情報保護學會論文誌
第18卷 第4號, 2008. 8

행렬 기반 랜덤화를 적용한 프라이버시 보호 기술의 안전성 및 정확성 분석*

강 주 성[1†], 안 아 론[1], 홍 도 원[2]
[1]국민대학교, [2]한국전자통신연구원

An Analysis of Privacy and Accuracy for Privacy-Preserving Techniques by Matrix-based Randomization*

Ju-Sung Kang[1†], Aron An[1], Dowon Hong[2]
[1]Kookmin University, [2]Electronics and Telecommunications Research Institute

요 약

실용적인 프라이버시 보호 기술 중의 하나인 행렬 기반 랜덤화 기법에 대하여 세밀한 분석을 실시한다. 최적의 변환 행렬을 찾기 위한 프라이버시 손상 관점의 요구조건 및 정확성 측도로 제안된 행렬의 조건수 개념과 연관된 파라미터들 간의 관계를 이론적으로 규명한다. 행렬 기반의 대표적 알고리즘인 랜덤 대치 기법의 효율적인 구현을 위하여 데이터 재구축 과정에서 필요한 역행렬을 간단히 구하는 공식을 제시하고, 행렬의 노름에 따른 변환 행렬의 조건수와 변환된 분포의 기댓값 및 분산을 계산함으로써 표준오차와 파라미터들 간의 관계식을 도출한다. 또한, 랜덤 대치 기법을 구현하여 다양한 시뮬레이션을 실시함으로써 이론적으로 얻은 결과를 실험적으로 검증한다.

ABSTRACT

We study on the practical privacy-preserving techniques by matrix-based randomization approach. We clearly examine the relationship between the two parameters associated with the measure of privacy breach and the condition number of matrix in order to achieve the optimal transition matrix. We propose a simple formula for efficiently calculating the inverse of transition matrix which are needed in the re-construction process of random substitution algorithm, and deduce some useful connections among standard error and another parameters by obtaining condition numbers according to norms of matrix and the expectation and variance of the transformed data. Moreover we give some experimental results about our theoretical expressions by implementing random substitution algorithm.

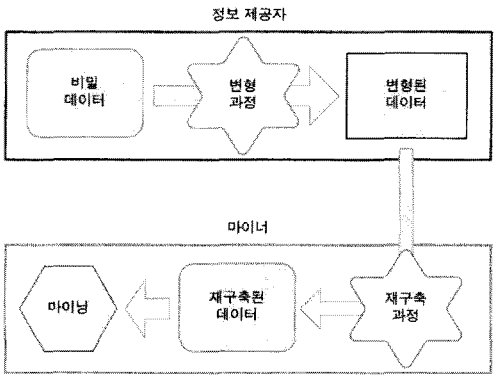
Keywords : privacy-preserving techniques, matrix-based randomization, privacy breach, condition number

접수일 : 2008년 1월 25일; 수정일 : 2008년 4월 17일;
채택일 : 2008년 5월 19일
* 본 연구는 지식경제부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2005-Y001-04, 차세대 시큐리티 기술 개발]
† 주저자, jskang@kookmin.ac.kr

Ⅰ. 서 론

랜덤화 또는 데이터 변형(data perturbation)이라 불리는 기법은 랜덤한 노이즈(noise)를 사용하여 기밀을 요하는 변인을 수정함으로써 프라이버시를 보호하고자

하는 방법이다. 랜덤화를 사용하는 프라이버시 보호 메커니즘에서는 데이터를 공개하는 정도에 따라 데이터 유용성(data utility)과 노출 위험성(disclosure risk) 사이에서 적절한 취사선택(trade-off)이 요구된다. 아무런 데이터도 공개하지 않으면 당연히 노출 위험성은 전혀 없지만 데이터의 유용성도 함께 사라지게 된다. 데이터를 공개하는 정도가 커지면 커질수록 노출 위험성은 증가하지만 데이터 유용성 또한 높아지게 되는 것이다. 그러므로 이러한 데이터 유용성과 노출 위험성 사이의 상관관계를 합리적인 측도에 의해서 결정할 수 있는 방식이 우수한 랜덤화 기법이라 할 수 있다.



〔그림 1〕 PPDM에서 랜덤화 기법의 적용

1.1 랜덤화 기법의 연구 동향

실용적인 프라이버시 보호 기술은 대부분 다양한 랜덤화 기법에 의존한다. 대표적인 응용 분야인 프라이버시 보존형 데이터 마이닝(Privacy-Preserving Data Mining, PPDM)에서는 정보제공자의 비밀 데이터를 보호하기 위해서 변형된(perturbed) 데이터를 마이너에게 제공한다. 데이터의 변형은 프라이버시 관련 정보를 노출시키지 않기 위함이며, 데이터 변형의 가장 실용적인 방법이 랜덤화 기법이다. 랜덤화 기법을 이용한 PPDM의 효시가 된 연구 결과는 Agrawal-Srikant[1]가 제시한 방법이다. 이들의 연구 결과를 기점으로 하여 랜덤화 기법을 이용한 여러 가지 PPDM 기술이 발표되었다. 랜덤화를 이용한 PPDM에서 프라이버시를 보존하기 위한 기본적인 가정은 정보제공자들이 민감한 속성(sensitive attribute)에 대해서 수정된 값을 변형된 데이터로 제공한다는 것이다. 그러면 데이터 마이너는 수신된 변형 데이터들로부터 원래의 데이터 자체를 복구할 수는 없지만 통계적 추정(estimation)에 의하여 확률분포만을 재구축함으로써 유용한 정보를 추출해낼 수 있다. 프라이버시 보호의 정도는 데이터를 변형하는 기법에 의존하고, 데이터 마이닝 결과의 정확성(accuracy)은 분포의 재구축 과정에서 발생하는 오차(error)에 따라 결정된다. PPDM에서 랜덤화 기법이 적용되는 과정을 도식화 한 것이 [그림 1]에 나타나 있다.

랜덤화 기법과 관련하여 최근에 발표된 의미 있는 연구 결과로는 랜덤 회전(random rotation)[2] 기법과 랜덤 사영(random projection)[3] 기법을 들 수 있다. 이 두 가지 방법은 비교적 높은 정확도를 가지지만 계산량적인 효율성과 실제 데이터에 적용하는 관점에서 실용성이 높다고 할 수 없다. 랜덤 회전과 랜덤 사영 기법 모두 노름(norm) 등과 같이 회전 변환이나 사영 변환에 의해서 불변이거나 변화가 미미한 특성에 대해서만 적용할 수 있다는 단점을 지닌다. 실제 응용 관점에서도 이 두 가지 기법은 모든 PPDM에 적용되지는 못하고, 분류(classification), 군집화(clustering) 등의 데이터마이닝 분야에서만 적용 가능하다. 가장 널리 사용되는 연관규칙 마이닝에 대해서는 랜덤 회전이나 랜덤 사영 기법을 적용하기에는 무리가 따른다. 또한, 랜덤 회전 기법에서는 사용되는 변환 행렬에 대한 안전성적인 문제점이 최근 지적되기도 하였다[3].

한편, Agrawal-Harista[4]는 PPDM을 위한 랜덤화 기법을 조직적으로 설계하기 위하여 행렬 이론이 적용된 프레임워크를 제안하였다. FRAPP(FRamework for Accuracy in Privacy-Preserving mining)로 명명된 이 프레임워크는 기존의 랜덤화 기법을 변환 행렬 관점에서 통합적으로 비교할 수 있다는 특징과 함께, 주어진 프라이버시 조건을 만족시키면서 정확도 관점에서 최적인 변환 행렬을 찾을 수 있다는 장점을 지니고 있다. Agrawal-Harista[4]는 이 FRAPP 개념을 연관 규칙 마이닝에 적용한 결과를 보여주었으며, Dowd-Xu-Zhang [5]은 결정 나무(decesion tree) 마이닝에 적용한 메커니즘을 소개하였다. 두 연구 결과에 사용된 랜덤화 알고리즘은 동일한 것으로 랜덤 대치(random substitution) 기법이라 부른다. 랜덤 대치 기법은 앞에서 살펴본 랜덤 회전 및 랜덤 사영 기법에서와 달리 데이터마이닝의 구체적인 방법에 의존하지 않고 데이터를 변형하고 재구축하기 때문에 모든 데이터마이닝 분야에 적용 가능하다는 장점을 지니고 있다. 그러므로 연관규칙 마이닝, 결정나무 마이닝과 같은 분류, 군집화, 아웃라이어 판별

등의 다양한 데이터마이닝 기술에서 프라이버시를 보호하기 위한 실용적인 방법으로 사용 가능하다.

1.2 연구 결과가 기여하는 점

본 논문에서는 실용적인 프라이버시 보호 기술 중 현재까지 가장 발전된 형태를 보이고 있는 행렬 기반 랜덤화 기법에 대하여 심도 있는 분석을 실시한다. 그 결과 첫째, 최적의 변환 행렬을 찾기 위한 조건수(condition number) 개념을 새롭게 적용하여 FRAPP 관련 파라미터들 간의 관계를 세밀하게 규명하여, 랜덤 대치에 사용되는 파라미터들과 정확도를 측정하기 위한 오차 한계에 대한 관계식을 이론적으로 도출해낸다. 이를 위하여 프라이버시 손상(breach) 관점의 파라미터 γ의 의미를 정확히 분석하고, 변환 행렬의 조건수(condition number)를 행렬 노름에 따라 계산하며, 변환된 데이터의 기댓값과 분산을 이론적으로 계산한다. 그 결과 파라미터 γ, 데이터 속성의 종류를 나타내는 파라미터 N, 그리고 데이터의 개수를 의미하는 파라미터 S가 유기적인 연관성을 가지고 오차 한계를 결정하게 된다는 사실을 입증한다. 원저자들은 이 파라미터들 간의 관계를 실험적으로 확인했을 뿐이며, 이론적으로 정확한 관계식을 밝히지는 못했다. 이러한 관계식은 프라이버시 정도와 계산 효율성을 취사선택할 때, 중요한 기준을 확실하게 제시해준다는 의미가 있다.

