연관규칙 마이닝에서 랜덤화를 이용한 프라이버시 보호 기법에 관한 연구

강 주 성[†] · 조 성 훈[††] · 이 옥 연[†††] · 홍 도 원[††††]

요 약

본 논문에서는 랜덤화 기법을 이용한 프라이버시 보존형 데이터 마이닝(PPDM) 기술에 대하여 논한다. 계산 효율성 때문에 실용화 되지 못하고 있는 안전한 다자간 계산(SMC) 기반 PPDM은 현재의 컴퓨팅 환경에서는 실용성 없는 다분히 이론적인 것이다. 그래서 우리는 실용적인 PPDM 기술에 집중하여 가장 널리 사용되고 있는 랜덤화 기법에 대한 연구 결과를 소개한다. 특히, 랜덤화를 이용한 실용적인 PPDM 분야에서 가장 중요한 프라이버시 측도 개념을 심도 있게 분석하였으며, 연관규칙 마이닝에서의 프라이버시 보호 기술에 초점을 맞춘다. Evfimievski 등이 제안한 select-a-size 범주에 속하는 새로운 랜덤화 작용소인 binomial-selector 개념을 제안하고, 적절한 파라미터를 찾기 위한 시뮬레이션 결과를 제시한다. 기존의 cut-and-paste 랜덤화 작용소는 아이템 집합이 큰 경우에는 매우 비효율적이며 복원된 지지도의 분산이 크다는 단점을 지니고 있다. 여기에서 제안하는 binomial-selector 랜덤화 작용소는 cut-and-paste 작용소가 갖는 단점들을 보완한다.

키워드 : 데이터마이닝, PPDM, 랜덤화, 연관규칙마이닝, 프라이버시 측도

On the Privacy Preserving Mining Association Rules by using Randomization

Ju-Sung Kang[†] · Sunghoon Cho[††] · Ok-Yeon Yi[†††] · Dowon Hong[††††]

ABSTRACT

We study on the privacy preserving data mining, PPDM for short, by using randomization. The theoretical PPDM based on the secure multi-party computation techniques is not practical for its computational inefficiency. So we concentrate on a practical PPDM, especially randomization technique. We survey various privacy measures and study on the privacy preserving mining of association rules by using randomization. We propose a new randomization operator, binomial selector, for privacy preserving technique of association rule mining. A binomial selector is a special case of a select-a-size operator by Evfimievski et al.[3]. Moreover we present some simulation results of detecting an appropriate parameter for a binomial selector. The randomization by a so-called cut-and-paste method in [3] is not efficient and has high variances on recovered support values for large item-sets. Our randomization by a binomial selector make up for this defects of cut-and-paste method.

Key Words : Data mining, PPDM, Randomization, Association rule mining, Privacy measure

1. 서 론

데이터 마이닝(data mining)은 많은 양의 데이터에 함축적으로 들어있는 지식이나 패턴을 찾아내는 기술이다. 인터넷과 전자상거래가 급속도로 보급되면서 소비자와 구매에 관련된 많은 양의 데이터가 자동으로 컴퓨터에 모이게 되었다. 이로 인해 과거에는 가능하지 않았던 거대한 양의 데이터를 우리 주변에서 쉽게 볼 수 있는 시대가 되었다. 하지만 이렇게 모아 놓은 데이터로부터 유용한 정보를 찾아내어 마케팅이나 회사의 이익을 효율적으로 증대시키기 위해서 사용하는 데는 아직 많은 어려움이 남아 있다. 데이터 마이닝 기술은 이러한 대용량의 데이터로부터 유용하고 값진 정보를 효율적으로 찾아내어 회사뿐만 아니라 개인의 일상생활에도 편리하게 도움을 줄 수 있다.

한편, 데이터 마이닝 기술의 유용성 이면에는 프라이버시 문제가 존재한다. 데이터를 모으고 이를 여러 가지 방법으로 분석하는 과정에서 프라이버시와 관련된 문제가 자연스

※ 본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음 (2005-Y001-04, 차세대 시큐리티 기술 개발).
† 정 회 원 : 국민대학교 수학과 부교수
†† 정 회 원 : 누리솔루션 엔지니어
††† 정 회 원 : 국민대학교 수학과 조교수
†††† 정 회 원 : 한국전자통신연구원 선임연구원
논문접수 : 2006년 12월 28일, 심사완료 : 2007년 5월 28일

럽게 대두된다. 기밀성을 요하는 데이터를 보호하고자 하는 욕구는 단지 개인의 문제만이 아니다. 경쟁 관계에 있는 회사들이 서로의 이윤 추구를 위해서 협력하는 경우에도 개별 회사의 중요 정보 노출은 꺼리게 된다. 국가 간의 협력을 도모하는 경우에도 이러한 문제는 여전히 중요한 이슈이다. 위와 같은 문제는 정보를 공유하는 것과 프라이버시를 유지하고자 하는 것의 취사선택(trade-off) 문제이다. 여기에서는 이와 같은 취사선택 문제를 해결하기 위한 기술적 관점의 연구 결과를 소개한다. 우리는 데이터 마이닝에서 프라이버시를 보호하기 위한 기술을 간단히 PPDM (Privacy-Preserving Data Mining)이라 부른다[1].

현재 상용 데이터 마이닝 소프트웨어에서 제공되는 알고리즘 중에서 최근 개발된 중요한 기술로는 연관 규칙(association rules), 분류(classification), 순차 패턴(sequential patterns), 군집화(clustering), 아웃라이어 판별(outlier discovery) 등이 있다. 특히, 연관 규칙은 데이터 마이닝을 소개할 때 대표적으로 언급되는 기술로서 백화점이나 슈퍼마켓에서 구매한 물건들에 관한 연관 규칙을 찾아내는 기술이다. 실제 데이터를 이용해 발견되었던 유명한 연관 규칙 중 하나는 미국의 대형 편의점에 연관 규칙 기술을 적용한 결과 일회용 기저귀를 사는 사람은 맥주도 같이 산다는 연관 규칙이다.

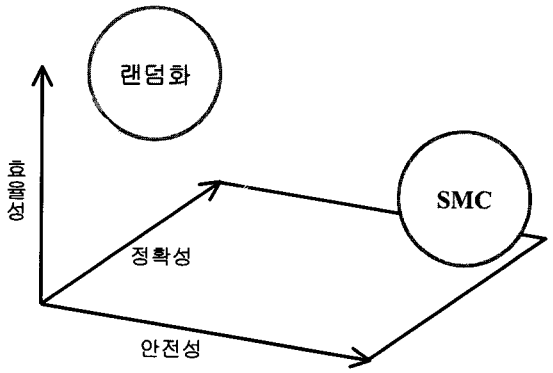
본 논문에서는 실용적인 PPDM 기술 중에서 가장 활발히 연구 되고 있는 분야인 연관규칙 마이닝 관련 프라이버시 보호 기술에 대하여 논한다. 최근의 연구 결과 중에서 주목할 만한 Evfimievski-Srikant-Agrawal-Gehrke[2,3]의 연구 결과를 면밀히 분석하여 보다 실용적인 기법을 제안한다. 그리고 제안한 기법에서 최적의 파라미터를 찾기 위한 시뮬레이션 결과도 보여준다.

논문의 구성은 다음과 같다. 2절에서는 프라이버시 보존형 데이터 마이닝 기술에 대한 전반적인 사항을 설명한다. 3절에서는 랜덤화 기법의 이론적 배경을 설명하고, 4절에서는 프라이버시를 측정하기 위한 여러 가지 측도에 대해서 논한다. 5절에서는 연관규칙 마이닝에 사용되는 프라이버시 보호 기술에 대해서 소개하고, 6절에서는 새롭게 제안하는 랜덤화 기법과 시뮬레이션 결과를 기술하며, 7절은 결론부이다.

2. PPDM 개요

PPDM 기술은 각 개체들이 소유한 데이터를 다른 개체들에게 노출시키지 않은 상태에서 여러 개체들의 결합된 데이터 집합으로부터 데이터 마이닝 관련 계산을 수행하는 것이다.

PPDM 관련 연구는 크게 두 가지로 대별된다. 먼저 프라이버시를 보존하는 통계적 데이터베이스로 종종 언급되는 것으로 데이터가 데이터마이닝으로 사용되기 이전에 변환하는 기술이다. 원래의 데이터에 노이즈를 더해주거나 다른 종류의 랜덤화(randomization)를 적용시키는 것이 이 방법의 예이다. 이 방법은 실용적으로 다양한 통계적 데이터를 위



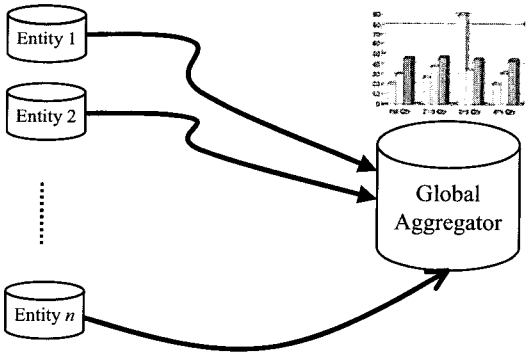
(그림 1) 두 가지 PPDM 방법의 효율성과 정확성

해서 널리 사용되었으나, 높은 안전성을 요하는 응용에는 적절하지 못하다. 랜덤화에 의한 프라이버시 보호 방법은 암호적 방법에 비해 안전성이 떨어지지만 매우 효율적이어서 실용화 되어 있다. 하지만 오리지널 데이터의 변형에 기인한 데이터 마이닝 결과의 정확성(accuracy)은 해결되어야 할 중요한 문제로 남는다.

