DOI: 10.3745/KIPSTC.2009.16C.5.563

랜덤대치 기반 프라이버시 보호 기법의 정확성 개선 알고리즘

강 주 성[†] · 이 창 우[††] · 홍 도 원[†††]

요 약

랜덤대치 기법은 실용적인 프라이버시 보호 방법으로 다양한 응용 가능성과 프라이버시 손상 관점의 안전성을 보장할 수 있다는 장점이 있다. 하지만 데이터 유용성을 위한 랜덤대치 기법의 정확성을 향상시키는 방법에 대해서는 그동안 면밀히 연구되지 않았다. 본 논문에서는 랜덤대치 기법의 표준오차에 대한 보다 진전된 이론적 분석을 실시함으로써 정확성을 개선할 수 있는 알고리즘을 제안한다. 다양한 실험을 통하여 균등분포와 정규분포를 따르는 원본 데이터에 대한 랜덤대치 기법의 적용이 실용적이지 못한 정확성을 나타낸다는 사실과 함께 개선된 알고리즘의 정확성 향상 정도를 확인한다. 우리가 제안하는 알고리즘은 기존의 랜덤대치 기법과 동일한 프라이버시 수준을 유지한 상태에서 정확성을 원하는 수준만큼 높일 수 있는 방법이며, 이를 위해 추가로 소요되는 계산량은 실용적인 면에서 여전히 수용 가능한 것임을 밝힌다.

키워드 : 랜덤화, 랜덤대치, 프라이버시, 정확성, 프라이버시 보존형 데이터마이닝

An Algorithm for Improving the Accuracy of Privacy-Preserving Technique Based on Random Substitutions

Ju-Sung Kang[†] · Chang-Woo Lee[††] · Dowon Hong[†††]

ABSTRACT

The merits of random substitutions are various applicability and security guarantee on the view point of privacy breach. However there is no research to improve the accuracy of random substitutions. In this paper we propose an algorithm for improving the accuracy of random substitutions by an advanced theoretical analysis about the standard errors. We examine that random substitutions have an unpractical accuracy level and our improved algorithm meets the theoretical results by some experiments for data sets having uniform and normal distributions. By our proposed algorithm, it is possible to upgrade the accuracy level under the same security level as the original method. The additional cost of computation for our algorithm is still acceptable and practical.

Keywords : Randomization, Random Substitutions, Privacy, Accuracy, Privacy Preserving Data Mining

1. 서 론

실용적인 프라이버시 보호 기술의 대표적인 응용 분야인 프라이버시 보존형 데이터 마이닝에서는 정보제공자의 비밀 데이터를 보호하기 위해서 변형된 데이터를 마이너에게 제공한다. 데이터의 변형은 프라이버시 관련 정보를 노출시키지 않기 위함이며, 데이터 변형의 가장 실용적인 방법이 랜덤화(randomization) 기법이다. 최근에 발표된 랜덤대치(random substitutions)는 랜덤화 기법 중의 하나로 안전성과 효율성이 높고, 다양한 분야에 응용 가능한 방법으로 알려져 있다. 하지만 데이터 유용성을 위한 랜덤대치 기법의 정확성을 높이는 문제는 지금까지 면밀히 연구되지 않았다. 본 논문에서는 랜덤대치 기법의 정확성에 대하여 심도 있는 분석을 실시하고, 다양한 실험을 통하여 랜덤대치 기법에 의한 균등분포와 정규분포를 따르는 데이터 변형 과정이 재구축 과정에서 비실용적인 정확성을 가진다는 점을 밝힌다. 또한, 랜덤대치 기법의 표준오차에 대한 이론적 분석을 진전시킴으로써 이에 기반한 정확성 개선 알고리즘을 제안하고 이를 실험적으로 확인한다. 우리가 제안하는 알고리즘은 원래의 랜덤대치 기법과 동일한 프라이버시 수준을 만족한 상태에서 원하는 수준만큼 정확성을 향상시킬 수 있는 방법이다. 물론 추가적인 계산량이 소요되지만 실용적인 면에서 이 계산량은 여전히 수용 가능한 것임을 밝힌다.

※ 본 연구는 국민대학교 교내연구비 지원과 지식경제부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음[2005-Y001-04, 차세대 시큐리티 기술 개발].
† 정 회 원 : 국민대학교 수학과 부교수
†† 정 회 원 : 국민대학교 수학과 이학석사
††† 정 회 원 : 한국전자통신연구원 지식정보보안연구부 선임연구원
논문접수 : 2009년 3월 27일
수 정 일 : 2009년 7월 6일
심사완료 : 2009년 7월 6일

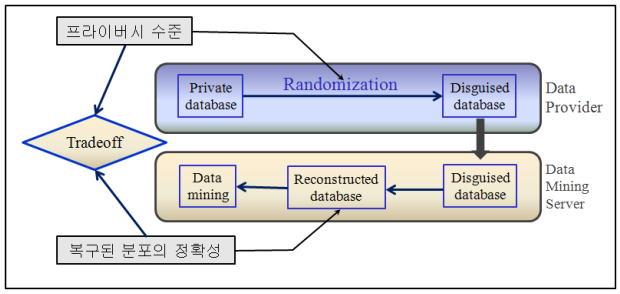
2. 프라이버시 보존형 데이터 마이닝

인터넷과 전자상거래가 급속도로 보급되면서 대용량의 데이터를 자동으로 수집하는 것이 용이하게 되었다. 컴퓨터 및 네트워크 기술의 발달로 과거에는 가능하지 않았던 거대한 양의 데이터를 우리 주변에서 쉽게 볼 수 있는 시대가 된 것이다. 하지만 이렇게 모아 놓은 데이터로부터 유용한 정보를 찾아내어 마케팅이나 회사의 이익을 효율적으로 증대시키기 위해서 사용하는 데는 아직 많은 어려움이 남아 있다. 데이터 마이닝(data mining) 기술은 이러한 대용량의 데이터로부터 유용하고 값진 정보를 효율적으로 찾아내어 회사뿐만 아니라 개인의 일상생활에도 편리하게 도움을 줄 수 있다. 한마디로 표현하면 데이터 마이닝은 대용량의 데이터에 함축적으로 들어있는 지식이나 패턴을 발견해내는 기술이다. 현재 상용 데이터 마이닝 소프트웨어에서 제공되는 중요 알고리즘에는 연관 규칙(association rules), 분류(classification), 순차 패턴(sequential patterns), 군집화(clustering), 아웃라이어 판별(outlier discovery) 등이 있다.