다음으로 우리가 얻은 결과는 랜덤 대치 기법의 효율적인 구현을 위하여 데이터 재구축 과정에서 필요로 하는 역행렬 구하는 공식이다. 이 역행렬 구하는 공식은 사용된 행렬의 특수성에 기인한 것으로 일반적인 가우스 소거법에 비하여 대단히 단순하고 효율적인 것이다. 이 공식은 우리가 시뮬레이션 과정에서 처음 발견하여 이론적으로 증명한 것으로 원전에는 나타나 있지 않은 내용이다.

위와 같이 이론적으로 얻은 연구 결과의 실증적 분석을 위하여 우리는 랜덤 대치 기법을 구현하고 다양한 시뮬레이션을 실시한다. 그 결과로 정규 분포와 균등 분포를 따르는 데이터에 대한 구현 결과로부터 변형된 데이터는 모두 균등 분포에 가까운 분포를 보인다는 것을 관찰하여, 균등 분포를 따르는 경우에는 재구축 과정이 필요하지 않다는 사실을 밝힌다. 이론적으로 구한 오차의 한계는 파라미터들인 γ, N, S의 다양한 값들에 대한 시뮬레이션에서도 유사한 결과를 관찰할 수 있음을 보인다. 그리고 랜덤 대치 알고리즘의 프라이버시 보호 정도는 사전확률 및 사후확률과 파라미터 γ의 관계식을 통한 분석과 알고리즘 수행 후에 실제로 변화된 데이터의 비율을 실험적으로 조사한 분석을 통하여 실증적으로 규명한다.

1.3 논문의 구성

본 논문은 서론을 포함하여 총 여섯 개의 장으로 이루어져 있다. 2장에서는 행렬 기반 랜덤화 기법과 프라이버시 손상에 대한 내용을 소개한다. 3장에서는 데이터 변환 행렬로 사용되는 γ-대각 행렬의 조건수와 정확도에 대한 결과를 제시하며, 4장에서는 랜덤 대치 기법의 정확성 및 프라이버시와 관련된 파라미터들 간의 관계식을 도출해낸다. 5장에서는 랜덤 대치 기법의 구현 방법 및 다양한 시뮬레이션 결과를 기술하고, 6장은 결론부이다.

Ⅱ. 행렬 기반 랜덤화 기법과 프라이버시

행렬 기반 랜덤화 기법은 기존의 랜덤화 기법들을 변환 행렬 관점에서 통합적으로 관찰할 수 있는 방법으로 현재까지 가장 발전된 형태의 랜덤화 기법이라 할 수 있다. 본 장에서는 행렬 기반 랜덤화 기법의 대표적 연구 결과인 Agrawal-Harista[4]의 FRAPP 개념을 기술한다. 이를 위하여 먼저 프라이버시 보호의 정도를 측정하는 측도로 널리 사용되고 있는 프라이버시 손상(privacy breach)에 대하여 살펴본다.

2.1 프라이버시 손상

프라이버시 손상(privacy breach) 개념은 기존의 상호정보(mutual information) 개념을 이용한 프라이버시 측도의 대안으로 Evfimievski-Gehrke-Srikant[7]가 제안한 것이다. Agrawal-Agrawal[8]은 신뢰구간의 길이로 프라이버시를 측정하는 기존 방법의 불합리성을 설명하면서 Shannon의 정보이론에 입각한 상호정보 기반의 측도를 제안하였다. 상호정보를 이용한 프라이버시 측도가 상당히 일반적이고 합리적인 것처럼 보이지만 모든 상황에서 합리적으로 적용할 수 있는 것은 아니다. 상호정보를 사용한 프라이버시 측도는 평균(average)의 의미가 강한 측도이지 드물게 발생하는 노출 가능성까지를 탐지해내지는 못하기 때문이다. 프라이버시 손상

개념은 이러한 노출 위험성을 탐지해낼 수 있다. 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. 프라이버시 관련 성질들이 지정된 경우라면 현재까지 제안된 측도 중에서는 프라이버시 손상 개념이 가장 보편적인 것으로 보인다.

프라이버시 손상이란 한 정보 제공자 C_i에 의해 데이터 마이너에게 제공된 랜덤화 이후의 정보 y_i가 높은 확률로 C_i의 프라이버시와 관련된 어떤 성질을 노출시킬 때의 상황을 의미한다. 정보 제공자 C_i의 개인정보를 x_i라 하고, 모든 x_i는 고정된 유한집합 V_X에 속한다고 가정한다. 개인정보 x_i가 랜덤화된 이후의 값을 y_i라 하고, 모든 가능한 결과물 y_i의 집합을 V_Y라 하자. 프라이버시 손상을 보다 명확하게 정의하기 위해서 C_i의 개인정보에 대해서 아무런 지식도 없는 상태에서 특정 성질에 대한 정보를 얻어낼 가능성을 사전확률(prior probability), y_i가 주어졌을 때 특정 성질에 대한 정보를 얻어낼 가능성을 사후확률(posterior probability)이라 정의한다.

데이터 마이너의 입장에서, C_i의 개인정보가 가질 수 있는 값 x는 확률 $p_X(x)$를 갖는 확률변수 X를 정의하자. 즉, $\Pr[X=x]=p_X(x)$이다. 확률변수 X는 x_i에 대한 마이너의 사전지식을 가장 잘 표현해준다. 정보 제공자 C_i가 마이너에 제공한 y_i는

$$\Pr[X=x,\ Y=y]=p_X(x)\cdot p[x\to y]$$

를 만족하는 확률변수 Y의 한 표본 값으로 볼 수 있다. 만일 y_i가 주어진다면 마이너는 C_i의 개인정보 관련 확률을 좀 더 효과적으로 계산할 수 있다. 즉, 베이즈 정리를 이용해서 다음의 사후확률을 계산한다.

$$\Pr[X=x\mid Y=y_i]=\frac{\Pr[X=x]\cdot p[x\to y_i]}{\Pr[Y=y_i]}.$$

더욱이 $Q: V_X\to\{true, false\}$를 만족하는 임의의 성질 $Q(X)$의 사후확률도 다음과 같이 찾을 수 있다.

$$\Pr[Q(X)\mid Y=y_i]=\sum_{Q(x),x\in V_X}\Pr[X=x\mid Y=y_i].$$

일반적으로 프라이버시 손상은 어떤 성질 Q에 대하여 마이너에게 제공된 y_i의 제공이 이 성질의 노출 확률을 눈에 띄게 증가시키는 상황을 의미한다. Evfimievski-Gehrke-Srikant[7]에 나타나 있는 프라이버시 손상의 엄밀한 정의는 다음과 같다.

정의 2.1. 성질 Q에 대하여 상향식(upward) $\rho_1\to\rho_2$ 프라이버시 손상이 발생한다는 의미는 $\Pr[Q(X)]\le\rho_1$이고 $\Pr[Q(X)\mid Y=y]\ge\rho_2$를 만족하는 어떤 $y\in V_Y$가 존재한다는 것이다.
반대로 하향식(downward) $\rho_2\to\rho_1$ 프라이버시 손상이 발생한다는 의미는 $\Pr[Q(X)]\ge\rho_2$이고 $\Pr[Q(X)\mid Y=y]\le\rho_1$을 만족하는 어떤 $y\in V_Y$가 존재한다는 것이다. 여기에서 $0<\rho_1<\rho_2<1$이고 $\Pr[Y=y]>0$이다.

다음으로 임의의 성질에 대하여 정보를 노출하지 않는 랜덤화 작용소(randomization operator)에 대한 조건을 정의한다. 실용적으로 유용한 랜덤화 작용소는 같은 $y\in V_Y$에 대해서 각각 다른 $x\in V_X$에 대한 작용소의 변환확률 $p[x\to y]$를 비교해봄으로써 정보 노출의 정도를 측정할 수 있다. 모든 x값이 유사한 가능성을 가지고 y로 랜덤화 된다면, 직관적으로 "$R(x)=y$"는 x에 관한 정보를 많이 노출하지 않는다고 볼 수 있다. 이러한 관점에서 랜덤화 작용소의 정보 노출 정도를 다음의 증폭(amplification) 개념으로 측정한다.

정의 2.2. 랜덤화 작용소 $R(x)$가 $y\in V_Y$에 대하여 기껏해야 γ-증폭(γ-amplifying)이라는 의미는

$$\forall x_1, x_2\in V_X,\ \frac{p[x_1\to y]}{p[x_2\to y]}\le\gamma$$

가 성립한다는 것이다. 여기에서 $\gamma\ge1$이고, $\exists x,\ p[x\to y]>0$이다. 그리고 모든 $y\in V_Y$에 대해서 기껏해야 γ-증폭일 경우에 작용소 $R(x)$는 기껏해야 γ-증폭이라고 말한다.

정리 2.1.(Evfimievski-Gehrke-Srikant[7]) R을 랜덤화 작용소, $y\in V_Y$를 $\exists x,\ p[x\to y]>0$을 만족하는 랜덤화된 값이라 하고 정의 2.1로부터의 $0<\rho_1<\rho_2<1$을 두 확률 값이라 하자. R은 y에 대해서 기껏해야 γ-증폭이라고 가정하자. 그

러면 "$R(x)=y$"로부터 노출되는 정보는

$$\frac{\rho_2}{\rho_1}\cdot\frac{1-\rho_1}{1-\rho_2}>\gamma$$

를 만족할 때, 상향식 $\rho_1\rightarrow\rho_2$ 프라이버시 손상과 하향식 $\rho_2\rightarrow\rho_1$ 프라이버시 손상이 발생하지 않는다.

정의 2.3. 랜덤화 작용소 R이 정리 2.1의 조건을 만족할 경우에 R은 (ρ_1, ρ_2) 프라이버시 보증을 지지한다고 말한다.

2.2 행렬 기반 랜덤화 기법

Agrawal-Harista[4]에 의해서 제안된 FRAPP 프레임워크는 기존의 랜덤화 기법을 행렬 이론 관점에서 통합적으로 기술할 수 있는 좋은 도구이다. 여기에서는 행렬 이론 관점에서 FRAPP 프레임워크를 보다 명확히 분석하기 위하여 원 저자들의 작업을 수학적으로 단순화 시켜서 이론을 전개하기로 한다. 먼저 정보 제공자는 이산(discrete) 형태의 정의역을 갖는 단일 속성 A에 대한 데이터 레코드들을 가지고 있다고 가정한다. 연속(continuous) 형태의 데이터는 적절한 구간으로 나눌 경우 이산 형태로 변환하는 것이 용이하다. 그리고 다수의 속성을 갖는 데이터 집합에 대해서는 단일 속성에서 전개한 논리를 데이터 레코드의 형태가 벡터 값인 경우로 확장함으로써 원하는 결과를 얻을 수 있다.