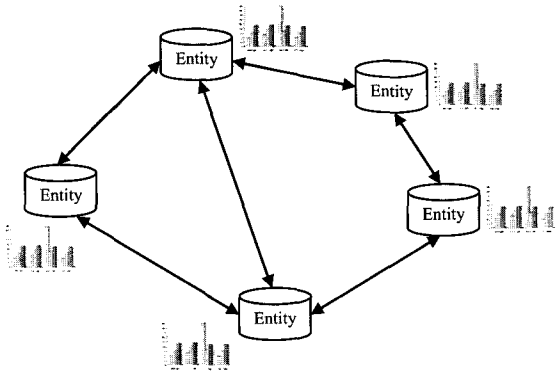
두 번째 방법은 데이터 마이닝에 암호 원천 기술이라 할 수 있는 SMC(secure multiparty computation)[4] 기술이 적용된 것으로, 이 경우의 모든 개체는 자신의 입력과 계산 결과 이외에는 어떠한 정보도 얻을 수 없다. SMC를 사용한 PPDM은 데이터 변형이 전혀 없는 것으로 가정되기 때문에 데이터 송신 전 단계에서 데이터를 변형시키는 첫 번째 방법에서 발생할 수 있는 정확성(accuracy) 문제는 발생하지 않는다. 그러나 SMC 기반 PPDM 기술은 계산 효율성이 매우 낮기 때문에 아직까지 실용적이지 못하다는 한계를 지닌다. (그림 1)은 랜덤화 기법을 기반으로 한 PPDM 기술과 SMC를 기반으로 한 PPDM 기술을 계산 효율성과 안전성 및 정확성 관점에서 비교한 일반적인 결과이다[1].

2.1 PPDM을 위한 시스템 구조

데이터 마이닝에서 프라이버시를 보호하기 위한 기술은 데이터 마이닝 시스템 구조에 따라서 달라진다. 이는 시스템 내의 각 개체들은 데이터 제공자(data provider) 역할만 수행하고, 중앙의 서버에서만 데이터 마이닝을 수행하는 경



(그림 2) 중앙집중식 시스템에서의 데이터 마이닝



(그림 3) 분산형 시스템에서의 데이터 마이닝

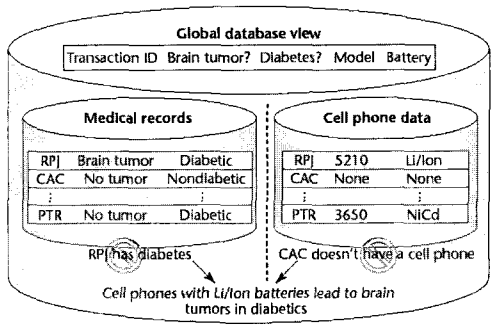
우에 사용되는 프라이버시 보호 기술과 각 개체에서도 독립적인 데이터 마이닝이 가능한 분산형 시스템에서의 프라이버시 보호 기술이 다르다는 의미이다. 중앙 집중식 시스템 구조에서의 데이터 마이닝은 (그림 2)에, 분산형 시스템 구조를 갖는 데이터 마이닝 모델은 (그림 3)에 각각 나타나 있다.

중앙 집중식 시스템 구조를 갖는 경우에 만족하고자 하는 프라이버시 관점의 목표와 분산형 시스템 구조를 갖는 경우에 프라이버시 관점에서 얻고자 하는 목표는 다를 수밖에 없다. 실용화 되어 있는 대부분의 시스템은 중앙 집중식이다. 중앙 집중식에서 서버가 TTP 역할을 수행한다면 프라이버시 문제는 간단히 해결된다. 하지만 현재의 인터넷 환경은 대부분의 중앙 서버에 대해서 신뢰감을 갖지 못하게 만들고 있다. 실제로 전자상거래의 경우 서버 역할을 수행하는 전자상거래 업체가 물건 대금만을 챙기고 난 후에 해당 사이트를 없애버리는 경우, 또는 회원들의 개인 정보를 서버가 팔아넘기는 행위 등이 빈번히 발생하고 있는 상황이다. 그러므로 중앙 집중식 네트워크 환경에서도 서버를 신뢰하지 않는다는 가정 하에서의 프라이버시 보호 기술이 필요하다.

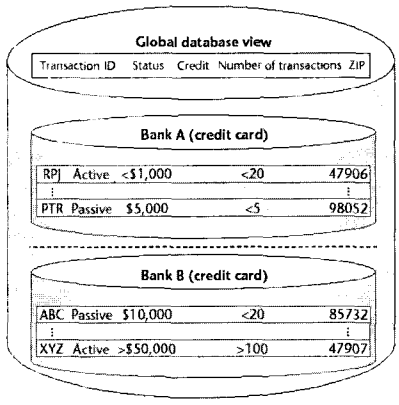
분산형 데이터 마이닝 시스템 구조에서 요구되는 프라이버시 관련 목표는 중앙 집중식 시스템에 비해서 상당히 복잡하다. 일반적으로 중앙 집중식 시스템은 하나의 서버가 수백만 이상의 개체를 거느릴 수 있는 큰 스케일인 반면, 분산형 시스템은 회사 간 거래, 정부 간 거래, 정부와 회사 간 거래 등과 같이 참여하는 개체 수 입장에서는 상대적으로 작은 스케일이다. 그러므로 효율성 높은 실용적인 프라이버시 보호 기술이 중앙 집중식 데이터 마이닝 시스템에서 요구되는 반면, 분산형 데이터 마이닝 시스템에서는 안전성이 높은 프라이버시 보호 기술이 요구된다고 볼 수 있다.

2.2 PPDM을 위한 데이터 베이스 구조

프라이버시를 보존하는 데이터 마이닝 기술에서 데이터베이스의 형태는 적용 기술을 달리해야 하는 중요한 지표가 된다. 데이터베이스의 형태는 수직적으로 분할된 데이터베이스(vertically partitioned database)와 수평적으로 분할된 데이터베이스(horizontally partitioned database)로 구분된다.



(그림 4) 수직 분할된 데이터베이스의 예



(그림 5) 수평 분할된 데이터베이스의 예

수직 분할된 데이터베이스의 예는 (그림 4)에 나타나 있고, 수평 분할된 데이터베이스의 예는 (그림 5)에 나타나 있다. 두 그림은 모두 참고문헌 [1]에서 인용한 것이다.

수직적으로 분할된 데이터베이스는 개체별로 서로 다른 형태의 데이터를 저장하고 있을 수 있기 때문에 이러한 환경이 고려된 프라이버시 보호 기술이 적용되어야 한다. 반면에 수평적으로 분할된 데이터베이스는 모든 개체가 같은 형태의 데이터를 가지고 있기 때문에 각 개체의 프라이버시를 보호하는 기술이 일률적으로 적용 가능하다. 수직 분할된 데이터베이스에서의 프라이버시 보호 기술은 안전성이 우선적으로 고려되는 분산형 시스템에 적용 가능한 PPDM 기술이 적합하다. 현재 가장 실용적인 PPDM 모델은 중앙 집중식 네트워크 구조 하에서 각 개체가 수평적으로 분할된 데이터를 소유하고 있는 것이 일반적이다.

2.3. 실용적인 PPDM 기술 개요

PPDM의 효시가 된 연구 결과는 IBM Almaden 연구소의 두 연구원인 Agrawal-Srikant[5]가 제시한 방법이다. 이들의 연구 결과를 기점으로 하여 랜덤화 기법을 이용한 여러 가지 PPDM 기술이 발표되었다. 본 절에서는 이를 바탕으로 실용적인 PPDM 기술의 세부 사항을 논하도록 한다.

SMC 기반 PPDM은 Lindell-Pinkas[6]의 연구 결과를 대표적인 것으로 들 수 있다. Lindell과 Pinkas의 결과는 포괄

적인 SMC(General SMC)로 해결 가능한 문제를 데이터 마이닝이라는 응용 관점에서 포괄적 SMC보다 효율적으로 계산할 수 있는 프로토콜을 제안한 것이다. 특히, 분류 문제에 집중하여 널리 사용되는 알고리즘인 ID3[7,8] 알고리즘을 효율적으로 수행할 수 있는 방식을 제안하였다. 그러나 이들이 제안한 프로토콜은 OT(Oblivious Transfer) 프로토콜에 기반한 것으로 포괄적 SMC보다는 매우 효율적인 것이지만 여전히 실용적이지 못하다. 그러므로 여기에서는 SMC에 기반한 PPDM 기술에 대해서 세부적인 내용을 기술하지 않기로 한다.

프라이버시를 보존하기 위한 기본적인 가정은 사용자들이 민감한 속성(sensitive attribute)에 대해서 수정된 값을 제공한다는 것이다. Agrawal-Srikant[5]에 언급되어 있는 데이터 값을 수정하는 두 가지 방법을 먼저 고려해보자.

먼저 Value-Class 멤버쉽 방법이다. 이 방법에서는 속성 값을 서로 배반인 집합으로 분할한다. 가장 손쉬운 방법은 속성 값들을 구간(interval)별로 이산화(discretization)하는 것이다. 이 때 모든 구간이 동일한 길이일 필요는 없다. 예를 들면, 연봉을 10K 단위의 구간으로 나눌 수 있고 상한선을 50K로 하여 마지막 구간은 50K 이상을 나타내는 구간으로 나누는 것이다. 사용자는 속성의 참값을 입력하는 것이 아니라 해당 구간만을 입력하기 때문에 본인 연봉에 대한 참값은 보호될 수 있는 것이다. 통계적 자료에서 이산화는 개인 정보를 숨기기 위해서 가장 널리 사용되는 방법이다.

다음으로 Value Distortion 방법이다. 데이터의 참값이 x_i인 경우에 어떤 분포로부터 랜덤하게 r을 선택하여 $x_i + r$을 보내는 방법이다. 사용되는 랜덤 분포로는 균일분포(uniform distribution)와 정규분포(Gaussian distribution)가 널리 사용된다. 균일 분포는 구간 $[-\alpha, \alpha]$에서 r을 임의로 선택하는 것으로 확률변수의 평균은 0이 된다. 정규분포는 평균이 $\mu = 0$이고 표준편차가 σ인 확률변수가 사용된다.

위 두 가지 방법에서 얻고자하는 프라이버시를 정량화(quantify)하기 위해서 Agrawal과 Srikant(2000)는 신뢰구간(confidence interval)의 길이를 사용하였다. 신뢰수준 c%의 신뢰구간의 길이는 이산화의 경우 구간의 길이를 W라고 할 때, $(c\%) \times W$가 되고, 균일분포인 경우는 $(c\%) \times 2\alpha$가 되며, 정규분포인 경우에는 신뢰수준에 해당하는 분위수에 표준편차 σ를 곱하는 값으로 정의된다. 신뢰구간 개념을 이용한 프라이버시의 정량화 노력은 어떤 측면에서 합리적으로 보일 수 있다. 이산화의 경우 구간 길이가 커질수록 참값의 범위는 넓어지므로 상대적인 프라이버시는 증가한다고 볼 수 있는 것이다. 그러나 이 신뢰구간 개념을 이용한 프라이버시의 정량화 방법은 다른 PPDM 기술에 적용하는 데 무리가 따른다는 사실이 차후에 밝혀졌다. 이에 대해서는 뒤에서 자세히 언급하기로 한다.

