, "k-Anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557-570, Oct, 2002.

〔27〕 D. Trinca, and S. Rajasekaran, "Towards a collusion-resistant algebraic multi-party protocol for privacy-preserving association rule mining in vertically partitioned data," Performance, Computing, and Communications Conference, IPCCC 2007, IEEE International, pp. 402-409, Apr, 2007.

〔28〕 Pang-Ning Tan, M. Steinbach, and V. Kumar, Introduction to data mining, Pearson, Mar, 2006.

〔29〕 J. Vaidya, and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," In The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644, Jul, 2002.

〔30〕 J. Vaidya, and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data," In The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 206-215, Aug, 2003.

〔31〕 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," SIGMOD Record, vol. 33, no. 1, pp. 50-57, Mar, 2004.

〔32〕 A. C. C. Yao, "How to generate and exchange secrets," Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, pp. 162-167, Oct, 1986.

〈著者紹介〉

이 상 훈 (Sang Hoon Lee) 학생회원
2010년 2월: 서울시립대학교 수학과 졸업
2010년 3월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 프라이버시향상기술(PET), 데이터베이스 보안, 빅데이터 보안

김 기 성 (Kee Sung Kim) 학생회원
2008년 2월: 서울시립대학교 수학과 졸업
2011년 2월: 고려대학교 정보보호대학원 석사 졸업
2011년 3월~현재: 고려대학교 정보보호대학원 박사과정
〈관심분야〉 프라이버시향상기술(PET), 데이터베이스 보안, 암호 이론

정 익 래 (Ik Rae Jeong) 정회원
1998년 2월: 고려대학교 전산학과 학사 졸업
2000년 2월: 고려대학교 전산학과 석사 졸업
2004년 8월: 고려대학교 정보보호대학원 박사 졸업
2006년 6월~2008년 2월: 한국전자통신연구원 암호기술연구팀 선임연구원
2008년 3월~2011년 8월: 고려대학교 정보경영공학전문대학원 조교수
2011년 9월~현재: 고려대학교 정보보호대학원 부교수
〈관심분야〉 프라이버시향상기술(PET), 데이터베이스 보안, 암호 이론