한편, 데이터 마이닝 기술의 유용성 이면에는 프라이버시 문제가 존재한다. 데이터를 모으고 이를 여러 가지 방법으로 분석하는 과정에서 프라이버시와 관련된 문제가 자연스럽게 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. 이와 같은 취사선택 문제를 해결하기 위한 기술적 관점의 노력들을 학계에서는 프라이버시 보존형 데이터 마이닝(PPDM, Privacy-Preserving Data Mining)[1-3]이라 부른다. PPDM 기술에서 중요한 사항은 참여자의 프라이버시 보호 수준과 공유된 유용한 정보의 품질을 결정하는 정확성(accuracy)을 논리적으로 취사선택하는 방법이라 할 수 있다. 실용적 측면에서의 PPDM 기술의 유용성은 계산 효율성이 좌우하므로 프라이버시 수준 및 정확성을 높이는 것과 함께 적용하려는 컴퓨팅 환경에 대한 적합성도 중요한 요소가 된다.

2.1 랜덤화 기반 PPDM

PPDM 관련 연구는 크게 두 가지로 대별된다. 먼저 프라이버시를 보존하는 통계적 데이터베이스로 종종 언급되는 것으로 데이터를 마이닝 대상으로 사용되기 이전에 변환하는 기술이다. 원래의 데이터에 노이즈를 첨가하거나 다른 종류의 랜덤화(randomization)를 적용시키는 것이 이 방법의 예이다. 랜덤화에 의한 프라이버시 보호 방법은 암호적 방법에 비해 안전성이 떨어짐에도 불구하고 매우 효율적이란 장점 때문에 실용적으로 널리 사용되고 있다. 하지만 오리지널 데이터의 변형에 기인한 데이터 마이닝 결과의 정확성(accuracy)은 해결되어야할 중요한 문제로 남는다. 랜덤화를 이용한 PPDM의 흐름도는 (그림 1)에 나타나 있다.



(그림 1) 랜덤화를 이용한 PPDM 모델

2.2 SMC 기반 PPDM

PPDM에서 프라이버시를 보호하는 두 번째 방법은 데이터 마이닝에 암호 원천 기술이라 할 수 있는 SMC(secure multiparty computation)[4] 기술이 적용된 것이다. 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 랜덤화 기법에서 발생할 수 있는 정확성(accuracy) 문제는 발생하지 않는다. 그러나 SMC 기반 PPDM 기술은 계산 효율성이 매우 낮기 때문에 아직까지 실용적이지 못하다는 한계를 지닌다. 여기에서는 아직까지는 다분히 이론적인 단계에 머물고 있는 SMC 기반 프라이버시 보호 기술에 대해서는 깊이 있게 다루지 않기로 한다.

3. 랜덤대치 기법

우리가 다루고자 하는 랜덤대치 기법은 행렬 기반 랜덤화의 일종으로 볼 수 있다. 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[5-7]의 FRAPP(Framework for High-Accuracy Privacy-Preserving Mining) 개념을 기술한다. 이를 위하여 먼저 프라이버시 보호의 정도를 측정하는 측도로 널리 사용되고 있는 프라이버시 손상(privacy breach)에 대하여 살펴본다.

3.1 프라이버시 손상

프라이버시 손상(privacy breach) 개념은 기존의 상호정보(mutual information) 개념을 이용한 프라이버시 측도의 대안으로 Evfimievski-Gehrke-Srikant[8]가 제안한 것이다. Agrawal-Agrawal[9]은 신뢰구간의 길이로 프라이버시를 측정하는 기존 방법의 불합리성을 지적하면서 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.

일반적으로 프라이버시 손상은 어떤 성질 Q에 대하여 변형되어 공개된 정보 y_i가 이 성질의 노출 확률을 눈에 띠게 증가시키는 상황을 의미한다. Evfimievski-Gehrke-Srikant[8]에 나타나 있는 프라이버시 손상의 엄밀한 정의는 다음과 같다.

[정의 3.1] X와 Y는 확률변수이고, V_X와 V_Y는 각각 X와 Y가 취할 수 있는 값의 영역이라 하자. 이 때, 성질 Q에 대하여 상향식 $\rho_1 \rightarrow \rho_2$ 프라이버시 손상이 발생한다는 의미는 $\Pr[Q(X)] \leq \rho_1$ 이고 $\Pr[Q(X) \mid Y=y] \geq \rho_2$를 만족하는 어떤 $y \in V_Y$가 존재한다는 것이다. 반대로 하향식 $\rho_2 \rightarrow \rho_1$ 프라이버시 손상이 발생한다는 의미는 $\Pr[Q(X)] \geq \rho_2$ 이고 $\Pr[Q(X) \mid Y=y] \leq \rho_1$을 만족하는 어떤 $y \in V_Y$가 존재한다는 것이다. 여기에서 $0 < \rho_1 < \rho_2 < 1$이고 $\Pr[Y=y] > 0$이다.

정의 3.1에서 기술하고 있는 프라이버시 손상의 직관적 의미는 다음과 같다. 성질 $Q(X)$는 랜덤화 되기 이전의 개인정보 분포를 의미하는 확률변수 X에 관한 사건(event)이며, 확률변수 Y는 데이터 마이닝 서버에게 공개된 마이닝 수행 직전의 데이터 분포를 의미한다. 그러므로 상향식 $\rho_1 \rightarrow \rho_2$ 프라이버시 손상이 발생했다는 의미는 비밀 상태의 원본 데이터에서 ρ_1이하의 확률이었던 사건이 마이닝 서버 입장에서 복구된 데이터 분포로부터 조건부 확률을 구할 경우 ρ_2이상의 확률을 갖는 사건으로 변한다는 것이다. 반대로 하향식 $\rho_2 \rightarrow \rho_1$ 프라이버시 손상이 발생한다는 것은 비밀 상태의 원본 데이터에서 ρ_2이상의 확률이었던 사건이 마이닝 서버 입장에서 복구된 데이터 분포로부터 조건부 확률을 구할 경우 ρ_1이하의 확률을 갖는 사건으로 변한다는 의미이다. 그러므로 비밀인 원본 데이터 분포로부터 구한 원확률과 서버에게 주어진 데이터 분포로부터 구하는 조건부확률의 차이가 클수록 프라이버시 손상이 많이 발생한다고 직관적으로 설명할 수 있다.

다음으로 임의의 성질에 대하여 정보를 노출하지 않는 랜덤화 작용소에 대한 조건을 정의한다. 실용적으로 유용한 랜덤화 작용소는 임의의 $y \in V_Y$에 대해서 서로 다른 $x \in V_X$에 대한 작용소의 변환확률 $p[x \rightarrow y]$를 비교해봄으로써 정보 노출의 정도를 측정할 수 있다. 모든 x 값이 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 개념으로 측정한다.