한편, 프라이버시를 보존하기 위해서 오리지널 데이터를 변형할 경우에 이를 토대로 한 데이터 마이닝 결과는 정확성이 떨어질 것이다. 그러므로 대부분의 PPDM 기술에서 랜덤화된 데이터로부터 오리지널 데이터의 분포를 복구해내는 것은 매우 중요한 요소이다.

(1) $f_X^0 :=$ Uniform distribution
(2) $j := 0$ // Iteration number
repeat
(3) $f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z) dz}$
(4) $j := j + 1$
until (stopping criterion met)

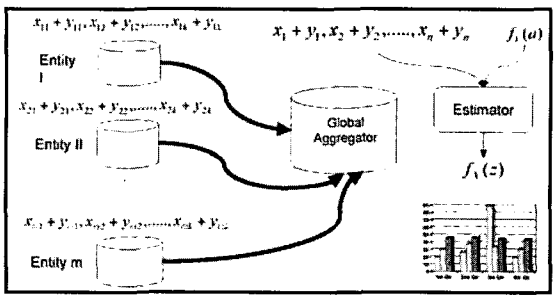
(그림 6) 오리지널 데이터의 분포 추정 알고리즘

오리지널 데이터의 확률분포로부터의 n개 표본을 $x_1, x_2, \cdots, x_n$이라 하면, 이 값들은 서로 독립이고 동일한 분포를 갖는 확률변수 $X_1, X_2, \cdots, X_n$의 실행값(realization)으로 볼 수 있다. 이들과 동일한 분포를 갖는 확률변수를 대표적으로 X라 하자. 원 데이터를 숨기기 위하여 다른 확률변수 Y와 동일한 분포를 갖는 서로 독립인 n개의 확률변수 $Y_1, Y_2, \cdots, Y_n$이 사용된다. y_i를 확률변수 Y_i의 실행값이라 할 때, $x_1 + y_1, \cdots, x_n + y_n$과 Y의 확률밀도함수(probability density function) f_Y가 주어졌을 때, X의 확률밀도함수 f_X를 추정하는 것이 오리지널 데이터를 복구한다는 의미이다. 오리지널 데이터의 실제값을 복원하는 것이 아니라 분포만을 추정하여 원하는 정보를 마이닝 한다는 것이다.

조건부 확률과 관련된 베이즈 공식(Bayes' rule)을 사용하면 f_X와 f_Y 사이의 관계식을 구할 수 있다. 그러나 우리는 f_Y는 알 수 있지만 f_X는 알 수 없다. 그래서 Agrawal-Srikant[5]는 초기의 확률밀도함수 f_X^0를 균일분포(uniform distribution)로 놓고 관계식을 반복적으로 사용함으로써 원하는 밀도함수를 얻는 방법을 제시하였다. 이 과정을 간단히 알고리즘으로 표현하면 (그림 6)과 같고, (그림 7)은 전체적인 과정을 간단히 도식화한 것이다.

3. 랜덤화 기법의 이론적 배경

랜덤화(randomization) 또는 데이터 섭동(data perturbation)이라 불리는 기법은 랜덤한 노이즈(noise)를 사용하여 기밀을 요하는 변인을 수정하여 프라이버시를 보호하고자 하는 방법이다.



(그림 7) 오리지널 데이터의 분포 추정 과정

데이터를 공개하는 정책에 따라서 랜덤화 기법 사용 시 데이터 유용성(data utility)과 노출 위험성(disclosure risk) 사이에는 항상 취사선택(trade-off)의 문제가 존재한다. 아무런 데이터도 공개하지 않으면 당연히 노출 위험성은 전혀 없다. 데이터를 공개하는 정도가 커지면 커질수록 노출 위험성은 증가하지만 데이터 유용성 또한 높아지게 된다. 본 절에서는 랜덤화 기법에서 추구하는 프라이버시의 개념을 확률론적 관점에서 살펴본다.

데이터 집합이 기밀성을 요하는 변수(confidential variable) X와 기밀성이 요구되지 않는 변수(non-confidential variable) S를 모두 포함하고 있다고 하자. 그리고 $f(\cdot)$는 확률밀도 함수를 나타낸 다고 하자. 랜덤화 기법은 다음 두 가지 요구조건 하에 마스킹된 값인 Z를 생성하는 것이다[9].

1. **데이터 유용성 또는 정확성(accuracy) 요구조건 :** Z의 통계적 특성은 X의 통계적 특성과 일치해야 한다. 즉, $f(X)=f(Z)$가 성립해야 한다. 그리고 Z와 S 사이의 관계는 X와 S 사이의 관계와 같아야 한다. 즉, $f(Z,S)=f(X,S)$이 성립해야 한다.

2. **노출 위험성 요구조건 :** X의 기밀성은 유지되어야 하고, 공개된 데이터 (Z,S)는 노출 위험성을 증가시키지 않아야 한다. 달리 표현하면, 다음 조건 $f(X|S,Z)=f(X|S)$이 성립해야 한다.

데이터 유용성과 노출 위험성 요구조건을 만족하는 랜덤화된 데이터를 생성하기 위한 일반적인 과정은 다음과 같다. 먼저 Z가 X와 서로 독립적인 확률변수일 때, $S=s_i$가 주어진 조건 하에서의 조건부 분포인 $f(X|S=s_i)$로부터 관측치(observation) z_i를 생성한다. 그러면

$$z_i \sim f(X|S=s_i),$$
$$f(X,Z|S=s_i)=f(X|S=s_i)f(Z|S=s_i)$$

가 성립한다. 이러한 생성 과정을 데이터 집합에 있는 모든 i에 대해서 반복한다. 이와 같이 생성된 i번째 관측치 z_i의 실제 값은 조건부 분포 $f(X|S=s_i)$로부터 독립된 구현값(realization)이다.

확률론적으로 이상적인 데이터 유용성 요구조건은 주변확률분포(marginal distribution)와 결합확률분포(joint distribution)를 사용한 표현으로는 다음과 같다.

$$f(Z)=f(X),\ \ f(Z,S)=f(X,S).$$

z_i는 $f(X|S)$로부터 생성되기 때문에

$$f(Z|S)=f(X|S)$$

이다. 그러므로

$$f(Z,S)=f(Z|S)f(S)=f(X|S)f(S)=f(X,S)$$

가 성립한다. 더욱이

$$f(Z)=\int_S f(Z,S)\,ds=\int_S f(X,S)\,ds=f(X)$$

도 성립하므로 위에서 언급한 일반적인 랜덤화 기법은 우리가 원하는 데이터 유용성을 만족하게 되는 것이다.

이제 노출 위험성 요구조건에 대해서 살펴보자. 노출 위험성에 관한 정의는 여러 가지가 있다. 먼저 데이터 접근 시의 노출 위험성에 대한 Dalenius[10]의 정의를 살펴보자. 이들의 이상적인 정의에서는 사용자들이 X와 S에 대한 최대 정보를 이미 가지고 있다고 가정한다. 기밀성이 요구되지 않는 데이터인 S에 대한 접근이 가능할 뿐만 아니라 $f(X|S)$를 가지고 있는 것이다. 그러므로 랜덤화된 데이터에 접근하기 이전 공격자(snooper)의 능력은 조건부 분포인 $f(X|S)$로부터 X를 예측하는 능력으로 정의된다.

사용자가 랜덤화된 마이크로 데이터에 대한 접근이 가능한 경우에는 추가적인 정보를 가지고 있으므로 조건부 분포 $f(X|S,Z)$를 이용하여 X를 예측함으로써 예측 능력을 높일 수 있다. 예를 들어 랜덤화의 방법으로 기밀성이 요구되는 정보에 노이즈를 첨가하는 $Z=X+e$ 형태의 변환을 사용했다고 하자. 이러한 경우에 $f(X|S,Z)$와 $f(X|S)$는 서로 다른 수식으로 표현된다는 것이 알려져 있으며, 이를 통하여 공격자의 예측 능력을 개선할 수 있는 것이다.

랜덤화된 데이터가 앞에서 언급한 일반적 과정에 의해서 생성된 경우라면 우리는 $f(X|S,Z)=f(X|S)$를 얻는다. 다시 말해서 랜덤화된 데이터 Z를 갖는다는 것이 X에 대한 정보를 얻어내는 데에 추가적인 정보를 제공해주지 못한다는 것이다. 그러므로 이상적인 방법으로 구현된 랜덤화 기법은 노출 위험성이 없는 것이지만, 실제로 구현된 랜덤화 기법은 노출 위험성이 존재한다는 예를 노이즈 추가 방법에서 찾을 수 있다.

한편, Duncan-Lambert[11]는 불확실성(*uncertainty*)에 기반한 노출 위험성 관련 정의를 제안하였다. 이들이 제안한 정의는 데이터에 접근하기 전과 후에 목표로 하는 예상 분포에 대한 음이 아닌 오목함수(concave function)를 적용함으로써 공격 목표에 대하여 공격자가 얻을 수 있는 지식(knowledge)을 기반으로 한 것이다. 이 경우에도 데이터에 접근하기 전과 후의 분포가 $f(X|S,Z)=f(X|S)$로 동일하다면 공격자가 얻는 지식은 없다. 그러므로 이상적인 방법에 의해서 생성된 랜덤화된 데이터는 위와 같은 여러 가지 의미의 노출 위험성에 대한 안전성을 제공하는 것이다.

4. 프라이버시 관련 측도

실용적인 PPDM 기술에서 고려하는 측도(measure)는 정확성(accuracy) 관련 측도와 프라이버시(privacy) 관련 측도가 있다. 정확성 관련 측도는 프라이버시 보호 기술을 적용

한 이후의 데이터로부터 데이터 마이닝 과정을 실행할 경우에 오리지널 데이터로부터 데이터 마이닝을 실시한 결과와 얼마나 차이가 있는지를 비교 평가하는 것이다. 이는 프라이버시를 보호하기 위한 방법에 따라서 차이가 나고, 각 방법에 따라서 사용되는 측도도 다양하다.