2.2.1 원본 데이터의 변형 과정

FRAPP 프레임워크에서 저자들은 프라이버시 손상 관점의 안전성 요구조건과 행렬의 조건수(condition number) 관점의 정확도 요구조건을 만족시키는 최적의 랜덤화 기법 중의 하나를 제안하였다. 이 랜덤화 기법을 Dowd-Xu-Zhang[5]은 랜덤대치 기법(random substitution)이라 명명하였다. 랜덤 대치의 기본적인 아이디어는 각 데이터 레코드의 속성 값을 어떤 확률 모델에 따라 속성의 정의역으로부터 랜덤하게 선택된 다른 값으로 바꾸는 것이다. 이 확률 모델은 각 속성 값이 바뀔 확률을 나타내는 전이행렬(transition matrix)을 생성하여 정의할 수 있다. 속성의 정의역을 $U=\{u_1,\cdots,u_N\}$라 가정하고 한 데이터의 속성 값 u_k가 u_h로 바뀔 확률을 다음과 같이 정의한다.

$$\Pr[u_k\rightarrow u_h]=m_{hk}$$

이렇게 정의된 확률 값 m_{hk}를 성분으로 하는 $N\times N$ 크기의 행렬을 M이라 놓는다. 각 속성 값은 자기 자신을 포함해서 반드시 U 안에 있는 값으로 바뀌기 때문에 각 열의 합은 1이 된다. 그러므로 행렬 M에서 각 열은 합이 1인 확률분포로 정의 될 수 있고, 열의 누적 분포함수(cumulative distribution function)를 이용함으로써 속성 값을 변형할 수 있다. 랜덤 대치 기법으로 데이터를 변형하는 방법을 알고리즘으로 표현하면 다음과 같다.

〔표 1〕 랜덤 대치에 의한 데이터 변형 알고리즘

알고리즘 1. 랜덤 대치 알고리즘

입력 : n 개의 레코드로 이루어진 원본 데이터 집합 O, 속성 A에 대한 정의역 $U=\{u_1,\cdots,u_N\}$, U에 대한 전이행렬 $M_{N\times N}$.

출력 : 변환된 데이터 레코드 집합 P.

랜덤화 과정 : 모든 레코드 $o\in O$ 에 대하여 다음 과정을 수행한다.

1. o가 가지는 속성의 인덱스 값 k를 구한다. 즉, o가 가지는 속성 값은 u_k이다.
2. (0, 1) 상의 균등(uniform) 분포로부터 랜덤수 r을 선택한다.
3. 다음을 만족하는 정수 $1\leq h\leq N$를 찾는다.
$$\sum_{i=1}^{h-1}m_{ik}<r\leq\sum_{i=1}^{h}m_{ik}$$
4. o에 대응되는 변환된 레코드 $p\in P$의 속성값을 인덱스 값이 h인 u_h로 결정한다.

랜덤 대치 데이터 변형 알고리즘의 계산복잡도는 $O(n\cdot N)$이다. 정리 2.1의 프라이버시 손상 관점의 안전성 요구조건을 만족하고, 최소 조건수 관점의 정확도 요구조건 하에서, 정의 2.2의 γ를 사용한 최적의 변형 행렬은 $M=xG$의 형태를 가지는 γ-대각 행렬이 된다는 사실이 알려져 있다[4]. 이 때, x와 G는 다음과 같이 주어진다.

$$x=\frac{1}{\gamma+N-1},\quad G=\begin{bmatrix}\gamma & 1 & 1 & \cdots\\ 1 & \gamma & 1 & \cdots\\ 1 & 1 & \gamma & \cdots\\ \vdots & \vdots & \vdots & \ddots\end{bmatrix}\tag{1}$$

행렬 $M=(m_{hk})_{N\times N}$을 정의역과 치역이 모두 $U=\{u_1,\cdots,u_N\}$인 랜덤화 작용소로 볼 때, 전이행렬의 조건

$$\forall\,1\leq h,k\leq N,\ m_{hk}\geq 0,\ \sum_{h=1}^{N}m_{hk}=1$$

을 만족하고, 정의 2.3의 (ρ_1, ρ_2) 프라이버시 보증을 지지하기 위하여 다음 조건

$$\forall 1 \le h, k_1, k_2 \le N, \ \frac{m_{h\,k_1}}{m_{h\,k_2}} \le \gamma \le \frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)}$$

을 만족한다. 정확도 관점에서는 행렬의 조건수를 고려할 경우에 전이행렬 $M=(m_{h\,k})_{N\times N}$ 의 조건수는 $1+N/(\gamma-1)$으로 계산되고, 대칭(symmetric) 전이행렬의 경우에는 이 값이 최소값이 된다는 사실이 밝혀져 있다[6]. 그러므로 조건수를 측도로 사용하는 정확도 관점에서는 적절한 제한 조건 아래에서 행렬 $M=(m_{h\,k})_{N\times N}$ 이 최적(optimal)의 선택 중 하나라는 사실을 알 수 있다. 행렬의 조건수에 대해서는 4절에서 좀 더 면밀한 분석이 이루어진다.

2.2.2 원본 데이터 분포의 재구축 과정

분포 재구축 과정은 변형된 데이터 집합으로부터 원본 데이터의 분포를 추정하는 과정이다[4,6]. 데이터 분포의 추정은 변형된 데이터 집합 P와 변형 행렬 M을 이용한다. 즉, M은 제공자와 마이너가 사전에 미리 설정하거나 공개되는 정보이다. 각 $u_i \in U$에 대해서, Y_i를 변형된 데이터 집합에서 u_i의 개수라 하고, X_i를 원본 데이터 집합 내에서 u_i의 개수라고 하자. 즉, $X=(X_1, \cdots, X_N)^T$ 와 $Y=(Y_1, \cdots, Y_N)^T$ 는 각각 원본과 변형된 데이터 집합에서 속성의 도수(frequency)를 표현하는 열벡터이다. 여기에서 원본 데이터와 변형 데이터의 전체 도수는

$$S=\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i$$

라 놓는다. 그러면 주어진 O에 대해서 Y에 대한 기댓값(expectation)은 다음과 같이 구할 수 있다.

$$E[Y]=(E[Y_1], \cdots, E[Y_N])^T = MX$$

만일 M이 가역(invertible)이고, $E[Y]$가 알려진 경우라면, 방정식 $X=M^{-1}E[Y]$를 풀어냄으로써 X를 구할 수 있다. 그러나 X의 분포는 공개되지 않는 정보이므로, 실제로는 $E[Y]$를 MX로부터 계산해낼 수가 없어서 M이 가역 행렬인 경우라도 정확한 X의 값은 알아내기 어렵게 된다.

서버는 정확한 X의 값을 계산해낼 수는 없지만 공개된 정보를 바탕으로 X의 추정값을 구할 수는 있다. X를 추정하기 위해서 변형된 데이터 집합에서 u_i들의 도수를 나타내는 벡터 Y의 관측값 $\boldsymbol{y}=(y_1, \cdots, y_N)$을 이용하면 X에 대한 추정량(estimator) $\hat{X}$의 추정값(estimate) $\hat{\boldsymbol{x}}$을 얻을 수 있다. 즉, 추정량을 확률변수

$$\hat{X}=(\hat{X_1}, \cdots, \hat{X_N})^T = M^{-1}Y$$

로 정의한다. 여기에서 $E[\hat{X}]=M^{-1}E[Y]=X$이므로, 추정량 $\hat{X}$은 기대값이 원래의 값 X와 일치하는 무편향 추정량(unbiased estimator)이 된다. 그러므로 $\hat{X}$를 재구축된 데이터 분포로 놓으면, 통계적으로 원본 데이터 분포와 유사한 분포를 얻게 되는 것이다.

2.2.3 효율적인 분포 재구축 방법

랜덤 대치 기법에서 원본 데이터의 분포를 추정하기 위한 통계량은 전 소절에서 논한 바와 같이 추정량 $\hat{X}=M^{-1}Y$로 주어지므로, 변형된 데이터 집합 P로부터 관측된 값인 $\boldsymbol{y}=(y_1, \cdots, y_N)$를 바탕으로 실제 추정값은 $\hat{\boldsymbol{x}}=M^{-1}\boldsymbol{y}$가 된다. 서버에서 이 추정값을 구하기 위해서는 M으로부터 역행렬 M^{-1}를 계산해내는 과정이 필수적이다. 가우스 소거법(Gaussian elimination) 등과 같은 일반적인 방법을 통하여 역행렬을 구할 수는 있지만[9], 우리는 M과 같이 특수한 형태의 행렬인 경우 매우 간단하고 효율적인 방법에 의해서 다음 정리에서와 같이 역행렬을 구할 수 있다는 사실을 발견하였다.