[정의 3.2] 랜덤화 작용소 $R(x)$가 $y \in V_Y$에 대하여 기껏해야 γ-증폭이라는 의미는

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

가 성립한다는 것이다. 여기에서 $\gamma \geq 1$이고, $p[x_i \rightarrow y] > 0$이다. 그리고 모든 $y \in V_Y$에 대해서 기껏해야 γ-증폭일 경우에 작용소 $R(x)$는 기껏해야 γ-증폭이라고 말한다.

랜덤화 작용소 $R(x)$가 기껏해야 γ-증폭이라는 직관적 의미는 임의의 x_1과 x_2에 대하여 각각의 변환확률이 $p(x_1 \rightarrow y) \leq \gamma\, p(x_2 \rightarrow y)$를 만족한다는 것이므로 랜덤화 과정에서의 모든 변환확률의 차이는 기껏해야 γ 배를 넘지 않는다는 것이다.

[정리 3.1] (Evfimievski-Gehrke-Srikant[8]) R을 랜덤화 작용소, $y \in V_Y$를 $\exists x, p[x \rightarrow y] > 0$을 만족하는 랜덤화된 값이라 하고 정의 3.1로부터의 $0 < \rho_1 < \rho_2 < 1$을 두 확률값이라 하자. R은 y에 대해서 기껏해야 γ-증폭이라고 가정하자. 그러면 "$R(x)=y$"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 3.3] 랜덤화 작용소 R이 정리 3.1의 조건을 만족할 경우에 R은 (ρ_1, ρ_2) 프라이버시 보증을 지지한다고 말한다.

3.2 랜덤대치 기법의 수행 과정

여기에서는 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 벡터값인 경우로 확장함으로써 원하는 결과를 얻을 수 있다.

3.2.1 원본 데이터의 변형 과정

Agrawal-Harista[5]에서 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[10]은 랜덤대치 기법(random substitutions)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 변환행렬을 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 분포함수를 이용함으로써 속성 값을 변형할 수 있다. 랜덤 대치 기법으로 데이터를 변형하는 방법을 알고리즘으로 표현하면 다음과 같다.

알고리즘 1. 랜덤대치 기법의 데이터 변형 방법

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

출력 : 변환된 데이터 집합 P

수행 과정 :

모든 레코드 $o\in O$에 대해 다음을 실행한다.

1. o가 가지는 속성 값의 인덱스 값 k를 구한다. 즉, o가 가지는 속성 값은 u_k이다.
2. (0, 1] 상의 균등분포로부터 랜덤수 r을 선택한다.
3. 다음을 만족하는 정수 $1 \leqq h \leqq N$ 를 찾는다

$$\sum_{i=1}^{h-1} m_{ik} < r \leqq \sum_{i=1}^{h} m_{ik}$$

4. o에 대응되는 변환된 레코드의 속성 값을u_h로 결정한다.

랜덤 대치 데이터 변형 알고리즘의 계산복잡도는 $O(n\cdot N)$이다. 정리 3.1의 프라이버시 손상 관점의 안전성 요구조건을 만족하고, 최소 조건수 관점의 정확도 요구조건 하에서, 정의 3.2의 γ를 사용한 최적의 변환행렬은 $M=xG$의 형태를 가지는 γ-대각 행렬이 된다는 사실이 알려져 있다[10]. 이 때, x와 G는 다음과 같이 주어진다.

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

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

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

을 만족하고, 정의 3.3의 (ρ_1,ρ_2) 프라이버시 보증을 지지하기 위하여 다음 조건

$$\forall\, 1\leq h,k_1,k_2\leq N,\ \frac{m_{hk_1}}{m_{hk_2}}\leq\gamma\leq\frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)}$$

을 만족한다. 정확도 관점에서는 행렬의 조건수를 고려할 경우에 변환행렬 $M=(m_{hk})_{N\times N}$ 의 조건수는 $1+N/(\gamma-1)$으로 계산되고, 대칭 변환행렬의 경우에는 이 값이 최소값이 된다는 사실이 밝혀져 있다[6]. 그러므로 조건수를 측도로 사용하는 정확도 관점에서는 적절한 제한 조건 아래에서 수식 (1)을 만족하는 행렬 $M=(m_{hk})_{N\times N}$ 이 최적의 선택 중 하나라는 사실을 알 수 있다.

3.2.2 원본 데이터 분포의 재구축 과정

분포 재구축 과정은 변형된 데이터 집합으로부터 원본 데이터의 분포를 추정하는 과정이다[5,6]. 데이터 분포의 추정은 변형된 데이터 집합 P와 변환행렬 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 도수를 표현하는 열벡터이다. 여기에서 원본 데이터와 변형 데이터의 전체 도수는

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

라 놓는다. 그러면 주어진 O에 대해서 Y에 대한 기댓값은 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 가역행렬인 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 $\hat{X}$의 추정값 $\hat{\boldsymbol{x}}$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 비편향 추정량이 된다. 그러므로 $\hat{X}$ 을 재구축된 데이터 분포로 놓으면, 통계적으로 원본 데이터 분포와 유사한 분포를 얻게 되는 것이다.

랜덤대치 기법의 효율적인 구현을 위하여 데이터 재구축 과정에서 필요로 하는 역행렬을 구하는 공식은 [11]의 정리 2.2에서와 같이 간단히 얻을 수 있고, 다음 정리 3.2에서와 같이 쉽게 $\hat{X}$ 를 재구축할 수 있다.

[정리 3.2] $\gamma > 1$, $N > 1$인 γ-대각 행렬 M에 대해서, 전체 $n = \sum_{i=1}^{N} Y_i$ 개의 레코드에 대한 $\hat{X_i}$는 다음과 같다.

$$\hat{X_i} = \frac{\gamma + N - 1}{\gamma - 1} Y_i - \frac{n}{\gamma - 1}.$$

[증명] $\hat{X} = (\hat{X_1}, \cdots, \hat{X_N})^T = M^{-1}\boldsymbol{y}$ 이므로

$$\begin{aligned}\hat{X_i} &= m_{i1}^{-1}Y_1 + m_{i2}^{-1}Y_2 + \cdots + m_{iN}^{-1}Y_N \\ &= \frac{\gamma + N - 2}{\gamma - 1} Y_i + \sum_{j \neq i} \frac{1}{1 - \gamma} Y_j \\ &= \frac{\gamma + N - 1}{\gamma - 1} Y_i - \sum_{j=1}^{N} \frac{1}{\gamma - 1} Y_j \\ &= \frac{\gamma + N - 1}{\gamma - 1} Y_i - \frac{n}{\gamma - 1}.\end{aligned}$$

4. 랜덤대치 기법의 정확성 분석

본 장에서는 랜덤대치 기법의 정확성을 이론적으로 분석하고, 그에 따른 실험을 통하여 랜덤대치 기법이 정확성 관점에서 취약함을 밝히고자 한다.