한편, 프라이버시 관련 측도는 프라이버시를 보존하고자 하는 방법을 고안할 당시부터 고려되어야 하는 항목이다. 다시 말해서 어떤 프라이버시 보호 기법을 제안하고자 할 때, 어떤 관점의 측도를 사용하여 우수한 프라이버시 보호 능력을 달성할 것인지가 중요하다. 랜덤화 기법을 적용한 실용적인 PPDM의 경우 대표적인 것이 통계학의 신뢰구간(confidence interval) 개념과 정보이론의 상호정보(mutual information) 개념, 그리고 프라이버시 손상(privacy breach) 개념을 들 수 있다. 이들을 포함한 대표적인 프라이버시 측도를 비교하여 조사 분석해 보자.

4.1 신뢰구간(confidence interval)

처음으로 랜덤화 기법을 이용한 PPDM 기술을 제안했던 Agrawal-Srikant[5]는 프라이버시 측도로 통계학적 개념인 신뢰구간(confidence interval)을 제안하였다. 오리지널 데이터 x_i가 모르는 분포일 때, 랜덤화의 결과값인 $Z_i = x_i + Y_i$의 분포에 대해서 x_i는 미지의 파라미터로 간주된다. $Z_i = z_i$인 값을 서버는 알 수 있기 때문에 서버는 $x_i \in I(z_i)$일 확률이 적어도 c%인 구간 $I(z_i)$를 계산할 수 있다. 이때 이 신뢰구간의 길이를 프라이버시의 측도로 삼았던 것이다.

이 신뢰구간 개념은 직관적으로 프라이버시 측도로서 의미가 있는 것 같지만 일반적인 경우의 척도로 확장 가능하지는 않다는 것이 알려져 있다. 다음과 같은 예에서 이러한 사실은 쉽게 엿볼 수 있다.

속성 X의 확률분포가 다음과 같다고 하자.

$$f_X(x) = \begin{cases} 1/2, & if\ -1 \le x \le 0\ or\ 3 \le x \le 4 \\ 0 & otherwise \end{cases}$$

그리고 랜덤화를 위한 확률변수 Y는 구간 $[-1, 1]$에서 균일하게(uniformly) 분포한다고 하자. 그러면 신뢰구간 측도를 사용할 경우 이 방식의 프라이버시 측도값은 100% 신뢰수준 하에서 2가 된다. 그러나 X의 값이 $[-1, 0] \cup [3, 4]$에 속한다는 사실을 안다고 하면, 신뢰구간의 길이는 2가 아닌 1이 된다. 실제로 계산되는 구간은 다음과 같다.

$$I(z) = \begin{cases} [-1, 0] & if\ -2 \le z \le 1 \\ [3, 4] & if\ 2 \le z \le 5 \end{cases}.$$

즉, z 값을 보고 x 값을 추정할 경우에 $-2 \le z \le 1$라는 사실을 알면 x의 값은 100%의 확률로 구간 $[-1, 0]$에 있고, $2 \le z \le 5$라는 사실을 알면 x는 구간 $[3, 4]$에 있음을 100% 확신할 수 있다. 그러므로 이 경우 신뢰 구간의 길이는 $[-1, 0]$ 또는 $[3, 4]$의 길이인 1이 되는 것이다. 이 외에도 여러 가지 경우에 신뢰 구간의 길이는 실제로 더 떨어질 수 있다. 그러므로 신뢰 구간의 길이를 프라이버시 측도로 삼는 것은 일반적이지 못한 것으로 생각된다.

4.2. 상호정보(mutual information)

Agrawal-Agrawal[12]은 신뢰구간의 길이로 프라이버시를 측정하는 것의 불합리성을 설명하면서 Shannon의 정보이론에 입각한 측도의 도입을 제안한다. 랜덤화 되기 이전의 속성인 X의 정보량 평균은 확률 분포에 의존하고 다음과 같은 미분엔트로피(differential entropy)에 의해서 측정된다.

$$\begin{aligned} h(X) &= E[-\log_2 f_X(X)] \\ &= -\int_{\Omega_X} \log_2 f_X(x)\, f_X(x)\, dx \end{aligned}$$

한편, 랜덤화된 속성 Z를 알고 난 이후의 X의 정보량은 다음과 같다.

$$\begin{aligned} h(X|Z) &= E[-\log_2 f_{X|Z=z}(X)] \\ &= -\int_{\Omega_{X,Z}} \log_2 f_{X|Z=z}(x)\, f_{X,Z}(x,z)\, dx\, dz \end{aligned}$$

그러므로 Z가 노출된 이후에 X의 정보량에 대한 평균 손실은 위 두 엔트로피 값의 차이가 된다. 이 값을 확률변수 X와 Z 사이의 상호정보(mutual information)라 하고 $I(X;Z)$로 표현한다. 즉,

$$\begin{aligned} I(X;Z) &= h(X) - h(X|Z) \\ &= E_{(x,z)\sim(X,Z)}\left[\log_2 \frac{f_{X|Z=z}(X)}{f_X(X)}\right] \end{aligned}$$

이다.

Agrawal-Agrawal[12]은 프라이버시의 양을 측정하는 측도로 $\Pi(X)$를 사용하고, 프라이버시 손실(loss)을 측정하는 측도로 $\wp(X|Z)$를 각각 사용할 것을 주장하였다. 이들의 정의는 다음과 같다.

$$\Pi(X) = 2^{h(X)},\quad \wp(X|Z) = 1 - 2^{-I(X;Z)}$$

앞 소절의 예제에서 이 값들은 다음과 같이 계산된다.

$$\begin{aligned} &\Pi(X) = 2, \\ &\Pi(X|Z) = 2^{h(X|Z)} \approx 0.84, \\ &\wp(X|Z) \approx 0.58. \end{aligned}$$

이 값의 의미는 직관적으로 다음과 같이 해석 가능하다. Z를 알기 이전의 X값의 위치는 크기 2인 집합 내에 속하지만, Z가 알려진 이후에 X값의 위치는 크기가 1보다도 작은 0.84 정도 크기의 집합에 속한다는 것이다.

정보 이론적 개념은 실용적인 환경에서 특수한 조건이 발

생할 경우에 예외적인 결과를 초래할 수는 있다. 하지만 프라이버시의 정도를 정량적으로 평가하기 위한 도구로는 상당히 일반적인 의미를 지닌다고 할 수 있다. 그러므로 상호정보를 이용한 프라이버시 측도는 세부적인 안전성 분석이 아닌 프로토콜 설계 당시에 안전성을 평가할 수 있는 측도로는 매우 유용할 것으로 사료된다.

4.3 프라이버시 손상(privacy breach)

Evfimievski 등[2]이 제안한 프라이버시 손상(privacy breach) 개념은 연관규칙 마이닝에서 프라이버시를 보존하는 기술을 논할 때 밀접한 관련성이 있다. 여기에서는 프라이버시 손상 개념이 특수한 경우에 상호정보 개념을 이용한 측도의 대안이 될 수 있음을 살펴보고, 이에 대한 엄밀한 정의는 5장에서 논하기로 한다.

상호정보를 이용한 프라이버시 측도가 상당히 일반적이고 합리적인 것처럼 보이지만 모든 상황에서 합리적으로 적용할 수 있는 것은 아니다. 4.1절에서 살펴본 예제에서 클라이언트들이 "$X \le -0.99$"와 같은 성질(property)을 노출시키고 싶지 않는다고 해보자. 이 성질에 대한 사전 확률(prior probability)은 $0.01 \times (1/2) = 0.005$, 즉 0.5%이다. 그런데 랜덤화된 값 Z가 구간 $[-2, -1.99]$에서 발생한다면, 이 성질에 대한 사후(posterior) 확률은

$$P(X \le -0.99 \mid Z = z) = 1$$

로 100%가 된다.

물론 Z가 구간 $[-2, -1.99]$에서 발생할 확률은 매우 낮다. 이 확률 값을 계산하면,

$$\begin{aligned} & P(-2 \le Z \le -1.99) \\ &= \int_{-2}^{-1.99} dz \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \\ &= \int_{-2}^{-1.99} dz \int_{0}^{z+1} \frac{1}{2} \cdot \frac{1}{2} dx \\ &= 0.0000125 \end{aligned}$$

이다. 그러므로 Z는 구간 $[-2, -1.99]$에 약 100,000번 중에 한 번 꼴로 발생한다. 그러나 이 사건이 10만 번 중의 한 번이라 할지라도 일단 발생하기만 하면 100%의 확률로 성질 "$X \le -0.99$"는 노출된다.

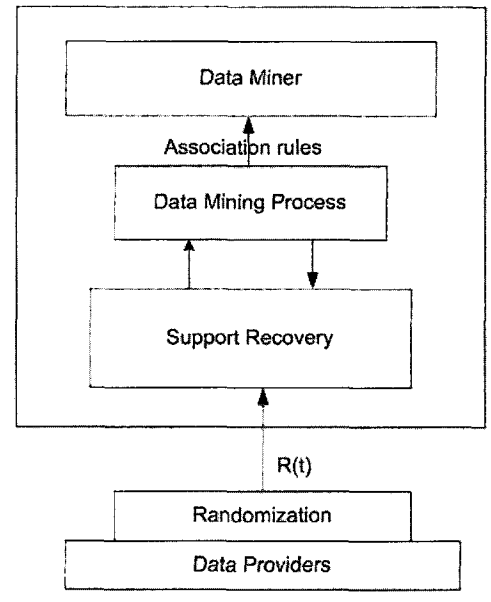
상호정보를 사용한 프라이버시 측도는 평균(average)의 의미가 강한 측도이지 이와 같이 드물게 발생하는 노출 가능성을 탐지해내지는 못한다는 것을 알 수 있다. 이 예제의 경우 "$X \in [-1, 0]$"인지 "$X \in [3, 4]$"인지의 사전 확률은 각각 50%이지만 Z를 알고 난 이후의 사후 확률은 100%가 된다. $Z \in [-2, 1]$이면 X는 첫 번째 구간에 속할 것이고, $Z \in [2, 5]$이면 X는 두 번째 구간에 속할 것이 100% 확실하기 때문이다. 그런데 이러한 성질에 대한 프라이버시 노출도 상호정보를 사용한 측도에서는 탐지하지 못한다.