정리 2.2. $\gamma>1$, $N>1$인 γ-대각 행렬 M에 대해서, M의 역행렬 $M^{-1}=(m_{ij}^{-1})_{N\times N}$은 다음과 같다.

$$m_{ij}^{-1} = \begin{cases} \dfrac{\gamma+N-2}{\gamma-1}, & i=j \\ \dfrac{1}{1-\gamma}, & i\neq j \end{cases}.$$

증명. 우리는 위와 같은 성분으로 주어진 M^{-1}와 γ-대각 행렬 M에 대해 $M^{-1}M=MM^{-1}=I$가 성립함을 보이면 된다. 먼저 M과 M^{-1}가 정리에서 가정한 것처럼 대각 성분과 비대각 성분이 각각 일정한 값을 갖는 행렬이라면, 곱셈에 대한 교환법칙

$M^{-1}M = MM^{-1}$가 성립함을 확인할 수 있다. 이제 $MM^{-1} = I$ 만을 확인하면 된다. $B = MM^{-1}$ 라고 하자. B의 임의의 대각 성분인 i행 i열 성분과 비대각 성분인 i행 j열($i \neq j$)의 성분을 각각 b_{ii}, b_{ij}라 하면 행렬 B는 다음 두 가지 형태의 성분들로 이루어진다.

$$\begin{aligned} b_{ii} &= m_{ii}m_{ii}^{-1} + \sum_{\substack{l=1 \\ l \neq i}}^{N} m_{il}m_{li}^{-1} \\ &= \frac{\gamma}{\gamma+N-1} \cdot \frac{\gamma+N-2}{\gamma-1} \\ &\quad + (N-1)\left(\frac{1}{\gamma+N-1} \cdot \frac{1}{1-\gamma}\right) \\ &= 1, \end{aligned}$$

$$\begin{aligned} b_{ij} &= m_{ii}m_{ij}^{-1} + m_{ij}m_{jj}^{-1} + \sum_{\substack{k=1 \\ k \neq i,j}}^{N} m_{ik}m_{kj}^{-1} \\ &= \frac{\gamma}{\gamma+N-1}\frac{1}{1-\gamma} + \frac{1}{\gamma+N-1}\frac{\gamma+N-2}{\gamma-1} \\ &\quad + (N-2)\left(\frac{1}{\gamma+N-1} \cdot \frac{1}{1-\gamma}\right) = 0. \end{aligned}$$

즉, B는 b_{ii}가 모두 1이고 b_{ij}가 모두 0인 항등행렬이 되므로 $B = MM^{-1} = I$ 가됨을 알 수 있다.

정리 2.2에서와 같이 M^{-1}를 계산할 경우, 실제로 데이터 분포 재구축 과정에서 가우스 소거법 등과 같은 복잡한 역행렬 계산 알고리즘 수행에 걸리는 시간을 크게 단축시켜 주므로 분포의 재구축 과정을 효율적으로 진행시킬 수 있다. 실제로 가우스 소거법에 의한 역행렬 계산 알고리즘의 복잡도는 행렬의 크기인 N에 민감하게 의존하여 곱셈 연산 기준으로 $O(N^3)$ 정도의 계산량을 나타낸다[9]. 하지만 정리 1을 사용할 경우에는 행렬의 크기 N과 무관하게 대단히 작은 계산량만으로 역행렬 M^{-1}을 얻을 수 있는 것이다.

Ⅲ. 변환 행렬의 조건수와 정확도

랜덤 대치 기법에서 사용되는 전이 행렬은 프라이버시 손상 관점의 안전성 조건을 만족하고 정확도 관점에서는 최소 조건수를 갖는다는 의미에서 최적인 γ-대각 행렬이다. 본 절에서는 행렬의 조건수를 보다 세밀하게 연구함으로써 구현 시에 중요한 요소인 행렬의 크기 N과 파라미터 γ 사이의 상관관계를 좀 더 명확히 규명하고자 한다.

3.1 행렬 노름과 변환 행렬의 조건수

행렬의 조건수는 일차연립방정식(linear system)과 연계하여 살펴보는 것이 자연스럽다. 일차연립방정식 $A\boldsymbol{x} = \boldsymbol{b}$가 주어지고 정방행렬 A가 정칙(nonsingular)일 때, 연립방정식의 계수를 약간 변화시킨 경우의 연립방정식을 $(A+B)\tilde{\boldsymbol{x}} = \boldsymbol{b}$라 하면, $\boldsymbol{x} = A^{-1}\boldsymbol{b}$이고 $\tilde{\boldsymbol{x}} = (A+B)^{-1}\boldsymbol{b}$가 된다. 행렬의 노름(norm)을 $\|\cdot\|$으로 표현할 경우

$$\begin{aligned} \|A^{-1} - (A+B)^{-1}\| &\approx \|A^{-1}BA^{-1}\| \\ &\leq \|A^{-1}\|\|B\|\|A^{-1}\| \end{aligned} \tag{2}$$

가 성립한다[9]. 여기에서 "$\approx$"는 근사값을 의미한다. 이 수식 (2)를 이용할 경우 우리는

$$\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}\|\|A\|\left(\frac{\|B\|}{\|A\|}\right) \tag{3}$$

라는 사실을 얻을 수 있다. 수식 (3)의 좌변은 일차연립방정식 해(solution)의 상대적 변화량을 나타내고, 우변은 계수 행렬의 상대적 변화량에 상수배를 곱한 양이다. 그러므로 수식 (3)은 일차연립방정식의 해의 변화 정도는 계수 행렬의 변화 정도와 상수 $\kappa = \|A^{-1}\|\|A\|$의 곱을 상계(upper bound)로 갖는다는 사실을 보여준다. 이를 바탕으로 행렬의 조건수를 다음과 같이 정의한다.

정의 3.1. A가 정칙(nonsingular) 행렬일 때, 일차연립방정식 $A\boldsymbol{x} = \boldsymbol{b}$에 대한 해(solution)의 민감성(sensitivity)은 $\kappa = \|A^{-1}\|\|A\|$에 의하여 측정되고, 이 κ를 행렬 A의 조건수(condition number)라 부른다.

한편, 행렬의 조건수는 그 행렬의 노름 값에 따라서 달라질 수 있음을 알 수 있다. 먼저 여기에서 사용할 두 가지 행렬의 노름을 정의하자.

정의 3.2. 행렬 A의 2-노름 $\|A\|_2$와 프로베니우스(Frobenius) 노름 $\|A\|_F$는 다음과 같이 정의된다.

$$\|A\|_2 = \max_{\|\boldsymbol{x}\|=1}\|A\boldsymbol{x}\|$$

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}.$$

여기에서 벡터의 노름은 유클리드 노름을 의미하고 a_{ij}는 행렬 A의 i행 j열 성분을 의미한다.

행렬의 2-노름은 그 행렬의 고유값(eigenvalue)에 의해서 다음과 같이 구해진다는 사실이 잘 알려져 있다. 즉, $\|A\|_2=\lambda_{\max}$, $\|A^{-1}\|_2=\frac{1}{\lambda_{\min}}$ 이 성립한다. 그러므로 행렬의 2-노름을 사용할 경우, 행렬 A의 조건수는 $\kappa=\frac{\lambda_{\max}}{\lambda_{\min}}$가 됨을 알 수 있다. 여기에서 $\lambda_{\max}$와 $\lambda_{\min}$은 각각 행렬 A의 최대 고유값과 최소 고유값을 의미한다. 행렬의 2-노름에 기반한 이와 같은 조건수에 대하여 Agrawal-Haritsa[6]는 2절에서 논한 γ-대각행렬이 대칭인 $N\times N$ 전이행렬 중에서 최소값인 $1+\frac{N}{\gamma-1}$를 조건수로 갖는다는 사실을 밝혔던 것이다. 이 값을 정확도에 관한 측도로 사용한다면, 행렬의 크기와 연관된 양인 N의 크기가 커지면, 이에 비례하여 조건수가 커지므로 정확도는 낮아질 것이다. 그리고 프라이버시 측도와 연관된 γ가 커지면 조건수가 작아져서 정확도가 높아진다는 직관적인 사실을 알 수 있다. 하지만 다른 행렬 노름을 사용할 경우에도 이러한 관계식들이 성립하는지는 확실치 않다. 우리는 행렬의 2-노름 대신에 프로베니우스 노름을 사용함으로써 이러한 궁금증을 해결해 보고자 한다.

정리 3.1. $\gamma>1$, $N>1$인 γ-대각 행렬 M에 대해서, 프로베니우스 노름에 의한 M의 조건수는 다음과 같이 주어진다.

$$\begin{aligned}\kappa &= \|M\|_F\|M^{-1}\|_F \\ &= O\left(\sqrt{\frac{N^5}{\gamma^2N^2+\gamma^3N+\gamma^4}}\right)\end{aligned}$$

증명. 행렬 $M=(m_{ij})_{N\times N}$의 성분은 $1\le i\ne j\le N$에 대하여,

$$m_{ii}=\frac{\gamma}{\gamma+N-1},\quad m_{ij}=\frac{1}{\gamma+N-1}$$

이므로

$$\begin{aligned}\|M\|_F^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N}m_{ij}^2 \\ &= N\left(\frac{\gamma}{\gamma+N-1}\right)^2+(N^2-N)\left(\frac{1}{\gamma+N-1}\right)^2\end{aligned}$$

이고, 역행렬 $M^{-1}=(m_{ij}^{-1})_{N\times N}$은 성분은 정리 2.1에 의하여 $m_{ii}^{-1}=\frac{\gamma+N-2}{\gamma-1}$, $m_{ij}^{-1}=\frac{1}{\gamma}$ 이므로

$$\begin{aligned}\|M^{-1}\|_F^2 &= N\left(\frac{\gamma+N-2}{\gamma-1}\right)^2 \\ &\quad +(N^2-N)\left(\frac{1}{1-\gamma}\right)^2\end{aligned}$$

이다. 이로부터 우리는

$$\begin{aligned}\kappa^2 &= \|M\|_F^2\cdot\|M^{-1}\|_F^2 \\ &= O\left(\frac{N^5}{\gamma^2N^2+\gamma^3N+\gamma^4}\right)\end{aligned}$$

임을 알 수 있다.

3.2 행렬 노름에 따른 조건수와 재구축 분포의 정확도

행렬의 조건수는 행렬 노름에 의존하여 우리가 관심있는 γ-대각행렬에 대해서도 행렬의 2-노름과 프로베니우스 노름에 따라 서로 다른 조건수가 계산됨을 알 수 있었다. γ-대각행렬 $M=(m_{ij})_{N\times N}$에 대하여 2-노름에 의한 조건수를 $\kappa_2(M)$이라 하고, 프로베니우스 노름에 의한 조건수를 $\kappa_F(M)$이라 놓자. 2-노름에 대한 계산 결과인 Agrawal-Haritsa[6]의 연구 결과와 프로베니우스 노름에 대한 결과인 정리 3.1에 의하면 다음과 같은 결과를 얻을 수 있다.