4.1 이론적 측면의 정확성 분석

변환 행렬로 γ-대각 행렬을 사용하는 랜덤 대치 기법의 정확성은 3장에서 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. 따라서 랜덤대치의 표준오차를 $\sigma_{\hat{X}} / \|X\|$ 로 정의한다. 여기에서 $\hat{X}$ 의 표준편차는 $\sigma_{\hat{X}} = \sqrt{Var(\hat{X})}$ $= \sqrt{E[\|\hat{X} - X\|^2]}$ 으로 주어지고, $\|\cdot\|$ 는 벡터의 유클리드 노름을 의미한다. 표준오차의 이론적인 상계를 계산하기 위하여 필요로 하는 $Var(Y)$는 [11]에 나타난 바와 같이 계산된다. 즉,

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

이다. 이제 표준오차의 이론적인 상계를 $Var(Y)$와 관련하여 다음과 같이 계산할 수 있다.

[정리 4.1] $X = (X_1, \cdots, X_N)^T$ 와 $\hat{X} = (\hat{X}_1, \cdots, \hat{X}_N)^T$ 는 각각 원본과 재구축된 데이터 집합에서 속성의 도수 분포를 표현하는 열벡터이고, $n = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i$ 일 때, $\hat{X}$의 표준오차 $\frac{\sigma_{\hat{X}}}{\|X\|}$는 변환 행렬 $M = (m_{hk})_{N \times N}$ 이 수식 (1)과 같은 γ-대각 행렬로 주어질 경우 다음과 같이 계산된다.

$$\frac{\sigma_{\hat{X}}}{\|X\|} \leq \frac{\sqrt{N(N-1)(N+2\gamma-2)}}{(\gamma-1)n^{1/2}}.$$

[증명] 알고리즘 1에서 랜덤수를 의미하는 파라미터 r를 원본 데이터 원소들마다 독립적으로 추출하기 때문에 확률변수 $\hat{X}_i\,(1 \leq i \leq n)$ 들은 서로 독립이므로,

$Var(\hat{X}) = \sum_{i=1}^{N} Var(\hat{X_i})$ 이다. [정리 3.2]로부터

$\hat{X_i} = \frac{\gamma + N - 1}{\gamma - 1} Y_i - \frac{n}{\gamma - 1}$ 이므로, $Var(\hat{X})$는 다음과 같이 계산할 수 있다.

$$\begin{aligned}Var(\hat{X}) &= \sum_{i=1}^{N} Var\left(\frac{\gamma+N-1}{\gamma-1} Y_i - \frac{n}{\gamma-1}\right) \\ &= \left(\frac{\gamma+N-1}{\gamma-1}\right)^2 \sum_{i=1}^{N} Var(Y_i) \\ &= \frac{(N-1)(N+2\gamma-2)n}{(\gamma-1)^2}\end{aligned}$$

$\|X\|$를 계산하기 위해, X_i와 관련하여 Cauchy-Schwarz 부등식을 이용하면,

$$\left(\sum_{i=1}^{N}(X_i \cdot 1)\right)^2 \leq \sum_{i=1}^{N} X_i^2 \cdot \sum_{i=1}^{N} 1^2$$

임을 만족한다. $\sum_{i=1}^{N} 1^2 = N$, $n^2 = \left(\sum_{i=1}^{N}(X_i \cdot 1)\right)^2$ 이므로,

$\|X\|$는 다음과 같다.

$$\|X\| = \sqrt{\sum_{i=1}^{N} X_i^2} \geq \frac{n}{\sqrt{N}}.$$

결과적으로 $\frac{\sigma_{\hat{X}}}{\|X\|}$는

$$\frac{\sigma_{\hat{X}}}{\|X\|} \leq \frac{\sqrt{(N-1)(N+2\gamma-2)n}}{\gamma-1} / \frac{n}{\sqrt{N}}$$

$$= \frac{\sqrt{N(N-1)(N+2\gamma-2)}}{(\gamma-1)n^{1/2}}$$

와 같이 계산된다.

랜덤화의 정확성을 결정하는 주된 요소는 추정량 $\hat{X}$의 표준편차 또는 분산이란 것은 주지의 사실이다. 하지만 정리 4.1에서 우리가 주목할 사항은 표준편차를 결정하는 파라미터들의 차수(order)이다. 즉, $\hat{X}$의 표준편차가

$$\sigma_{\hat{X}} = \sqrt{(N-1)(N+2\gamma-2)n} / (\gamma-1)$$

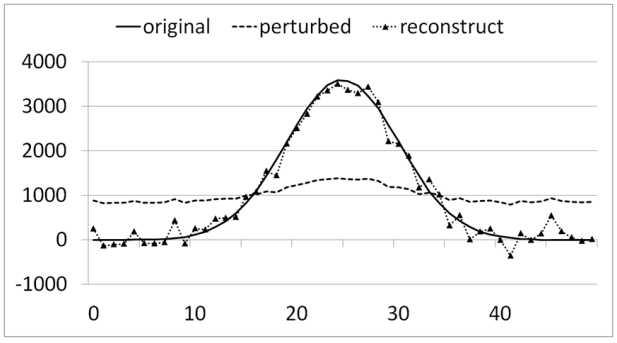
로 계산되기 때문에 $\hat{X}$의 표준편차는 근사적으로 N과 $\sqrt{n}$에 비례하고, $\sqrt{\gamma}$에 반비례함을 알 수 있다. Dowd-Xu-Zhang[10]는 연속형 자료를 이산형 자료로 변환(discretize)하여 랜덤대치 기법을 적용함으로써 프라이버시 보존형 의사결정나무 분석에 적용하였다. 참고문헌 [10]에서 저자들은 데이터의 차원을 의미하는 파라미터 N을 적절히 선택하여 이산화를 수행함으로써 정확성과 프라이버시 수준 및 계산 효율성을 취사선택할 수 있는 방안을 제시하였다. 하지만 프라이버시 보존형 연관규칙 마이닝 등에 랜덤대치 기법을 적용할 경우 파라미터 N은 데이터 속성의 개수를 의미함으로 고정된 경우가 대부분이다. 그러므로 파라미터 N을 조정함으로써 정확성을 확보하고자 하는 방법은 제한적일 수 밖에 없다. 정리 4.1에서 우리가 얻은 결과는 데이터의 총량을 의미하는 파라미터 n이 정확성에 영향을 미친다는 것을 단적으로 보여준다. 파라미터 γ는 프라이버시 수준을 먼저 고려할 경우 최적값으로 고정시킬 수 있으므로 n을 조정함으로써 원하는 수준의 정확성을 만족시키는 방법은 대단히 유용한 것이다.