한편, 프라이버시 손상 개념은 이러한 노출 위험성을 탐지해낼 수 있다. 프라이버시 손상의 수준이 $\rho < 100\%$로 설정된 경우라 한다면 위와 같은 사후 확률은 모두 프라이버시 손상 수준을 넘기 때문에 프라이버시가 손실된 것으로 판단하는 것이다. 그러나 프라이버시 손상 개념의 문제점은 프라이버시 손상 수준 이하로 보존되어야할 민감한(privacy-sensitive) 성질들(properties)이 구체적으로 무엇인지가 정의되어야 한다는 점이다. 너무 많은 성질들을 민감한 것으로 규정할 경우 랜덤화에 의한 손상 정도가 심해서 정확도에 문제가 발생할 수 있을 것이다. 이러한 문제점들 때문에 아직까지 가장 올바른 프라이버시 측도가 무엇인가 하는 문제는 미해결 과제로 남아 있다.

5. 연관규칙 마이닝에서 프라이버시 보호 기술

데이터마이닝에 사용되는 프라이버시 보호 기술 중에서 가장 활발히 연구 되고 있는 분야는 연관규칙 마이닝 관련 프라이버시 보호 기술이다. 연관규칙 마이닝은 전자상거래 등에서 매우 유용한 기술이기 때문이다. 최근의 연구 결과 중에서는 Evfimievski 등[2]의 연구 결과가 주목할 만하다. 본 절에서는 이들의 결과를 분석해보고 결과 개선 및 시뮬레이션을 위한 이론적 배경으로 삼고자 한다. 랜덤화를 사용하여 프라이버시를 보존하는 일반적인 연관규칙 마이닝의 절차는 (그림 8)에 나타나 있다.

그림 상의 과정을 간단히 설명하면 다음과 같다. 먼저 데이터 제공자는 랜덤화 과정을 통해서 프라이버시가 보호된 정보를 데이터 마이너에게 보낸다. 연관규칙 마이닝 알고리즘에서 핵심 요소는 지지도(support) 계산 과정이므로 데이터 마이너는 랜덤화된 데이터를 기반으로 지지도를 계산한다. 이 때, 어떤 아이템 집합의 오리지널 데이터에서의 지지도와 랜덤화된 데이터에서의 지지도 값은 일반적으로



(그림 8) 랜덤화를 이용한 연관규칙 마이닝 과정

다르다. 다른 정도를 최소화 하는 기법이 랜덤화 작용소의 정확도(accuracy)와 직접적인 관계가 있는 것이다. 지지도 값이 계산된 이후에는 보통의 Apriori 알고리즘[13]과 같은 연관규칙 마이닝과 동일한 절차에 의해서 원하는 연관규칙을 이끌어내면 데이터 마이너는 프라이버시가 보존된 상태에서 연관규칙 마이닝 알고리즘을 수행한 것이 된다. 그러므로 프라이버시 보존형 연관규칙 마이닝에서도 핵심적인 과정은 프라이버시 유지와 원래 지지도를 복구하는 것이다.

5.1 연관규칙 마이닝 관련 정의

Evfimievski 등[2]은 연관규칙 마이닝 알고리즘에서 프라이버시를 보호하는 방법으로 기존의 균일한 랜덤화(uniform randomization) 기법이 사용될 경우에 안전성적인 문제가 있음을 지적하였다. 이때 사용된 개념이 프라이버시 손상(privacy breach)이라는 측도이다. 이들의 연구 결과는 프라이버시 손상 관점에서 균일한 랜덤화보다 안전한 효율적인 방법을 제안한 것이다.

연관규칙 마이닝 알고리즘의 핵심은 지지도(support)를 계산하는 것이기 때문에 프라이버시 확보를 위한 랜덤화 기법에서도 가장 중요한 단계는 지지도 계산과 관련된 부분이다. 지지도는 다음과 같이 정의된다.

[정의 5.1.] $I=\{a_1,\cdots,a_n\}$을 n개의 아이템으로 이루어진 아이템 집합이라 하자. $T=(t_1,\cdots,t_N)$은 각각의 t_i가 I의 부분집합으로 구성된 트랜잭션일 때, N개의 트랜잭션으로 구성된 수열이라 하자. 임의의 아이템 집합 $A\subset I$에 대하여 A의 지지도(support)는 다음과 같이 정의된다.

$$supp^T(A)=\frac{\#\ \{t\in T\mid A\subseteq t\}}{N}$$

그리고 어떤 아이템 집합이 사용자가 정의한 파라미터인 τ에 대하여 $supp^T(A)\geq\tau$를 만족할 때, 아이템 집합 $A\subset I$를 T에서 빈번한(frequent) 아이템 집합이라 부른다.

여기에서는 하나의 서버가 여러 개체를 갖는 (그림 2)의 중앙 집중식 시스템 모델을 고려한다. 정의 5.1에서의 기호를 사용한다면, 총 N개의 개체가 있고, 서버가 소유한 전체 물품 목록은 n개의 아이템 집합 $I=\{a_1,\cdots,a_n\}$이 되며, 트랜잭션 t_i는 i번째 개체가 거래한 내역이 된다.

프라이버시 보호를 위해서 사용하는 랜덤화 작용소(randomization operator)는 다음과 같이 정의된다.

[정의 5.2.] (Ω,ϑ,P)를 확률공간(probability space)이라 할 때, 랜덤화 작용소(randomization operator)는 N개의 트랜잭션을 다른 N개의 트랜잭션으로 랜덤하게 변환시키는 가측함수(measurable function)

$$R:\ \Omega\times\{all\ possible\ T\}\rightarrow\{all\ possible\ T\}$$

이다. 랜덤화 이후의 변환된 N개의 트랜잭션을 $T'=R(T)$으로 나타낸다.

이제 프라이버시 손상 개념을 엄밀히 정의할 수 있다.

[정의 5.3.] 랜덤화 되지 않은 수열 T가 알려져 있는 분포로부터 추출된 것이라 하고, T 안의 i번째 트랜잭션을 $t_i\in T$라 하자. 성질(property) $P(t_i)$에 대하여 수준(level) ρ의 프라이버시 손상(privacy breach)이 발생했다는 의미는 다음과 같다.

$$\exists\ T':\ P(P(t_i)\mid R(T)=T')\geq\rho$$

또한 성질(property) $P(t_i)$에 대하여 성질 $Q(T')$이 수준 ρ의 프라이버시 손상을 야기한다(cause a privacy breach)는 의미는 다음이 성립한다는 것이다.

$$P\big(P(t_i)\mid Q(R(T))\big)\geq\rho.$$

트랜잭션 $t_i\in T$와 아이템 집합 $A\subset I$, 그리고 하나의 아이템 $a\in A$에 대하여 성질 "$A\subset t_i'\in T'$"이 성질 "$a\in t_i$"에 대하여 프라이버시 손상을 야기하는 경우를 고려해보자. 다시 말해서 랜덤화된 트랜잭션 속에 A가 들어있다는 사실이 랜덤화 되지 않은 오리지널 트랜잭션 속에 원소 a가 들어 있음을 어느 정도 확신할 수 있는 경우이다. 이 경우에 아이템 집합 A가 프라이버시 손상을 야기한다고 말한다. 정확한 정의는 다음과 같다.

[정의 5.4.] 적당한 아이템 $a\in A$에 대해서

$$P\big(a\in t_i\mid A\subset t_i'\big)\geq\rho$$

를 만족하는 $i\in\{1,\cdots,N\}$이 존재하면, 아이템 집합 A는 수준 ρ의 프라이버시 손상을 야기한다고 말한다.

5.2 랜덤화 작용소

랜덤화 작용소 R이 트랜잭션별(per-transaction)로 이루어지고 아이템에 따라서 변하지 않는다는 조건이 있으면 오리지널 데이터 복구에 대한 이론을 명확히 전개할 수 있다.

$$R(t_1,\cdots,t_N)=(R(1,t_1),\cdots,R(N,t_N))$$

일 때, 랜덤화가 트랜잭션별로 이루어진다고 한다. 여기에서 $R(i,t_i)$들은 서로 독립이고, 분포가 i에는 무관하고, t에만 의존하므로

$$t_i'=R(i,t_i)=R(t_i)$$

로 나타낸다. 랜덤화가 아이템 불변(item-invariant)이라는

의미는 임의의 순열 $\pi: I \rightarrow I$에 대하여, $\pi^{-1}R(\pi T)$의 분포가 $R(T)$의 분포와 동일하다는 것이다.

Evfimievski 등[2]이 제안한 랜덤화 작용소는 개수 선택(select-a-size)으로 명명된 작용소로 오리지널 트랜잭션 중 랜덤하게 선택된 몇 개의 항목들을 그 트랜잭션에 들어 있지 않은 아이템으로 대치하는 방법이다. 여기에서 개수를 선택한다는 의미는 오리지널 트랜잭션에서 대치되는 아이템의 개수를 선택한다는 것이다. select-a-size 랜덤화 작용소의 정의는 다음과 같다.

[정의 5.5.] 각 트랜잭션 내의 크기(아이템의 개수)가 m으로 모두 동일할 때, select-a-size 랜덤화 작용소는 다음과 같은 파라미터를 갖는다.

- 랜덤화 수준으로 불리는 아이템의 기본 확률(default probability)은 $\rho_m \in (0, 1)$으로 주어진다.
- 트랜잭션의 부분집합의 크기를 선택하는 확률 $p_m[j] (0 \le j \le m)$은 다음을 만족한다.

$$p_m[j] \ge 0,\ p_m[0] + p_m[1] + \cdots + p_m[m] = 1 .$$

오리지널 트랜잭션의 수열 $T = (t_1, \cdots, t_N)$이 주어졌을 때, 랜덤화 작용소는 각각의 t_i에 독립적으로 작용하여 t_i'을 다음과 같은 과정으로 생성한다. 이때, 각 트랜잭션 집합의 크기는 m으로 일정하다. 즉, $|t_i| = m$이다.