정리 3.2. $\gamma>1$, $N>1$인 γ-대각 행렬 M에 대해서, 2-노름과 프로베니우스 노름에 의한 M의 조건수는 다음을 만족한다.

$$\kappa_2(M)=1+\frac{N}{\gamma-1},\quad \kappa_F(M)\approx 1+\frac{N^{3/2}}{\gamma^2}.$$

증명. $\kappa_2(M)$에 대한 계산 결과는 Agrawal-Haritsa[6]에 나타나 있다. 정리 3.1로부터 $\kappa_F(M)$는 γ와 N에 대한 분자, 분모의 최고차항을 고려하면,

$$\kappa_F(M)\approx\left(\frac{N^5+\gamma^4N^2}{\gamma^4N^2}\right)^{1/2}\approx 1+\frac{N^{3/2}}{\gamma^2}$$

이 성립함을 알 수 있다.

행렬 이론에 기반한 랜덤화 기법인 FRAPP 프레임워

크에서는 정확도에 대한 측도로 행렬의 조건수를 고려한다. 직관적으로 행렬의 조건수가 1에 가까워지면 행렬에 의해서 변환된 벡터의 노름 값이 별로 변하지 않으므로 정확도가 높아진다는 것이 행렬의 조건수에 의한 정확도 측정 방법이다. 변환 행렬의 조건수가 증가하게 되면 변환된 벡터의 노름과 원본 벡터의 노름 값의 차이도 증가하여 분포 재구축 시 정확도는 떨어질 것이다.

프라이버시 손상 관점의 안전성 조건을 만족한 상태에서 최소의 조건수를 가진 변환 행렬은 γ-대각 행렬이었다. 행렬의 조건수를 최소값으로 만드는 변환 행렬의 제약 조건은 대칭인 확률 전이 행렬이어야 한다는 것이고, 여기에 사용된 행렬의 노름은 2-노름이었다. 이와 같이 행렬의 2-노름을 사용할 경우에 이론 전개가 확실하고, 행렬의 조건수 관점에서 최적인 변환 행렬을 γ-대각 행렬과 같이 명시적으로 얻을 수 있다는 장점이 있다. 그런데 행렬의 2-노름은 원본 벡터의 노름 값이 1인 벡터에 대하여 최대의 변화량만을 탐지해내는 측도임을 주목해볼 필요가 있다. 변환된 벡터의 방향별로 변화량을 비교하여 최대값만을 고려하는 것보다는 모든 성분들에 대한 변화량을 탐지하는 것이 정확도 측정에 보다 공정한 측도가 될 수 있다. 그리고 모든 성분들에 대한 변화량을 고려하기 위해서는 변환 행렬의 모든 성분이 참여하는 프로베니우스 노름을 사용하는 것이 합리적이다.

변환 행렬인 γ-대각 행렬에 관하여 정리 3.2는 행렬의 2-노름과 프로베니우스 노름에 대한 조건수가 약간 다르게 계산됨을 보여준다. 조건수가 두 가지 행렬 노름에 대하여 행렬의 크기 N에 비례하고 프라이버시 관련 파라미터인 γ에는 반비례한다는 사실은 변하지 않지만, N과 γ에 대한 변화 비율은 두 가지 노름에 대하여 확연히 다른 차이를 정리 3.2에서 발견할 수 있는 것이다. 일반적으로 조건수가 작을 때 좀 더 높은 정확도를 보이므로, 행렬의 2-노름을 사용할 경우에는 정확도는 N에 반비례하고 γ에 비례하며, 행렬의 프로베니우스 노름을 사용할 경우에는 정확도가 $N^{3/2}$에 반비례하고 γ^2에 비례한다고 말할 수 있다. 실제로 변환 행렬의 조건수가 재구축 분포의 정확도에 얼마나 직접적인 영향를 미치는지에 대해서는 확실히 알려진 바가 없다. 하지만 다양한 형태의 원본 데이터를 변환시킨다는 관점에서 행렬의 2-노름보다는 프로베니우스 노름이 정확도에 영향을 미치는 파라미터인 N과 γ에 대한 변화 정도를 측정하는 데에는 좀 더 합리적인 도구가 될 것이란 사실은 자명하다. 다음 절에서 우리는 다양한 실험을 통하여 이 문제를 규명해보고자 한다.

Ⅳ. 랜덤 대치 파라미터들 사이의 관계식

변환 행렬로 γ-대각 행렬을 사용하는 랜덤 대치 기법의 정확도는 2절에서 살펴본 바와 같이 $MX = E[Y]$를 만족하는 원본 데이터의 도수를 나타내는 벡터 X와 $\hat{X} = M^{-1}Y$로 계산되는 추정량 $\hat{X}$의 차이를 측정함으로써 알 수 있다. 이를 위해서 행렬의 조건수에 대하여 3절의 수식 (3)의 발전된 형태를 이용하기로 한다. 일차연립방정식 $A\boldsymbol{x} = \boldsymbol{b}$와 $(A+B)\tilde{\boldsymbol{x}} = \boldsymbol{b}$가 주어졌을 때, $(A+B)\tilde{\boldsymbol{x}} = \boldsymbol{b}$는 $A\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}} = \boldsymbol{b} + \boldsymbol{e}$를 만족한다고 볼 수 있어서 두 방정식의 해인 $\boldsymbol{x}$와 $\tilde{\boldsymbol{x}}$ 사이에는 행렬 A의 조건수를 κ라 할 때, 다음 수식이 성립한다는 사실이 선형대수학에서 잘 알려져 있다[9].

$$\frac{1}{\kappa}\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|} \le \frac{\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} \le \kappa\frac{\|\boldsymbol{e}\|}{\|\boldsymbol{b}\|} \tag{4}$$

수식 (4)를 이용하기 위하여 벡터 X와 $\hat{X}$에 관련된 일차연립방정식을 각각 $MX = E[Y]$와 $M\hat{X} = Y$로 표현하면, 다음 정리를 얻을 수 있다.

정리 4.1. $X=(X_1,\cdots,X_N)^T$와 $Y=(Y_1,\cdots,Y_N)^T$는 각각 원본과 변형된 데이터 집합에서 속성의 도수 분포를 표현하는 열벡터이고, 행렬 $M=(m_{hk})_{N\times N}$는 알고리즘 1에 사용된 전이 행렬이며, $\hat{X} = M^{-1}Y$은 분포 재구축 과정에 사용되는 추정량이라 하자. 그러면 행렬 M의 조건수를 κ라 할 때, 다음 관계식이 성립한다.

$$\frac{1}{\kappa}\frac{\|Y-E[Y]\|}{\|E[Y]\|} \le \frac{\|X-\hat{X}\|}{\|X\|} \le \kappa\frac{\|Y-E[Y]\|}{\|E[Y]\|}.$$

정리 4.1은 원본 데이터와 재구축 과정에 의하여 추정된 값 사이의 오차(error)는 두 가지 요소에 의해 영향을 받고 있음을 의미한다. 오차를 결정하는 한 가지 요소는 변환 행렬의 조건수이며, 다른 한 가지 요소는 변형된 데이터 집합에서 얻을 수 있는 확률벡터(random vector) Y의 분산(variance)이다. 확률벡터 Y의 분산을 계산하기 위하여 성분을 이루고 있는 확률변수

$Y_j\,(1 \le j \le N)$를 다음과 같이 표현할 수 있다.

$$Y_j = \sum_{i=1}^{S} Y_j^{\alpha_i},\quad Y_j^{\alpha_i} = \begin{cases} 1, & \text{if } \alpha_i \to u_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

여기에서 원본 데이터 배열은 데이터의 전체 도수가 $S = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i$ 일 때, $O = (\alpha_1, \alpha_2, \cdots, \alpha_S)$로 표현될 수 있음을 가정한 것이다. 즉, $Y_j^{\alpha_i}$는 $\alpha_i = u_k$ 일 때, 성공 확률 $m_{jk} = \Pr[u_k \to u_j]$를 갖는 베르누이(Bernoulli) 확률변수이고, 이들의 합으로 Y_j를 나타낼 수 있다. 한편, 알고리즘 1에서 파라미터 r을 원본 데이터 원소들마다 독립적으로 추출하기 때문에 확률변수 $Y_j^{\alpha_i}\,(1 \le i \le S)$ 들은 서로 독립(independent)이다. 그렇지만 위에서 살펴본 바와 같이 $Y_j^{\alpha_i}\,(1 \le i \le S)$ 들의 분포는 동일하지 않다(not identical). 확률변수 Y_j와 같이 독립적이지만 동일하지 않은 분포를 따르는 확률변수들의 합으로 표현되는 형태의 확률변수를 포아송-이항(Poisson-Binomial) 분포라고 한다[10]. 변형 행렬이 γ-대각 행렬인 경우에 우리는 Y의 분산을 다음 정리와 같이 계산해낼 수 있다.

정리 4.2. $X = (X_1, \cdots, X_N)^T$ 와 $Y = (Y_1, \cdots, Y_N)^T$ 는 각각 원본과 변형된 데이터 집합에서 속성의 도수 분포를 표현하는 열벡터이고, $S = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i$ 일 때, 변환 행렬 $M = (m_{hk})_{N \times N}$ 이 수식 (1)과 같은 γ-대각 행렬로 주어질 경우 Y의 분산은 다음과 같이 계산된다.

$$Var(Y) = E[\| Y - E[Y] \|^2] = \frac{(N-1)(N+2\gamma-2)S}{(\gamma+N-1)^2}$$

여기에서 $\| \cdot \|$는 벡터의 유클리드 노름이다.