한편, n을 증가시키면 $\sqrt{n}$에 비례하여 표준편차는 증가하여 단순히 정확성이 떨어질 것이라 예상할 수 있지만, 데이터 총량을 고려한 상대적 정확성 측도인 표준오차 $\frac{\sigma_{\hat{X}}}{\|X\|}$를 고려할 경우 n에 따른 표준편차의 증가 비율이 더욱 중요한 요소가 된다. 직관적으로 $\|X\|$는 n에 비례하여 증가하므로 표준오차는 전체적으로 $\sqrt{n}$에 반비례할 것이란 것을 알 수 있으며, 이를 이론적으로 분석해낸 결과가 정리 4.1이다. 주어진 데이터에서 총량 n을 l배 증가시킨다는 의미는 랜덤대치 알고리즘을 수행할 때 동일한 데이터 레코드에 대하여 l번의 랜덤대치를 적용한다는 것이다. 이러한 기본적인 생각을 적용하여 새롭게 제안한 정확성 향상 알고리즘의 세부 내용은 5장에서 다룬다.

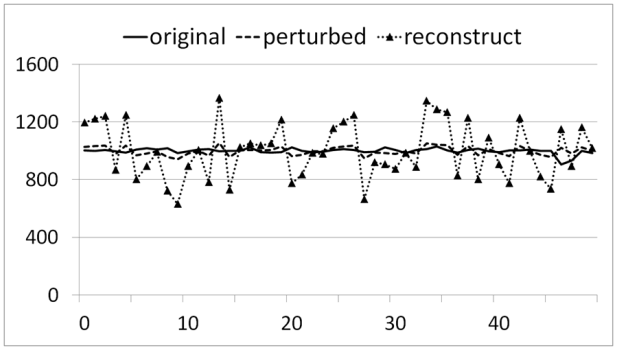
4.2 실험적 측면에서의 정확성 분석

실험을 통해서 정확성에 영향을 미치는 세 개의 파라미터 (n, N, γ) 중에서 n의 감소와 N의 증가에 따른 정확성의 변화를 측정하고자 한다. 실험 후의 실제 표준오차는 10회 평균값으로 측정 할 것이며, 실험 결과를 판단하기 위한 두 개의 대조군을 선택하였다. 하나는 실제 표준오차가 0.1185인 정규분포를 따르는 데이터 집합이고, 다른 하나는 실제 표준오차가 0.1929인 균등분포를 따르는 데이터 집합으로 선택하였다. 각 대조군에 사용된 파라미터는 $N=50$, $n=50{,}000$, $\gamma=10$으로 동일하고, 이론적인 표준오차의 상계는 0.2028이다. 정규분포를 따르는 데이터에 대해 랜덤대치를 수행한 결과는 (그림 2)와 같고, 균등분포를 따르는 데이터에 대해 랜덤대치를 수행한 결과는 (그림 3)과 같다.

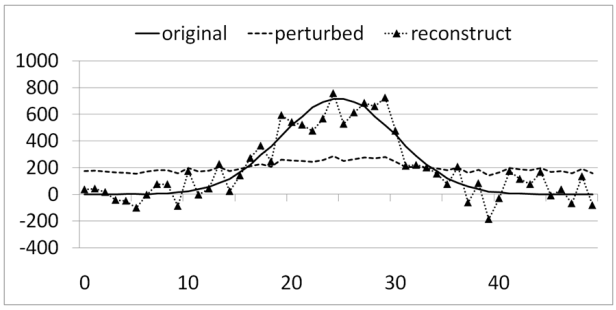
여기서 n의 값을 10,000으로 감소시키고 나머지 조건을 동일하게 유지하여 이론적인 표준오차의 상계가 0.4535인 파라미터에 대한 정규분포 데이터 집합의 실험 결과는 (그림 4)와 같고, 균등분포 데이터 집합의 실험 결과는 (그림 5)와 같다.



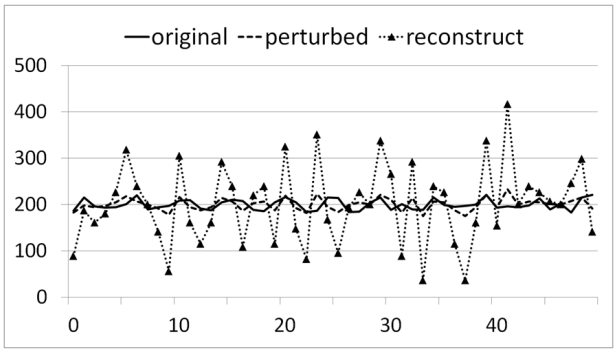
(그림 2) $N=50$, $n=50{,}000$일 때, 속성값 분포(정규)



(그림 3) $N=50$, $n=50{,}000$일 때, 속성값 분포(균등)



(그림 4) $N=50$, $n=10{,}000$일 때, 속성값 분포(정규)



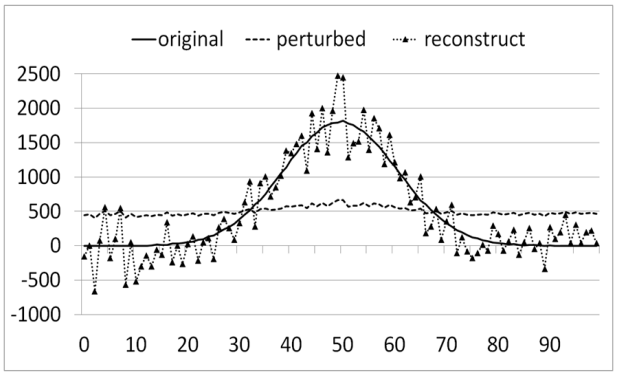
(그림 5) $N=50$, $n=10{,}000$일 때, 속성값 분포(균등)

(그림 2)와 (그림 4), (그림 3)과 (그림 5)를 각각 비교하면 두 데이터 집합 모두에서 정확성이 현저히 떨어졌음을 쉽게 확인할 수 있다. 한편, 정리 4.1의 이론적인 결과는 n의 증가 비율 $\sqrt{n}=\sqrt{5}$에 따라, (그림 4)와 (그림 5)의 실제 표준오차는 각각 $\sqrt{5}$배 만큼 증가하여 0.2650와 0.4313일 것이라고 예상할 수 있다. 실제로 각각의 표준오차는 0.2793과 0.4204로 예측 값에 근사함을 확인할 수 있다.