1. 작용소는 집합 $\{0, 1, \cdots, m\}$으로부터 정수 j를 랜덤하게 선택한다. 여기에서 j가 선택될 확률은 $p_m[j]$이다.
2. 트랜잭션 t_i로부터 균일하게(uniformly) j개의 아이템을 비복원 추출하여 t_i'에 놓고, 추출되지 않는 아이템들은 t_i'에 놓지 않는다.
3. 트랜잭션 t_i에 들어 있지 않은 각 아이템에 대하여 ρ_m의 앞면 성공 확률로 동전던지기를 실시하여, 앞면 결과를 나타내는 아이템들을 t_i'에 포함시킨다. 즉, $a \notin t_i$인 각 아이템에 대하여 성공률 ρ_m인 베르누이 시행을 독립적으로 반복 실시한다.

[정의 5.5]에 기술되어 있는 select-a-size 랜덤화 작용소는 상당히 광범위한 것으로 볼 수 있다. 구체적인 확률 분포가 제시되어 있지 않고, 방법도 매우 일반적인 것이기 때문이다. 오리지널 트랜잭션에서 랜덤화 이후에도 유지되는 아이템의 개수 j를 랜덤하게 선택하고, 나머지는 오리지널 트랜잭션에 들어 있지 않은 아이템들 각각에 성공 확률 ρ_m인 베르누이 시행을 독립적으로 $n-m$번 반복 수행한다는 것이 select-a-size 랜덤화 작용소의 수행 과정이다. 여기에서 전체 아이템 집합 I의 개수는 n으로 가정하였다.

Evfimievski 등[2]은 select-a-size 랜덤화 작용소의 특수한 예로 볼 수 있는 cut-and-paste 랜덤화 작용소에 대해서 이론적인 분석과 실제 데이터를 바탕으로 한 시뮬레이션 결과를 제시하였다. Cut-and-paste 랜덤화 작용소는 처음에 선택하는 파라미터 j의 상한선을 정하고, 베르누이 시행을 할 때, 오리지널 트랜잭션에 남아 있는 아이템들도 포함시킨다는 점이 추가되었다.

[정의 5.6.] Cut-and-paste 랜덤화 작용소는 select-a-size의 특수한 경우로 랜덤화 수준 $\rho_m \in (0, 1)$과 컷오프(cutoff)를 의미하는 정수 $K_m > 0$을 파라미터로 갖는다.

오리지널 트랜잭션의 수열 $T = (t_1, \cdots, t_N)$이 주어졌을 때, 랜덤화 작용소는 각각의 t_i에 독립적으로 작용하여 t_i'을 다음과 같은 과정으로 생성한다. 이때, 각 트랜잭션 집합의 크기는 m으로 일정하다. 즉, $|t_i| = m$이다.

1. 작용소는 집합 $\{0, 1, \cdots, K_m\}$으로부터 정수 j를 균일 랜덤하게(uniformly random) 선택한다. 이때 $j > m$이면, $j = m$으로 놓는다.
2. 트랜잭션 t_i로부터 균일하게(uniformly) j개의 아이템을 비복원 추출하여 t_i'에 놓는다.
3. 트랜잭션 t_i에 들어 있지 않은 아이템들과 2 단계에서 추출되지 않은 아이템들에 대하여 ρ_m의 앞면 성공 확률로 동전던지기를 실시하여, 앞면 결과를 나타내는 아이템들을 t_i'에 포함시킨다.

프라이버시 손상에 관한 안전성 조건과 랜덤화된 트랜잭션들로부터 연관규칙을 이끌어내는 정확성 관련 조건을 동시에 만족하는 최적의 파라미터를 찾는 것이 랜덤화 작용소 관련 연구의 중요한 과제이다. 일반화된 정의인 select-a-size 랜덤화 작용소에서 최적의 파라미터를 찾는다는 것은 미해결 문제이다. Cut-and-paste 작용소에 대해서 최적의 파라미터를 찾는 문제는 Evfimievski 등[2]에 의해서 해결되었다. 그러나 cut-and-paste 방법에 의한 랜덤화 기법에서 오리지널 지지도를 추정하기 위해서 사용되는 복원 기법은 아이템 집합이 큰 경우에 매우 비효율적이다. 확률 및 통계 이론을 적용하여 편향되지 않은(unbiased) 추정량(estimator)을 제시하였지만, 이는 아이템 개수에 의존하여 커지게 되는 행렬의 역행렬 계산을 요한다. 또한, 아이템 개수가 클 경우에 지지도 추정량의 분산(variance)도 커지게 되어 직관적으로 지지도 추정량의 정확도가 떨어지는 결과를 초래한다. 실제로 [14]에서 저자들은 cut-and-past 랜덤화 작용소의 이러한 단점을 지적하고 있다. 본 논문에서는 cut-and-paste 방법의 단점을 보완하여 아이템 집합이 큰 경우에도 효율적으로 지지도를 복구해낼 수 있는 랜덤화 기법을 제안한다.

6. 제안하는 랜덤화 기법

우리는 select-a-size 랜덤화 작용소의 한 종류를 제안하고자 한다. Cut-and-paste 랜덤화 작용소는 프라이버시 손

상 관점의 안전성 조건과 랜덤화된 트랜잭션으로부터 연관 규칙 마이닝의 정확성을 얻기 위한 조건을 만족시키는 파라미터를 정할 수 있다는 장점을 지닌다. 그러나 오리지널 트랜잭션 중에서 랜덤화 이후에도 그대로 남아 있는 아이템들의 개수를 균일 분포(uniform distribution)로부터 추출한다는 것은 직관적으로 부자연스럽다. 랜덤화 이후 오리지널 트랜잭션 중의 아이템이 하나도 남아 있지 않을 확률과 오리지널 트랜잭션의 아이템들이 모두 남아 있을 가능성이 동일하다는 사실이 정확성과 안전성 관점 모두에서 좋지 않을 수 있다는 생각이다. 여기에서 우리는 오리지널 트랜잭션에서 랜덤화 이후에도 남아 있는 아이템의 수를 적절히 예측하기 위해서 균일 분포가 아닌 이항 분포(binomial distribution)를 사용할 것을 제안한다.

6.1 Binomial-selector 랜덤화 작용소

우리가 제안하는 랜덤화 작용소는 이항분포에 기반한 것이기 때문에 "binomial-selector"라 이름 붙이기로 한다. 이에 대한 자세한 정의는 다음과 같다.

[정의 6.1.] Binomial-selector 랜덤화 작용소는 select-a-size의 특수한 경우로 랜덤화 수준 $\rho_m \in (0,1)$ 과 이항 분포의 추출 확률 $p \in (0,1)$ 를 파라미터로 갖는다.

오리지널 트랜잭션의 수열 $T=(t_1,\cdots,t_N)$ 이 주어졌을 때, 랜덤화 작용소는 각각의 t_i 에 독립적으로 작용하여 t_i' 을 다음과 같은 과정으로 생성한다. 이때, 각 트랜잭션 집합의 크기는 m 으로 일정하고, 전체 아이템 집합의 총 개수는 n 이다. 즉, $|t_i|=m$, $|I|=n$ 이다.

1. 작용소는 트랜잭션 t_i 에 속하는 각 아이템에 대하여 성공 확률 p의 베르누이 시행을 독립적으로 반복 실행하여, 성공하면 t_i' 에 놓고 실패하는 아이템은 버린다. 이렇게 하면 t_i 에 속하는 m 개의 아이템 중에서 t_i' 에도 속하는 아이템의 개수가 $j \in \{0,1,\cdots,m\}$ 일 확률은 이항 분포 $B(m,p)$ 를 따른다. 즉, 다음이 성립한다.

$$p_m[j]=\binom{m}{j}p^j(1-p)^{m-j},\ \ j=0,1,\cdots,m$$

2. 트랜잭션 t_i 에 들어 있지 않은 $n-m$ 개의 아이템들에 대하여 성공 확률 ρ_m 인 베르누이 시행을 독립적으로 반복 수행하여, 성공하는 아이템들만 t_i' 에 포함시킨다.

[정의 6.1]의 단계 1에서 고려하는 이항분포 $B(m,p)$ 를 따르는 확률변수를 X , 단계 2에서 반복적인 베르누이 시행의 결과 선택되는 아이템의 개수를 확률변수 Y로 나타내고, 랜덤화 이후의 트랜잭션 t_i' 의 아이템 개수를 나타내는 확률변수를 Z라고 놓으면, $Z=X+Y$ 가 성립한다.

확률변수 Y는 총 $n-m$ 번의 베르누이 시행을 성공 확률 ρ_m을 가지고 독립적으로 반복하기 때문에 이항분포 $B(n-m,\rho_m)$ 을 따른다. 그러므로

$$E[Z]=E[X]+E[Y]=mp+(n-m)\rho_m$$

이 성립한다. 즉, binomial-selector 랜덤화 작용소 적용 이후의 트랜잭션 하나에 포함된 아이템의 평균 개수는 $mp+(n-m)\rho_m$이 된다는 것이다. 우리는 랜덤화 수행 이후의 트랜잭션에 포함된 아이템의 개수를 랜덤화 이전의 그것과 비슷하게 유지되도록 하는 조건을 추가하기로 한다. 즉,

$$\begin{aligned}E[|t_i'|]=E[Z]&=mp+(n-m)\rho_m\\&=m=E[|t_i|]\end{aligned}$$

가 성립하기를 원하는 것이다. 이는 연관규칙의 정확성(accuracy) 관점에서 직관적으로 적절한 조건으로 생각된다. 랜덤화 이후의 트랜잭션들이 랜덤화 이전의 트랜잭션들과 아이템의 개수부터 현저히 차이가 난다면, 연관규칙의 정확성이 저하될 것은 분명해 보이기 때문이다. 이 조건을 만족하는 ρ_m의 값은 다음과 같이 주어진다.

$$\rho_m=\frac{m\cdot(1-p)}{n-m}$$

이와 같이 ρ_m의 값을 설정하기로 하고, 초기 이항분포에서 발생하는 파라미터 p를 변화시키면서 프라이버시 관련 조건과 연관규칙의 정확성 관점에서 적절한 수준의 p 값을 시뮬레이션을 통해서 살펴보고자 한다.