증명. 수식 (5)로부터 Y_j는 서로 독립인 확률변수들인 $Y_j^{\alpha_i}$ 들의 합이기 때문에

$$Var(Y_j) = \sum_{i=1}^{S} Var(Y_j^{\alpha_i})$$

가 성립하므로 $Var(Y_j^{\alpha_i})$를 먼저 계산한다. $Y_j^{\alpha_i}$는 베르누이 확률변수 이므로 $\alpha_i = u_k$ 일 때,

$$Var(Y_j^{\alpha_i}) = m_{jk}(1 - m_{jk})$$

이다. 그러므로 원본 데이터 속성의 도수 분포가 $X = (X_1, \cdots, X_N)^T$ 임을 이용하여

$$\begin{aligned} Var(Y_j) &= X_1 \cdot m_{j1}(1-m_{j1}) + X_2 \cdot m_{j2}(1-m_{j2}) + \cdots + X_N \cdot m_{jN}(1-m_{jN}) \\ &= X_j \cdot \frac{\gamma}{\gamma+N-1}\left(1 - \frac{\gamma}{\gamma+N-1}\right) + (X_1 + \cdots + X_{j-1} + X_{j+1} + \cdots X_N) \cdot \frac{1}{\gamma+N-1}\left(1 - \frac{1}{\gamma+N-1}\right) \\ &= X_j \cdot \frac{(N-1)\gamma}{(N+\gamma-1)^2} + (S - X_j) \cdot \frac{N+\gamma-2}{(N+\gamma-1)^2} \end{aligned}$$

을 얻을 수 있다. 결과적으로 Y의 분산은

$$\begin{aligned} Var(Y) &= E[\| Y - E[Y] \|^2] \\ &= \sum_{j=1}^{N} E[(Y_j - E[Y_j])^2] \\ &= \sum_{j=1}^{N} Var(Y_j) \\ &= \frac{(N-1)\gamma}{(N+\gamma-1)^2} \sum_{j=1}^{N} X_j + \frac{N+\gamma-2}{(N+\gamma-1)^2} \sum_{j=1}^{N} (S - X_j) \\ &= \frac{(N-1)S}{(N+\gamma-1)^2}(\gamma+N+\gamma-2) \\ &= \frac{(N-1)(N+2\gamma-2)S}{(N+\gamma-1)^2} \end{aligned}$$

와 같이 계산된다.

정리 4.1과 4.2로부터 우리는 직관적으로 분포 재구축 과정으로부터 추정된 데이터 분포의 정확도는 변환 행렬의 조건수와 Y의 분산에 의존하고, 이들은 다시 파라미터 γ, N, S등과 유기적으로 연관되어 있음을 알 수 있다. FRAPP 프레임워크[4,6]의 원저자들은 행렬의 2-노름에 기반하여 변환 행렬의 조건수를 계산해내는 데에 그쳤지만 우리는 정리 4.1과 4.2에서 보는 바와 같이 변환 행렬의 조건수 뿐만 아니라 변환된 분포의 분산에 대해서도 파라미터들과의 연관성을 밝힐 수 있었다. 2절에서 살펴본 분포 재구축 과정과 정리 3.2, 정리 4.1, 정리 4.2를 종합적으로 고려함으로써 분포 재구축 과정에서 발생할 수 있는 오차에 대한 다음 정리를 얻을 수 있다.

정리 4.3. $X=(X_1,\cdots,X_N)^T$ 와 $Y=(Y_1,\cdots,Y_N)^T$ 는 각각 원본과 변형된 데이터 집합에서 속성의 도수 분포를 표현하는 열벡터이고, 변환 행렬 $M=(m_{hk})_{N\times N}$ 이 수식 (1)과 같은 γ-대각 행렬이며, $\widehat{X}=M^{-1}Y$ 은 분포 재구축 과정에 사용되는 추정량이라 하자. 그러면 원본 데이터와 추정된 데이터 분포 사이의 오차를 나타내는 다음 관계식을 만족한다.

$$\frac{sd(\widehat{X})}{\|X\|} \le \kappa_2 \frac{sd(Y)}{\|E[Y]\|} \approx \frac{N(N+2\gamma)^{1/2}}{\gamma S^{1/2}} = O\left(\frac{N^{3/2}}{\gamma^{1/2}S^{1/2}}\right),$$

$$\frac{sd(\widehat{X})}{\|X\|} \le \kappa_F \frac{sd(Y)}{\|E[Y]\|} \approx \frac{(N^{3/2}+\gamma^2)N(N+2\gamma)^{1/2}}{(N+\gamma)\gamma^2 S^{1/2}} = O\left(\frac{N^2}{\gamma^{1/2}S^{1/2}}\right),$$

여기에서 $sd(\widehat{X})=\sqrt{Var(\widehat{X})}=\sqrt{E[\|\widehat{X}-X\|^2}$ 이고, $sd(Y)=\sqrt{Var(Y)}$ 으로 각각 확률벡터 $\widehat{X}$ 와 Y 의 표준편차(standard deviation)를 의미한다.

증명. 정리 3.2, 정리 4.1, 그리고 정리 4.2로부터 다른 항들은 계산 가능하므로 $\|E[Y]\|$ 에 대한 계산 결과를 이끌어내면 증명은 끝난다. $E[Y]=MX$ 라는 사실과 행렬의 각 성분에 관한 수식 (1)로부터

$$\begin{aligned}
&\|E[Y]\|^2 \\
&= \left(\frac{\gamma}{N+\gamma-1}X_1+\frac{1}{N+\gamma-1}(S-X_1)\right)^2 + \\
&\quad \cdots+\left(\frac{\gamma}{N+\gamma-1}X_N+\frac{1}{N+\gamma-1}(S-X_N)\right)^2 \\
&= \frac{1}{(N+\gamma-1)^2} \\
&\quad \{(S+(\gamma-1)X_1)^2+\cdots+(S+(\gamma-1)X_N)^2\} \\
&= \frac{1}{(N+\gamma-1)^2} \\
&\quad \{NS^2+2(\gamma-1)S^2+(\gamma-1)^2(X_1^2+\cdots+X_N^2)\} \\
&\ge \frac{1}{(N+\gamma-1)^2} \\
&\quad \left\{NS^2+S^2\left(2(\gamma-1)+\frac{(\gamma-1)^2}{N}\right)\right\} \\
&\approx \frac{(N^2+2\gamma N+\gamma^2)S^2}{N(N+\gamma}
\end{aligned}$$

을 계산해낼 수 있다. 위 수식의 계산 과정에서 Cauchy-Schwarz 부등식에 의하여 얻을 수 있는

$$(X_1^2+\cdots+X_N^2)\cdot N \ge S^2$$

이 사용되었다. 그러므로 $\|E[Y]\| \ge S/\sqrt{N}$ 임을 이용하여 원하는 결과를 얻을 수 있다.

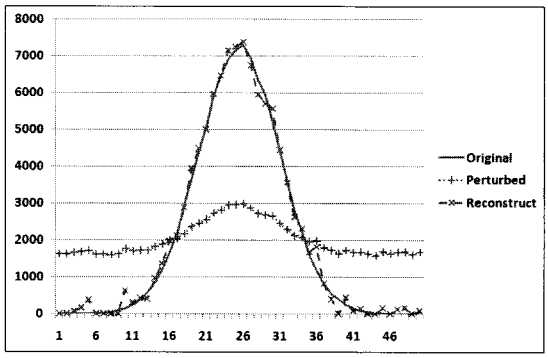
우리는 정리 4.3에서 고려하는 오차 $sd(\widehat{X})/\|X\|$ 를 표준오차(standard error)라 부르기로 한다. 정리 4.3의 내용은 γ-대각 행렬을 변환 행렬로 사용하는 랜덤 대치 기법의 정확성과 사용되는 파라미터 사이의 상관관계를 세밀하게 보여주고 있다. 사용되는 행렬의 노름에 따라서 조건수가 다르게 계산되기 때문에 두 가지의 관계식을 얻었다. 두 관계식 모두 공통으로 프라이버시 관련 파라미터 $\gamma^{1/2}$ 과 처리되는 데이터의 전체 개수 관련 파라미터인 $S^{1/2}$ 에 비례하여 표준오차의 한계는 감소함을 보여준다. 그리고 데이터 속성의 개수와 관련된 파라미터 $N^{3/2}$ 또는 N^2 에 비례하여 표준오차의 한계는 증가한다는 사실을 얻을 수 있다. 이론적으로 $N^{3/2}$ 이 좀 더 좋은(tight) 경계치이기 때문에 행렬의 2-노름을 사용하는 것이 여러 면에서 합리적이라는 것도 알 수 있다.

V. 구현 및 시뮬레이션 결과

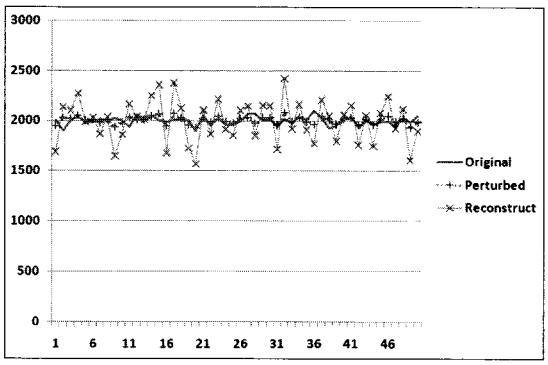
5.1 랜덤 대치 기법의 구현

랜덤 대치 기법의 정확성은 원본 데이터의 분포와 재구축된 분포와의 차이로 정의되는 오차를 통하여 측정될 수 있다. 재구축 분포에 대한 벡터를 $R=M^{-1}\boldsymbol{y}$ 라고 하자. 정리 2.1에서 볼 수 있는 바와 같이 M^{-1}의 비대각 성분이 모두 음수이기 때문에 재구축 과정에서 R의 성분이 음수 값을 가지게 되는 경우가 발생할 수 있다. 미지의 원본 데이터 분포를 나타내는 벡터 X의 성분은 각 속성의 도수를 나타내는 양의 정수이므로 재구축 분포인 R에서 음수 값을 갖는 성분을 0으로 바꿔주면 오차를 좀 더 줄일 수 있다[5]. 이와 같이 오차를 줄이고 속성의 개수를 나타내는 정수 값을 부여하기 위하여 R 을 보정한 추정량 $\widehat{X}$ 을 다음과 같이 정의한다.

$$\forall i \in \{1,\cdots,N\},\ \widehat{X}_i = \begin{cases} 0 & , R_i \le 0 \\ \lfloor R_i \rfloor & , R_i > 0 \end{cases}. \qquad (6)$$



〔그림 2〕 각 데이터 속성 값에 대한 분포 (정규분포)



〔그림 3〕 각 데이터 속성 값에 대한 분포 (균등)

구현에서 데이터 레코드는 정규분포와 균등분포를 따르는 10만개의 레코드로 γ를 2에서 21까지, N은 10부터 100까지 등으로 변화시키면서 실험하였다. [그림 2]는 원본(Original) 데이터의 속성 값이 정규분포를 따르는 경우에 변형된(Perturbed) 데이터의 속성 값 분포와 재구축(Reconstruct) 분포를 나타낸 것 중의 하나이고, [그림 3]은 균등분포에 대한 각 데이터 속성 값의 분포를 나타낸 것 중의 하나이다.