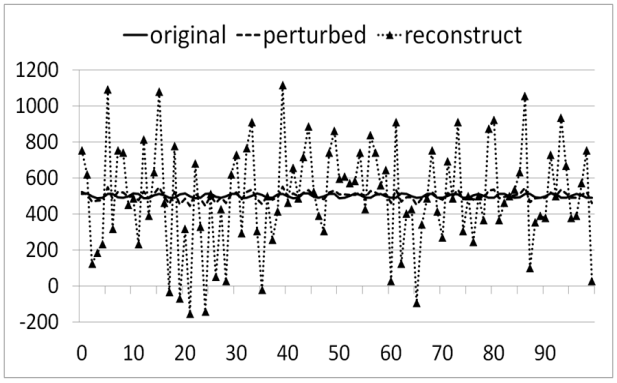
직관적으로 랜덤대치의 랜덤화는 균등한 분포로부터 r값이 랜덤하게 선택되어진다는 것에 기인한다. 하지만 동일한 50개의 속성값으로 랜덤화가 이루어짐에도 (그림 1)에서는 각 속성별 데이터 수가 대부분 1000개 이상이었지만, (그림 4)에서는 대부분 200개도 넘지 않는다. 이렇듯 전체 데이터 수의 감소로 r값이 충분히 균등하게 선택되어질 만큼 각 속성별 데이터의 수가 많지 않아서 정확성이 떨어진 것을 확인할 수 있었다.

다음으로 N의 값을 100으로 증가시키고 나머지 조건을 동일하게 유지하여 이론적인 표준오차의 상계가 0.5371인 파라미터에 대한 실험을 하였다. 정규분포 데이터 집합의 실험 결과는 (그림 6)과 같고, 실제 표준오차 0.3372을 가진다. 균등분포 데이터 집합의 실험 결과는 (그림 7)과 같고, 실제 표준오차는 0.5708과 같다.

각각 (그림 2), (그림 3)과 비교해보면 가시적으로도 정확성이 많이 떨어지는 현상을 발견할 수 있다. 정리 4.1의 이론적 결과는 N의 증가비율 $N^{3/2}=2^{3/2}$에 따라 정규분포 데이터 집합은 0.3352, 균등분포 데이터 집합은 0.5456의 표준오차가 발생할 것이라고 예상할 수 있다. 이는 실험적



(그림 6) $N=100$, $n=50{,}000$일 때, 속성값 분포(정규)



(그림 7) $N=100$, $n=50{,}000$일 때, 속성값 분포(정규)

으로 얻은 표준오차 값인 0.3372와 0.5708에 매우 가까운 수임을 확인할 수 있다.

5. 정확성 개선을 위한 l-확장 랜덤대치 기법

랜덤대치 기법의 정확성은 세 개의 파라미터 (N, γ, n)에 의해 결정된다. 우리는 N이 원본 데이터 집합의 속성 도메인의 기수(cardinality)이므로 주어진 데이터에 대해 항상 일정하게 유지되는 값으로 놓는다. 연속형 자료인 경우 이산화 정도에 따라 N이 변할 수 있으나, 연관규칙 마이닝 등에 적용할 경우의 이산형 자료는 원본 데이터의 형태를 유지하는 것이 합리적이므로 N을 고정된 수로 보는 것이 타당하다. 또한, γ는 프라이버시를 위해 일정한 값 이하로 유지해주어야 하므로 정확도를 올리기 위하여 데이터 총량인 n의 값을 증가시키는 방안을 고려한다. 하지만 주어진 데이터에 임의로 새로운 데이터를 만들어 추가를 할 수는 없으므로, 주어진 데이터를 l배 확장하여 랜덤대치를 한 후에 각 속성 값 별 분포를 l로 나누어 원본 데이터의 추정값을 구하고자한다. 이로 인해 각 속성별 데이터의 수는 l배가 되어 4장에서 지적되었던 많은 속성 개수를 가지는 적은 양의 데이터 집합에 대해서도 고른 랜덤성을 확보할 수 있게 되어 결과적으로 정확성을 높일 수 있다는 것이 우리의 직관적인 생각이다.

5.1 프라이버시를 고려한 l-확장 랜덤대치 기법

랜덤대치의 안전성은 γ-증폭과 관련한 (ρ_1, ρ_2) 프라이버시 보증에 의존한다. 그런데 l-확장 랜덤대치에서는 하나의 데이터에 대하여 l개의 랜덤화된 데이터를 생성하고자 하므로 랜덤대치와는 달리 l개의 랜덤화된 데이터에 같은 데이터가 중복되어 생성되는 경우가 발생할 수 있다. 이러한 상황을 고려하여 l-확장 랜덤대치의 γ-증폭을 다시 계산하여야 한다. l-확장 랜덤대치에서 임의의 한 원소를 l번 랜덤화하여 생성된 데이터 l개로 구성된 다중집합(multi-set)을 y^*라 하자. 다중집합 y^* 내의 l개 원소 중에서 임의의 속성 u_i가 중복되어 나타난 회수를 $s_i (0 \le s_i \le l)$라 할 때, 속성 u_i가 집합 y^*로 랜덤화 될 전이확률(transition probability)이

$$p[u_i \rightarrow y^*] = \frac{\gamma^{s_i}}{(\gamma+N-1)^l}$$

과 같이 계산된다. $\gamma \ge 1$ 이므로, 모든 u_i에 대하여 랜덤화될 확률이 가장 큰 y^*는 $s_i = l$ 일 때이고, 그 확률이 가장 작은 때는 $s_i = 1$인 y^*이다. 따라서 l-확장 랜덤화의 랜덤화된 집합 y^*에 대한 γ-증폭은 다음과 같다.

$$\forall u_1, u_2 \in U, \ \frac{p[u_1 \rightarrow y^*]}{p[u_2 \rightarrow y^*]} \le \frac{\gamma^l/(\gamma+N-1)^l}{1/(\gamma+N-1)^l} = \gamma^l.$$

따라서 [정리 3.1]에 의해 l-확장 랜덤대치는 다음을 만족하는 ρ_1^*, ρ_2^*에 의해 (ρ_1^*, ρ_2^*) 프라이버시 보증을 만족한다.

$$\frac{\rho_2^*}{\rho_1^*} \cdot \frac{1-\rho_1^*}{1-\rho_2^*} > \gamma^l .$$

이는 사전확률이 ρ_1^*이하인 어떤 성질도 l-확장 랜덤대치를 통해 사후확률이 ρ_2^*를 넘을 수 없다는 것을 의미한다. 그런데 $\gamma^l \ge \gamma$ 이므로, 동일한 사전 확률에 대해서 l-확장 랜덤대치로 인한 사후확률의 상계가 더 크거나 같다. 즉, 하나의 원본 데이터에 대한 l개의 랜덤화된 데이터에 대해 중복이 적을수록 기존의 랜덤대치와 비슷한 안전성을 보이고, 중복이 없다면 동일한 안전성을 보이게 된다. 중복된 데이터가 발생할 확률은 다음과 같이 계산할 수 있다.