6.2 시뮬레이션 결과

연관규칙 마이닝 알고리즘에서 핵심 요소는 지지도(support)를 구하는 것이다. 그러므로 프라이버시 보존형 연관규칙 마이닝 프로토콜에서도 랜덤화 전과 후의 지지도가 얼마나 변하는가 하는 것이 정확도(accuracy)를 측정하는 기준이 된다. 랜덤화 이후의 트랜잭션들에서 어떤 아이템 집합의 지지도가 랜덤화 이전과 별로 차이가 없다면 그 프로토콜의 정확도는 높다고 볼 수 있는 것이다.

실험 조건은 아이템 집합의 크기 n은 10개, 트랜잭션의 개수 N은 100,000개, 각 트랜잭션 당 가지고 있는 아이템의 개수 m 은 3과 4로 설정하였다. 랜덤화 이전의 오리지널 트랜잭션들에는 Apriori 알고리즘에서와 같은 방법으로 지지도를 구했으며, 랜덤화 이후의 트랜잭션들에 Apriori 알고리즘을 적용한 결과를 비교해보았다.

시뮬레이션에 사용된 데이터는 $m=3,4$에 대하여 각각 랜덤하게 생성된 10만개의 트랜잭션 자료이며, binomial-selector로 랜덤화된 이후의 지지도에 대한 기대값(expectation)을 추정하기 위해서 100번의 랜덤화 과정을 거쳐서 지지도의 평균값(average)을 구하였다. 또한 랜덤화 수

〈표 1〉 $m=3$인 경우의 오리지널 지지도 ($N=100,000$)

아이템 집합	지지도
A_1={1}	28,930
A_2={1,2}	8,660
A_3={1,2,3}	1,216

준을 결정하는 확률값 p에 대한 지지도의 추이를 관찰하기 위해서 $p=0.1$부터 $p=0.9$까지 0.1 간격으로 조사하였다. 이 때, 또 다른 랜덤화 수준 값인 ρ_m은 6.1절에서 논한 바와 같이 $\rho_m = m(1-p)/(n-m)$으로 계산하였다. 지지도 계산에 사용된 아이템 집합은 개수가 1개, 2개, 3개인 경우에 대하여 조사하였다.

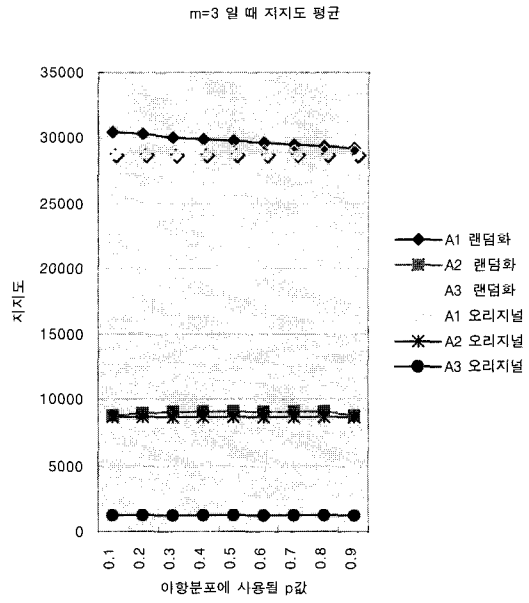
아래 <표 1>은 $m=3$인 경우에 각 아이템 집합에 대한 오리지널 데이터의 지지도를 표현한 것이다. 원래 지지도는 전체 트랜젝션 중에서 해당 아이템 집합을 포함하고 있는 트랜젝션의 비율을 의미한다. 여기에서는 편의상 지지도의 분자에 해당하는 수, 즉, 아이템 집합을 포함하고 있는 트랜젝션의 개수를 지지도로 표현한다. 정의 5.1에서의 표현법으로 <표 1>에서 의미하는 바는 다음과 같다.

$$supp^T(A_1) = \frac{28930}{100000} = 0.2893,$$
$$supp^T(A_2) = \frac{8660}{100000} = 0.0866,$$
$$supp^T(A_3) = \frac{1216}{100000} = 0.001216.$$

<표 2>는 $m=3$인 경우에 각 아이템 집합에 대한 랜덤화 이후의 데이터에 대한 지지도의 평균값을 랜덤화 수준 p를 변화시키면서 얻은 결과를 표현한 것이다. 랜덤화는 100회를 수행하여 지지도의 평균값을 구함으로써 랜덤화 이후의 지지도에 대한 기대값을 추정해본 것이다. <표 2>에 나타난 결과를 살펴보면 대체적으로 랜덤화 수준 p가 1에 가까울수록 랜덤화 이후의 지지도가 오리지널 지지도에 근접함을 알 수 있다. 이는 랜덤화 수준 p의 의미를 생각해볼 때, 지극히 자연스러운 결과이다. 일반적으로 랜덤화 수준 p가 작은 값인 경우 프라이버시 수준이 높다고 할 수 있으므로 프라이버시 수준과 정확도는 반비례함을 볼 수 있다. (그림 9)는 <표 1>과 <표 2>의 결과를 그래프로 표현한 것으로 랜덤화 수준 p의 변화에 따른 지지도의 추이를 시각적으로 보여준 것이다.

〈표 2〉 $m=3$인 경우의 랜덤화된 지지도 평균값 (100회)

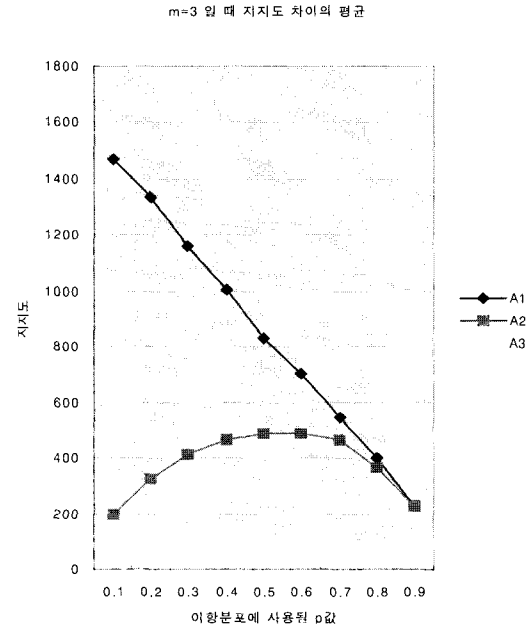
A \ p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
A_1	30,398	30,265	30,094	29,932	29,758	29,631	29,474	29,333	29,158
A_2	8,857	8,985	9,075	9,123	9,149	9,149	9,124	9,028	8,887
A_3	2,771	2,772	2,741	2,662	2,533	2,373	2,174	1,901	1,594



(그림 9) $m=3$일 때 p에 따른 지지도 변화

〈표 3〉 $m=3$인 경우의 지지도 차이의 평균값 (100회)

A \ p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
A_1	1,468	1,335	1,164	1,002	828	701	544	403	228
A_2	198	325	415	463	489	489	464	368	227
A_3	1,555	1,556	1,525	1,446	1,317	1,157	958	685	378



(그림 10) $m=3$일 때 p에 따른 지지도 오차의 변화

〈표 4〉 $m=4$인 경우의 오리지널 지지도 (N= 100,000)

아이템 집합	지지도
A_1={1}	39,059
A_2={1,2}	16,311
A_3={1,2,3}	5,244

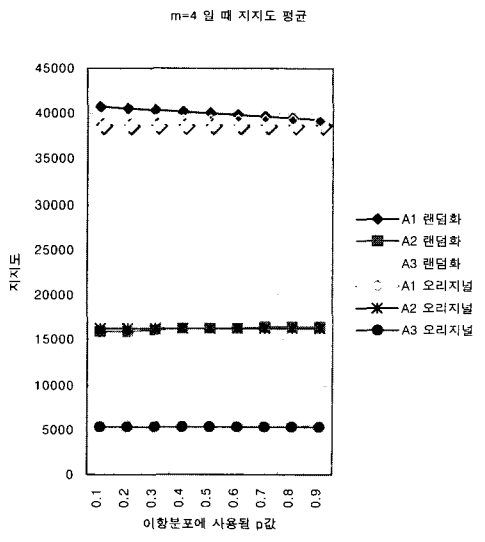
〈표 5〉 $m=4$ 인 경우의 랜덤화된 지지도 평균값 (100회)

A \ p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
A_1	40,690	40,498	40,330	40,195	39,995	39,821	39,680	39,489	39,309
A_2	15,916	15,996	16,082	16,214	16,270	16,356	16,443	16,455	16,442
A_3	6,676	6,650	6,601	6,559	6,423	6,270	6,074	5,829	5,564

<표 3>에서는 랜덤화 수준 p에 따른 지지도 차이의 평균값을 표현하고 있다. 이는 랜덤화 이후의 자료에서 지지도를 계산할 경우에 발생할 수 밖에 없는 오차를 의미한다. 예를 들면, 아이템 집합 A_2의 오리지널 지지도는 0.0866인데, 랜덤화 수준 p가 0.4인 경우 랜덤화된 데이터로부터 추정된 지지도는 약 0.09123이 되어 0.00463의 오차를 보이게 된다는 뜻이다. <표 3>에서는 이 지지도의 오차에 트랜잭션의 전체 개수인 N= 100000을 곱한 값이 나타나 있고, 이를 그래프로 표현한 것이 (그림 10)이다.

<표 4>는 $m=4$인 경우에 각 아이템 집합에 대한 오리지널 데이터의 지지도를 표현한 것이다. 표현 방법은 $m=3$인 경우에서와 동일하다.