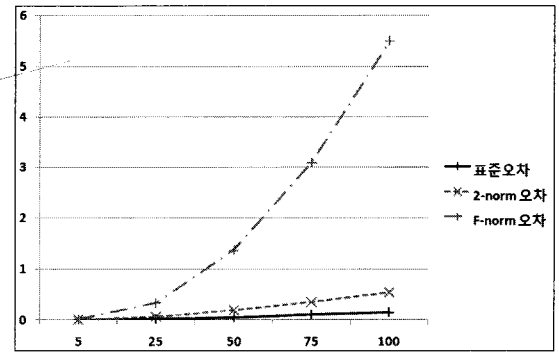
구현을 통한 실험 결과 원본 데이터가 정규분포인 경우 변형된 데이터는 균등 분포와 유사한 분포를 이루고, 재구축된 분포는 원본 데이터와 유사한 분포가 됨을 확인할 수 있다. 하지만 원본 데이터의 분포가 균등분포를 이룰 때는 변형된 데이터의 분포가 원본 데이터와 매우 유사함을 알 수 있다. 물론 이 경우에도 실제 데이터는 많은 차이를 보이고 있으며, 단지 분포만 유사하다는 것이다. 또한, 균등분포의 경우에 재구축된 데이터의 분포는 전체적으로 원본 데이터와 유사한 형태를 이루지만 그 오차는 변형된 데이터 자체보다 오히려 크다는 사실을 알 수 있다. 이러한 현상은 랜덤 대치 알고리즘이 균등분포에 기반을 두고 있다는 점과 최적 행렬로 선택된 γ-대각 행렬이 자기 자신으로 변환된 확률을 나타내는 대각 성분 외에는 동일한 확률 값을 갖기 때문에 발생하는 현상으로 보인다. 이론적으로도 이러한 현상은 간단히 규명할 수 있다. 만일, 원본 데이터 분포가 $X_i \equiv a\,(1 \le i \le N)$로 일정한 값을 갖는 균등분포라 하면 $Y = MX$에 의하여 변형된 데이터 분포 Y는

$$Y_j = \sum_{i=1}^{N} m_{ji} X_i = a\left(\sum_{i=1}^{N} m_{ji}\right) = a, \quad 1 \le j \le N$$

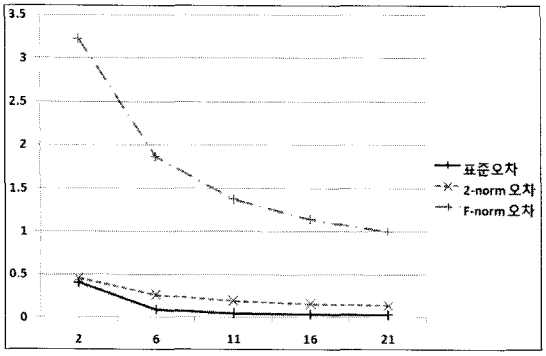
을 만족하여 X와 동일한 균등분포를 따르게 된다는 것을 알 수 있다. 그러므로 원본 데이터가 균등분포를 따른다는 가정이 합리적인 경우에는 재구축 과정을 생략하고 변형된 데이터 자체를 추정값으로 사용하는 것이 오히려 오차를 작게 할 수 있을 것이다.

5.2 정확도 측정에 관한 실험 결과

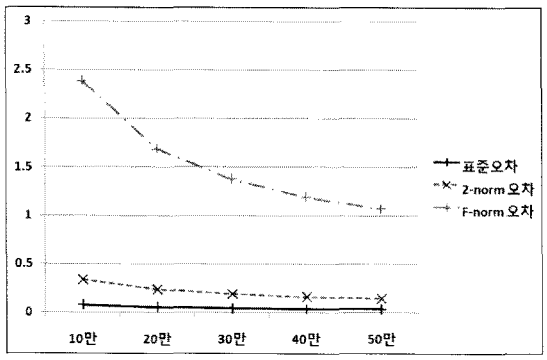
정리 4.3에서 우리는 추정량의 상대적 표준편차를 고려하여 추정 오차인 $sd(\hat{X})/\|X\|$에 대한 계산 결과를 이론적으로 도출하였다. 이 계산 결과에 의하면 표준오차에 대한 한계는 행렬의 2-노름을 사용할 경우에 근사값 $N^{3/2}/\sqrt{\gamma S}$를 갖고, 행렬의 프로베니우스 노름을 사용할 경우에는 근사값 $N^2/\sqrt{\gamma S}$을 갖는다. 이를 실험적으로 규명하기 위하여 N, γ, S를 적절히 변화시킨 경우의 표준오차 값을 관찰하였다. 실험에서 N은 5부터 100 사이의 정수, γ는 2부터 21 사이의 정수에 대하여 관찰하였으며, 데이터의 개수 S는 10만부터 50만까지 10만 단위로 5 가지 경우를 관찰하였다. [그림 4]부터 [그림 6]까지는 각 파라미터에 대한 오차의 변화를



〔그림 4〕 N의 변화에 따른 표준오차와 오차의 한계



〔그림 5〕 γ의 변화에 따른 표준오차와 오차의 한계



〔그림 6〕 S의 변화에 따른 표준오차와 오차의 한계

그래프로 나타낸 것이다. 그림에서 보는 바와 같이 실험 결과는 정리 4.3의 내용을 잘 보여주고 있다. 그림들에서 실제로 구한 표준오차의 값이 정리 4.3에서 구한 두 가지 오차의 한계와 약간의 차이가 있음을 발견할 수 있다. 이는 랜덤 대치 알고리즘을 실제 구현할 때, 수식 (6)에서와 같은 보정을 통하여 오차가 줄어들도록 했기 때문에 발생한 현상일 것으로 생각된다.

우리는 보다 구체적으로 오차를 측정해보기 위하여 합리적으로 생각할 수 있는 오차에 관한 측도를 추가로 고려하여 랜덤 대치 기법의 정확성을 다양한 관점에서 실험하였다. 여기에서 고려한 오차는 다음 추정값과 원본 사이의 표준절대편차, 평균의 차이, 표준편차의 차이 등 세 가지이다.

① 표준절대편차($error_1$) : 전체 레코드 중에서 원본 데이터의 각 속성 값의 개수 X_i와 재구축된 데이터의 각 속성 값의 개수 $\hat{X}_i$ 사이의 표준절대편차.

$$error_1 = \frac{1}{S}\sum_{j=1}^{N}|\hat{X}_i - X_i| \ .$$

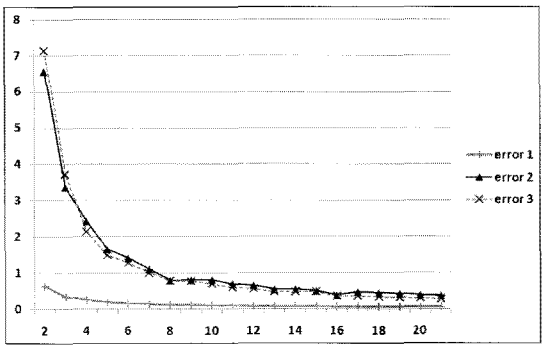
② 평균의 차이($error_2$) : 원본 데이터의 각 속성 값에 대한 평균 μ와 재구축된 데이터의 각 속성 값에 대한 평균 $\hat{\mu}$의 차이.

$$\mu = \frac{1}{S}\sum_{i=1}^{N} u_i X_i \ , \quad \hat{\mu} = \frac{1}{S}\sum_{i=1}^{N} u_i \hat{X}_i \ ,$$
$$error_2 = |\mu - \hat{\mu}| \ .$$

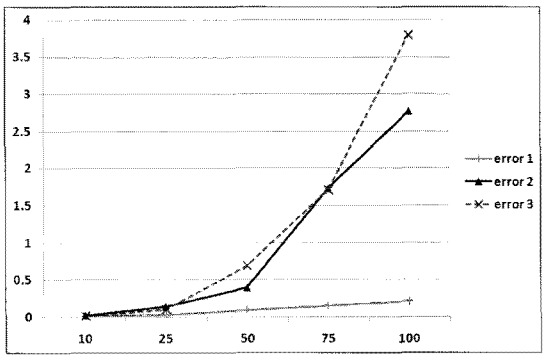
③ 표준편차의 차이($error_3$) : 원본 데이터의 각 속성 값에 대한 표준편차 σ와 재구축된 각 속성 값에 대한 표준편차 $\hat{\sigma}$의 차이.

$$\sigma = \sqrt{\frac{1}{S}\sum_{i=1}^{N}(u_i - \mu)^2} \ , \quad \hat{\sigma} = \sqrt{\frac{1}{S}\sum_{i=1}^{N}(u_i - \hat{\mu})^2}$$
$$error_3 = |\sigma - \hat{\sigma}|$$

[그림 7]에서 정규분포 상의 고정된 N에 대해 γ가 감소할수록 에러가 증가하는 것을 볼 수 있다. 오차의 증가는 정확도의 감소를 뜻하는데 이것은 γ가 감소할수록 변형행렬 M의 대각성분 $\gamma/(\gamma+N-1)$는 작아지고, 비대각 성분 $1/(\gamma+N-1)$은 커지기 때문이다.



〔그림 7〕 정규분포에서 γ에 대한 여러 가지 오차 (N=50)



〔그림 8〕 정규분포에서 N에 대한 여러 가지 오차 (γ=11)

[그림 8]에서 정규분포상의 고정된 γ에 대해 N이 증가할수록 오차는 증가하는 것을 볼 수 있는데, 그 이유는 역행렬의 비대각성분이 N에 관계없이 일정하기 때문이다. N이 크면 각 속성 값의 개수는 전체적으로 작은 값이 많아지게 되고, 주변 값보다 상대적으로 작은 값에 대한 재구축은 음수 값을 갖는 값을 도출해낼 가능성이 커진다. 그러므로 오차 역시 증가하게 되고 정확도가 감소하게 된다.