[정리 5.1] s를 임의의 원본 데이터 x가 l-확장 랜덤대치를 통해, l개의 데이터 y^*로 랜덤화 될 때의 중복수라고 하면, y^*에 x와 동일한 데이터가 중복되어 존재할 확률은 $P[s \ge 2]$와 같고, 다음과 같이 계산된다.

$$P[s \ge 2] = 1 - \frac{N-l+l\gamma}{(\gamma+N-1)^l} \prod_{i=1}^{l-1} (N-i).$$

[증명] $s \ge 1$ 이므로, $P[s \ge 2] = 1 - P[s=1]$이고, $P[s=1]$은 x의 l개의 램덤화된 데이터가 서로 다를 확률이므로, 원본 데이터 x의 포함여부와 관련하여 $P[s=1, x \in y^*]$와 $P[s=1, x \notin y^*]$로 나누어 계산할 수 있다. x가 x로 랜덤화될 확률은 $\gamma/(\gamma+N-1)$이고, 중복이 일어나지 않고 x가 아닌 데이터로 랜덤화될 확률은 랜덤화가 수행되는 순서에 따라 적당한 $i (1 \le i \le l-1)$에 대하여 $(N-i)/(\gamma+N-1)$이다. l번의 랜덤화가 진행되는 동안 각각의 랜덤화 과정은 서로 독립이고, y^* 내에서 x가 발생하는 l가지 순서의 경우를 모두 고려하면,

$$P[s=1, x \in y^*] = \frac{l\gamma \prod_{i=1}^{l-1}(N-i)}{(\gamma+N-1)^l}$$

이 된다. 유사한 방법으로 $P[s=1, x \notin y^*]$는

$$P[s=1, x \notin y^*] = \frac{\prod_{i=1}^{l}(N-i)}{(\gamma+N-1)^l}$$

이다. 그리고 $P[s=1]$는 $P[s=1, x \in y^*]$와 $P[s=1, x \notin y^*]$의 합이므로,

$$P[s=1] = \frac{N-l+l\gamma}{(\gamma+N-1)^l} \prod_{i=1}^{l-1}(N-i)$$

이다. 결국 중복된 데이터가 발생할 확률은

$$P[s \ge 2] = 1 - \frac{N-l+l\gamma}{(\gamma+N-1)^l} \prod_{i=1}^{l-1}(N-i)$$

와 같이 계산된다.

정리 5.1에 의하면 중복된 데이터가 발생할 확률은 N에 반비례하고, γ와 l에 비례하여 증가한다. l-확장 기법은 N의 값이 비교적 클 경우에 사용이 되므로 중복이 일어날 확률은 적을 것이다. 하지만 중복이 일어날 확률이 적다는 것이지 없다는 것은 아니다. 정확성은 직관적으로 l 값에 비례하고, 프라이버시는 l 값에 반비례하므로 주어진 데이터 집합과 목적에 따른 최선의 l 값을 선택하여야 한다. 그리고 어느 정도 발생할 것으로 기대되는 중복 데이터에 대해서는 동일한 안전성을 위해 중복이 발생하지 않도록 중복이 발생한 속성 값에 대해서 재랜덤화를 하여 중복이 일어나지 않도록 한다. 이러한 아이디어를 적용한 l-확장 랜덤대치 알고리즘은 다음과 같다.

l-확장 랜덤대치의 재구축은 정리 3.2와 동일한 방법으로 분포 $\widehat{X^*}$를 계산하면,

$$E[\widehat{X^*}] = l \cdot E[\widehat{X}]$$

알고리즘 2. l-확장 랜덤대치 기법

입력 : n개의 레코드로 이루어진 원본 데이터 집합 O, 속성 A에 대한 정의역 $U=\{u_1,\cdots,u_N\}$, U에 대한 변환행렬 $M_{N\times N}$, $l \geq 1$.

출력 : 변환된 데이터 집합 P

수행 과정 :

모든 $o \in O$에 대해 다음 과정을 실행한다.

1. o가 가지는 속성 값의 인덱스 값 k를 구한다. 즉, o가 가지는 속성 값은 u_k이다.
2. $1 \leq j \leq l$ 에 대하여 다음을 실행한다.($U_0 = \varnothing$)
   2.1. (0, 1] 상의 균등분포로부터 랜덤수 r을 선택한다.
   2.2. 다음을 만족하는 정수 $1 \leqq h \leqq N$ 를 찾는다.

$$\sum_{i=1}^{h-1} m_{ik} < r \leqq \sum_{i=1}^{h} m_{ik}.$$

   2.3. if $u_h \notin U_{j-1}$
   then o의 j번째 랜덤화 속성값 = u_h ,
   $U_j = U_{j-1} \cup \{u_h\}$.
   else go to 2.1.
3. o에 대응되는 변환된 레코드의 속성값의 집합을 U_l로 결정한다.

이 성립한다. 즉, 원본데이터 집합을 l배하여 랜덤화를 수행하였으므로, 최종적으로 재구축된 분포는 $\widehat{X^*}$를 l로 나누어 계산을 해준 값인 $\widehat{X^*}/l$로 정의하면 타당한 추정량(estimator)이 된다. 이 추정량은 l-확장 기법을 적용하지 않은 기존의 랜덤대치 추정량에 비하여 정확성이 개선된 것이다. 다음 소절에서 우리는 개선된 정확성에 대하여 논한다.

5.2 l-확장 랜덤대치 기법의 정확성 분석

l-확장 랜덤대치는 기존 n개의 레코드를 l배 확장 하는 것이므로, 다음과 같은 O^*에서 랜덤대치를 수행한다.

$$\begin{aligned} O^* &= \left[o_1^*, o_2^*, \ldots, o_{ln}^*\right] \\ &= \left[o_1, \ldots, o_1, o_2, \ldots, o_2, \ldots, o_n, \ldots, o_n\right] \end{aligned}$$

따라서 O^*의 속성별 빈도수 X^*는 다음과 같이 표현할 수 있다.

$$X^* = (lX_1, \ldots, lX_N) = l\,X\ .$$

$\widehat{X^*}/l$의 정확도가 X와 비교하여 얼마나 향상 되었는지를 계산하기 위하여, $Var(\widehat{X^*}/l)$ 와 $Var(\hat{X})$를 비교하면 다음과 같다.