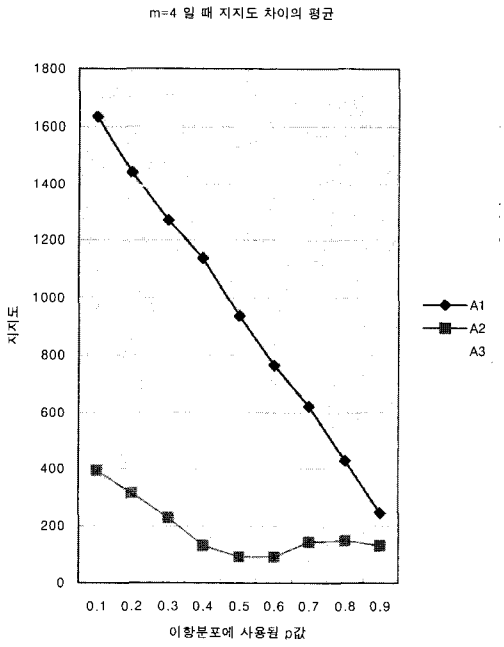
<표 5>는 $m=4$ 인 경우에 각 아이템 집합에 대한 랜덤화 이후의 데이터에 대한 지지도의 평균값을 랜덤화 수준 p를 변화시키면서 얻은 결과를 표현한 것이다. 랜덤화 방법과 횟수는 $m=3$인 경우와 같다. <표 4>와 <표 5>의 내용을 그래프로 표현한 것이 (그림 11)에 나타나 있다.



(그림 11) $m=4$ 일 때 p에 따른 지지도 변화

〈표 6〉 $m=4$인 경우의 지지도 차이의 평균값 (100회)

A \ p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
A_1	1,631	1,439	1,271	1,136	936	762	621	430	250
A_2	394	314	228	135	90	94	143	147	133
A_3	1,432	1,406	1,357	1,315	1,179	1,026	830	585	320



(그림 12) $m=4$ 일 때 p에 따른 지지도 오차의 변화

<표 6>에서는 $m=4$인 경우에 랜덤화 수준 p에 따른 지지도 차이의 평균값을 표현하고 있으며, 이를 그래프로 표현한 것이 (그림 12)에 나타나 있다.

6.3 Cut-and-paste와의 비교 분석

본 논문에서 제안한 binomial-selector와 참고문헌 [3]에 제안되어 있는 cut-and-past 랜덤화 작용소는 모두 [3]에 있는 select-a-size 랜덤화 작용소의 특수한 예에 속한다. 그러므로 select-a-size 방식에 대해서 전개되어 있는 모든 이론이 binomial-selector와 cut-and-paste에 대해서도 적용 가능하다.

한편, cut-and-paste 방식에 대해서 [3]의 저자들은 프라이버시 손상 판점의 안전성 조건을 만족하는 최적의 파라미터들을 제시하였다. 통계적으로 편향되지 않은 추정량(unbiased estimator)을 제안함으로써 정확도를 보장받고, 확률 이론적인 분석에 의해서 안전성 조건을 만족하는 파라미터들을 결정한 것이다. 그런데 cut-and-paste 방식에서 사용되는 지지도 추정량은 트랜잭션의 전이확률(transition probability)을 성분으로 갖는 정칙행렬의 역행렬 계산을 필

요로 한다. 이 역행렬 계산은 아이템 집합이 큰 경우에는 막대한 효율성 저하의 요인이 된다. 또한, 여기에 사용된 지지도 추정량의 분산은 아이템 집합이 큰 경우에 커지게 되므로 정확도가 떨어질 수 있다는 단점을 지닌다.

직관적으로도 cut-and-paste 방식은 오리지널 트랜젝션에서 랜덤화 이후 남는 아이템의 개수는 균일분포(uniform distribution)를 이루고, 새롭게 들어오는 아이템의 개수는 이항분포를 이루기 때문에 자연스럽지 못한 측면이 있다. Binomial-selector는 랜덤화 이후에 남아 있는 아이템의 개수와 새롭게 들어오는 아이템의 개수 모두 이항분포를 따르도록 하는 방식이다.

랜덤화된 이후의 데이터로부터 지지도를 추정하는 방식은 binomial-selector의 경우 오리지널 데이터의 지지도 계산 방식과 같으므로 매우 효율적이다. Cut-and-paste 방식에서 사용하는 역행렬 계산 과정이 필요하지 않다는 의미이다. 또한, 아이템 집합이 큰 경우에도 지지도 추정 방식의 오리지널 데이터에 대한 계산과 동일하여 효율성은 떨어지지 않는다.

시뮬레이션 결과 binomial-selector는 랜덤화 수준 파라미터인 p를 조절함으로써 정확도와 안전성 수준을 적절히 선택할 수 있다. 즉, p의 값이 0에 가까우면 프라이버시 보호 수준은 높아지지만 정확도는 떨어진다. 반면에 p의 값이 1에 가까우면 프라이버시 보호 수준이 낮아지고 정확도는 높아진다. 그러므로 필요에 따라 안전성 수준과 정확도를 하나의 파라미터에 의해서 취사선택(trade-off)할 수 있는 것이다. 향후 이러한 시뮬레이션 결과를 바탕으로 좀 더 면밀한 이론적 분석을 실시하여 정확한 관계식에 의해서 안전성 수준과 정확도를 결정할 수 있는 연구가 진행되어야 한다.

7. 결 론

본 논문에서 우리는 데이터 마이닝에서 프라이버시 보호 기술에 대하여 살펴보았다. PPDM이라 불리는 프라이버시 보존형 데이터 마이닝 기술 개발 동향을 조사 분석하였다. 계산 효율성 때문에 실용화 되지 못하고 있는 SMC 기반 PPDM은 현재의 컴퓨팅 환경에서는 다분히 이론적인 것에 머무른다고 할 수 있다. 그래서 우리는 실용적인 PPDM 기술에 집중하여 가장 널리 사용되고 있는 랜덤화 기법에 대한 연구를 진행하였다.

특히, 실용적인 PPDM 분야에서 가장 중요한 프라이버시 측도 개념을 심도있게 분석하였으며, 연관규칙 마이닝에서의 프라이버시 기법에 초점을 맞추었다. Evfimievski 등[3]이 제안한 select-a-size 범주에 속하는 새로운 랜덤화 작용소인 binomial-selector 개념을 제안하고, 적절한 파라미터를 찾기 위한 시뮬레이션 결과를 제시하였다. 아이템 개수가 큰 경우에 계산 효율성이 낮은 기존의 cut-and-paste 방식에 비해서 binomial-selector는 아이템의 개수에 상관 없이 효율적으로 지지도를 추정할 수 있다는 장점을 지닌다. 여기에 제안된 binomial-selector에 대해서는 추가적으로 정확도와 프라이버시 보호 관점의 이론적 분석이 면밀히 수행되어야 할 것으로 사료된다.

참 고 문 헌

[1] J. Vaidya, C. Clifton, "Privacy-Preserving Data Mining: Why, How, and When", IEEE Security & Privacy, November/December 2004, www.computer.org/security/

[2] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules", Proc. ACMSIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2002, pp. 217-228.

[3] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules", Information Systems, Vol. 29, 2004, pp. 343-364.

[4] O. Goldreich, "Secure Multi-Party Computation (Final Draft, Version 1.4)", http://www.wisdom.weizmann.ac.il/home/oded/public_html/foc.html, 2002.

[5] R. Agrawal, R. Srikant, "Privacy preserving data mining", ACM SIGMOD Conference on Management of Data, Dallas, TX, 2000, pp. 439-450.

[6] Y. Lindell, B. Pinkas, "Privacy preserving data mining", CRYPTO 2000, pp. 36-54.

[7] J. R. Quinlan. "Discovering rules by induction from large collection of examples", Expert Systems in the Micro Electronic Age, Edinburgh University Press, pp. 168-201.

[8] J. R. Quinlan, "Induction of decision trees", Machine learning, Vol. 1, No. 1, 1986, pp. 81-106.

[9] K. Muralidhar, R. Sarathy, "A theoretical basis for perturbation methods", Statistics and Computing, Vol. 13, 2003, pp. 329-335.

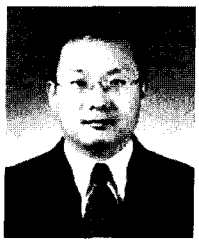
[10] T. Dalenius, "Towards a methodology for statistical disclosure control", Statistisktidskrift , Vol. 5, 1977, pp. 429-444.

[11] G. T. Duncan, D. Lambert, "Disclosure limited data dissemination", Journal of the Americal Statistical Association, Vol. 81, 1986, pp. 10-18.

[12] D. Agrawal, C. C. Agrawal, "On the design and quantification of privacy preserving data mining algorithms", Proceedings of the 20th Symposium on Principles of Database Systems, May 2001.

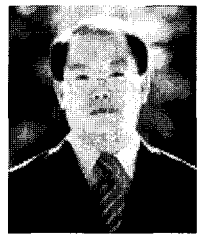
[13] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the ACM SIGMOD Conference on Management of Data, 1993, pp. 207-216.

[14] N. Zhang, S. Wang, W. Zhao, "A new scheme on privacy preserving association rule mining", PKDD 2004, LNAI 3202, 2004, pp. 484-495.



강 주 성

e-mail : jskang@kookmin.ac.kr
1989년 고려대학교 수학과(학사)
1991년 고려대학교 일반대학원 수학과 (이학석사)
1996년 고려대학교 일반대학원 수학과 (이학박사)
1997년~2004년 한국전자통신연구원 선임연구원
2004년~현재 국민대학교 수학과 부교수
관심분야 : 암호이론, 정보보호이론, 응용확률론 등

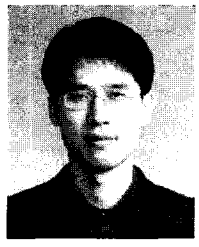


이 옥 연

e-mail : oyyi@kookmin.ac.kr
1988년 고려대학교 수학과
1990년 고려대학교 일반대학원 수학과 (이학석사)
1996년 University of Kentucky 수학과 (이학박사)
1999년~2001년 한국전자통신연구원 선임연구원, 팀장
2001년~현재 국민대학교 수학과 조교수
관심분야 : 정보보호, 이동통신, 암호론

조 성 훈

e-mail : mysho2n@hotmail.com
2004년 국민대학교 수학과(학사)
2007년 국민대학교 일반대학원 수학과(이학석사)
2007년~현재 누리솔루션 엔지니어
관심분야 : 암호이론, 정보보호이론, 금융보안 등



홍 도 원

e-mail : dwhong@etri.re.kr
1994년 고려대학교 수학과(학사)
1996년 고려대학교 일반대학원 수학과 (이학석사)
2000년 고려대학교 일반대학원 수학과 (이학박사)
2000년~현재 한국전자통신연구원 선임연구원
관심분야 : 암호이론, 정보보호이론, 이동통신 정보보호 등