$error_1$으로는 분포 사이의 전체적인 오차에 대한 측정이 가능하지만, 속성 값들이 얼마만큼의 차이로 변하는가와 같은 구체적인 차이는 알아내기 어렵다. 정규분포에서 $error_2$와 $error_3$는 분포 그래프의 위치와 모양이 얼마나 변하는 지를 보여준다. 즉, 이 두 오차는 원본 분포와 재구축 분포의 구체적인 차이를 나타내는 파라미터가 된다. 위의 오차에 관한 실험은 균등분포에서도 비슷한 결과를 가진다.

5.3 프라이버시의 측정

정확도 측정에서 살펴본 바와 같이 오차는 γ가 증가할수록 전체적으로 감소한다. 즉, 정확도가 증가하는 것이다. 하지만 γ가 증가할수록 프라이버시 보호 정도는 감소하게 되는데, 이것은 정리 2.1에 나타난 수식과 동치인 다음 부등식을 살펴보면 알 수 있다.

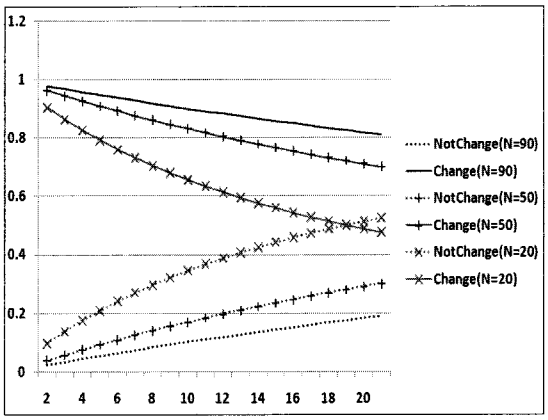
$$\rho_2 > \frac{\gamma\rho_1}{1-\rho_1+\gamma\rho_1}$$

이 부등식에서 원본 데이터의 사전확률인 ρ_1과 γ가 고정된 값이라면 랜덤 대치 알고리즘이 가지는 사후확률 ρ_2는 우변의 값보다 큰 값이 된다. [표 2]는 ρ_1과 γ에 대한 ρ_2 값의 하계(lower bound)를 나타낸 것이다.

[표 2]에서 보는 바와 같이 원본 데이터의 사전확률

〔표 2〕 ρ_1과 γ에 대한 ρ_2의 하계

γ \ ρ_1	5%	10%	15%
2	0.095238	0.181818	0.260870
6	0.240000	0.400000	0.514286
11	0.366667	0.550000	0.660000
15	0.441176	0.625000	0.725806
19	0.500000	0.678571	0.770270
24	0.558140	0.727273	0.808989



〔그림 9〕 N과 γ에 대한 속성 값의 변화 비율

이 5%라면 사후확률이 50% 이상이 되도록 하기위해서는 γ를 19로 선택해야 한다. γ가 커진다면 오차는 줄어들겠지만 사후확률이 증가하게 되어 프라이버시 보호 정도가 낮아지게 된다. 한편, 사전확률이 10%라면 사후확률이 50% 이상 되게 하기위해서 γ를 9 정도로 선택하면 되기 때문에 사전확률이 5%일 때보다 정확도는 상대적으로 낮아진다. 그러나 이 경우 프라이버시 손상 정도는 상대적으로 작기 때문에 프라이버시의 보호 정도는 사전확률이 5%일 때보다 10%일 때가 더 높다고 할 수 있다.

한편, 프라이버시 보호 정도는 γ와 N의 변화에 관련된 오차 측정방법과 사전확률 및 사후확률 관점의 분석 외에 실험적인 관점에서도 측정해볼 수 있다. 각 원본 데이터의 속성 값이 변형될 때, 자기 자신이 아닌 다른 속성 값으로 변하는 데이터의 개수가 많다면 프라이버시 보호 정도가 높다고 볼 수 있다. 실제로 우리는 시뮬레이션을 통하여 이러한 관점의 분석을 실시하였다. [그림 9]는 γ와 N의 변화에 따라 자기 자신이 아닌 다른 속성 값으로 변화된 데이터의 비율을 표현한 것이다. 이 그림에서 보면 γ가 증가할수록 속성 값이 바뀐 비율이 점점 감소하고 그에 따라 바뀌지 않은 비율이 점점 증가함을 알 수 있다.

Ⅵ. 결 론

우리는 본 논문에서 행렬 기반 랜덤화 기법에 대하여 깊이 있는 분석을 실시하였다. 최적의 변환 행렬을 찾기 위한 행렬의 조건수 개념을 새롭게 적용하여 FRAPP 관련 파라미터들 간의 관계를 세밀하게 규명하였다. 행

렬 기반의 대표적 알고리즘인 랜덤 대치 기법의 효율적인 구현을 위하여 데이터 재구축 과정에서 필요한 역행렬을 구하기 위한 단순한 공식을 제시하였다. 이 공식은 변환 행렬의 특수성에 기인한 것으로 가우스 소거법에 비해 매우 효율적이며 행렬의 크기에 영향을 받지 않는다는 장점을 지닌다.

프라이버시와 정확도에 대한 취사선택 기준을 확실하게 제시하기 위하여 랜덤 대치에 사용되는 파라미터들과 정확도를 측정하기 위한 오차 한계에 대한 관계식을 이론적으로 도출해내었다. 이 사실을 밝혀내기 위하여 프라이버시 손상(breach) 관점의 파라미터 γ의 의미를 정확히 분석하고, 변환 행렬의 조건수를 행렬 노름에 따라 계산하였으며, 변환된 데이터의 기댓값과 분산을 이론적으로 계산하였다. 그 결과 파라미터 γ, 데이터 속성의 종류를 나타내는 파라미터 N, 그리고 데이터의 개수를 의미하는 파라미터 S가 유기적인 연관성을 가지고 오차 한계를 결정하게 된다는 사실을 입증하였다.

더욱이 위와 같이 이론적으로 얻은 연구 결과의 실증적 분석을 위하여 랜덤 대치 기법을 구현하고 다양한 시뮬레이션을 실시하였다. 정규 분포와 균등 분포를 따르는 데이터에 대한 구현 결과로부터 변형된 데이터는 모두 균등 분포에 가까운 분포를 보인다는 것을 관찰하여, 균등 분포를 따르는 경우에는 재구축 과정이 필요하지 않다는 사실도 밝혔다. 이론적으로 구한 오차의 한계는 파라미터들인 γ, N, S의 다양한 값들에 대한 시뮬레이션에서도 유사한 결과를 관찰할 수 있음을 발견할 수 있었다. 그리고 프라이버시 보호 정도는 사전확률 및 사후확률과 파라미터 γ의 관계식을 통한 분석과 랜덤 대치 알고리즘에서 실제로 변화된 데이터의 비율을 실험적으로 조사한 분석을 통하여 실증적으로 규명하였다.

랜덤 대치 기법은 원본 데이터 분포를 재구축한다는 측면에서 연관규칙 마이닝, 결정나무 마이닝, 군집화, 아웃라이어 판별 등 여러 가지 형태의 데이터 마이닝 기법을 위한 프라이버시 보호 기술로 활용될 수 있다는 장점이 있다. 하지만 각각의 응용 기술에 랜덤 대치 기법을 적용할 때에는 각 응용 환경에 적합한 효율적인 구현 방법과 프라이버시 측정 방법들에 대한 연구가 좀 더 면밀히 수행되어야 한다.

참고문헌

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

[2] Keke Chen, and Ling Liu, "Privacy-Preserving Data Classification with Rotation Perturbation", *Proc. of IEEE Intl. Conf. on Data Mining (ICDM05)*, 2005.

[3] Kun Liu, Hillol Kargupta, and Jessica Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", *IEEE Transactions on Knowledge and Data Engineering archive*, Vol. 18, Issue 1, 2006.

[4] S. Agrawal, and J. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining", *Proc. of ICDE 2005*, 2005.

[5] Jim Dowd, Shouhuai Xu, and Weining Zhang, "Privacy-Preserving Decision Tree Mining Based on Random Substitutions", *ETRICS2006, LNCS 3995*, Springer-Verlag, pp. 145-159, 2006.

[6] S. Agrawal, and J. Haritsa, "A framework for high-accuracy privacy-preserving mining", Technical Report TR-2004-02, *Database Systems Lab, Indian Institute of Science*, 2004. (http://dsl.serc.iisc.ernet.in/pub/TR/R-2004-02.pdf).

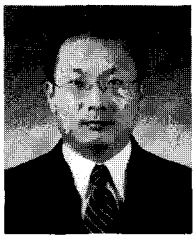
[7] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining", *Proc. of ACM Symp. on Principles of Database Systems (PODS)*, 2003.

[8] 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.

[9] C. D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM(Society for Industrial and Applied Mathematics), Philadelphia, 2000.

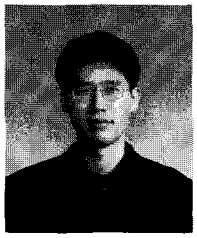
[10] Y. Wang, "On the number of success in independent trials", *Statistica Silica* 3, 1993.

〈著者紹介〉



강 주 성 (Ju-Sung Kang) 종신회원
1989년 : 고려대학교 수학과(학사)
1991년 : 고려대학교 일반대학원 수학과 (이학석사)
1996년 : 고려대학교 일반대학원 수학과 (이학박사)
1996년~1997년 : 과학재단 박사후연구원
1997년~2004년 : 한국전자통신연구원 선임연구원, 팀장
2001년~2002년 : 벨기에 루벤대학 COSIC 방문연구원
2004년~현재 : 국민대학교 수학과 부교수
<관심분야> 암호 알고리즘, 정보보호 프로토콜

안 아 론 (Aron An)
2006년 : 국민대학교 수학과 졸업
2008년 : 국민대학교 일반대학원 수학과 (이학석사)
2008년~현재 : (주)누리솔루션
<관심분야> 정보보호, 암호론



홍 도 원 (Dowon Hong) 정회원
1994년 : 고려대학교 수학과(학사)
1996년 : 고려대학교 일반대학원 수학과 (이학석사)
2000년 : 고려대학교 일반대학원 수학과 (이학박사)
2000년~현재 : 한국전자통신연구원 선임연구원, 팀장
<관심분야> 암호 이론, 정보보호 이론, 이동통신 정보보호