[정리 5.2] O^*의 속성별 빈도수를 X^*, 변형된 데이터의 속성별 빈도수를 Y^*라 했을 때, 기존의 랜덤대치와 비교하여 l-확장 랜덤대치의 재구축된 데이터의 분포의 분산 $Var(\widehat{X^*}/l)$ 은 다음과 같다.

$$Var(\widehat{X^*}/l) \;=\; \frac{1}{l}\,Var(\hat{X}).$$

[증명] 우선 Y^*의 분산을 계산하기 위해, 랜덤변수 Y_j^*를 성공 확률 $p_{kj} = \Pr(u_k \rightarrow u_j)$를 가지는 Bernoulli 랜덤 변수 $Y_j^{o_i^*}$로 표현하자. $(o_i^*[A] = u_k)$

$$Y_j^* = \sum_{i=1}^{ln} Y_j^{o_i^*}\ , \quad Y_j^{o_i^*} = \begin{cases} 1, & \text{만일 } o_i^*[A] \rightarrow u_j \\ 0, & \text{그렇지 않은 경우} \end{cases}$$

그러면 $Var(Y_j^{o_i^*}) = p_{kj}(1-p_{kj})$ 이므로, Y_j^*의 분산은 Y_j의 분산과 관련하여 다음을 만족한다.

$$\begin{aligned} & Var(Y_j^*) \\ &= \sum_{i=1}^{ln} Var(Y_j^{o_i^*}) \\ &= (lX_1)p_{1j}(1-p_{1j}) + \cdots + (lX_N)p_{1N}(1-p_{1N}) \\ &= l\,(X_1 p_{1j}(1-p_{1j}) + \cdots + X_N p_{1N}(1-p_{1N})) \\ &= l\left(\sum_{i=1}^{n} Var(Y_j^{o_i})\right) \\ &= l\;Var(Y_j) \end{aligned}$$

따라서 Y^*의 분산은

$$\begin{aligned} Var(Y^*) \;&=\; \sum_{j=1}^{N} Var(Y_j^*) \\ &= l\sum_{j=1}^{N} Var(Y_j) \\ &= l\;Var(Y) \end{aligned}$$

이 된다. 한편, $\widehat{X_j^*} = \frac{\gamma+N-1}{\gamma-1}\,Y_j^* + \frac{nl}{\gamma-1}$ 이므로, 결과적으로

$$\begin{aligned} Var(\widehat{X_j^*}/l) \;&=\; Var\left(\frac{\gamma+N-1}{(\gamma-1)l}\,Y_j^* - \frac{n}{\gamma-1}\right) \\ &= \frac{1}{l^2}\left(\frac{\gamma+N-1}{\gamma-1}\right)^2 Var(Y_j^*) \end{aligned}$$

$$= \frac{1}{l^2}\left(\frac{\gamma+N-1}{\gamma-1}\right)^2 l\, Var(Y_j)$$

$$= \frac{1}{l} Var(\hat{X}_j)$$

와 같이 계산된다.

위의 정리로 부터 $\widehat{X^*}/l$의 X에 대한 정확도는 기존의 랜덤대치와 비교하여 다음과 같이 향상되었음을 알 수 있다.

$$\frac{\sigma_{\widehat{X^*}/l}}{\|X\|} = \frac{1}{\sqrt{l}} \frac{\sigma_{\hat{X}}}{\|X\|}.$$

5.3 l-확장 랜덤대치 기법의 계산량 분석

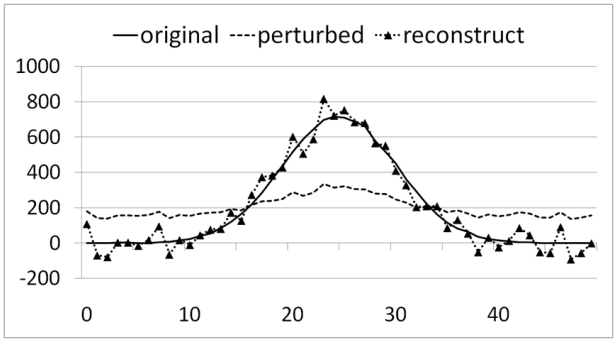
l-확장 랜덤대치 알고리즘은 하나의 데이터 레코드에 대하여 l개의 랜덤화된 데이터를 생성하고, 랜덤화 과정에서 중복이 발생하지 않을 경우 원래의 랜덤대치 방식에 비해 l배 정도의 계산량을 요한다. 정리 5.1에서 보는 바와 같이 중복이 발생할 확률은 파라미터 l, γ, N에 의존하지만, γ와 N은 고정된 값으로 보아야 하므로 l값에 따라 중복 확률은 변한다고 할 수 있다. l값이 커지면 중복이 발생할 가능성은 증가하지만, 실용적인 면에서 정확성 향상이 필요한 데이터는 N이 큰 경우이므로 중복 발생 확률은 상대적으로 낮아지게 된다. 이러한 환경을 고려해볼 때, 실용적 관점에서 l-확장 랜덤대치 알고리즘의 계산복잡도는 기존 랜덤대치 기법에 비하여 l배 보다 심각히 증가하는 현상은 발생하지 않을 것이다. 실제로 6장에서 기술할 시뮬레이션 결과 l의 값이 N값의 10%를 넘지 않을 경우 거의 l배만의 계산량 증가가 발생함을 확인할 수 있다.

6. 시뮬레이션 결과

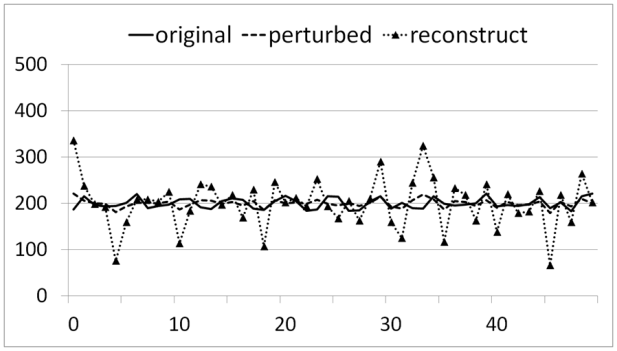
실험은 랜덤대치에서와 동일한 환경에서 l값에 변화를 주어 관찰하였다. 다음 각각의 그림은 $l=4$일 때의 l-확장 랜덤대치 결과를 나타낸다.

(그림 4)와 (그림 5)의 실험을 l-확장 랜덤대치로 수행한 결과는 (그림 8)과 (그림 9)에 나타나 있다. 5.1절의 l-확장 랜덤대치의 정확성 분석에 의하여, 원본 데이터 집합을 $l=4$로 확장하여 랜덤대치를 수행할 경우, 정확성이 2배 향상될 것이라 기대할 수 있다. 따라서 각각 0.2793과 0.4204의 1/2배인 0.1396과 0.2102의 표준오차가 생길 것이라 예상된다. 실제로 (그림 8)과 (그림 9)의 표준오차는 0.1655와 0.2281으로 예측 값에 거의 일치함을 확인할 수 있다. 또한 (그림 6)과 (그림 7)의 실험을 l-확장 랜덤대치로 수행한 결과는 (그림 10)과 (그림 11)로 나타난다. 각각의 실제 표준오차는 이전 랜덤대치에 비해 거의 1/2배인 0.1880과 0.2529임을 확인할 수 있다.

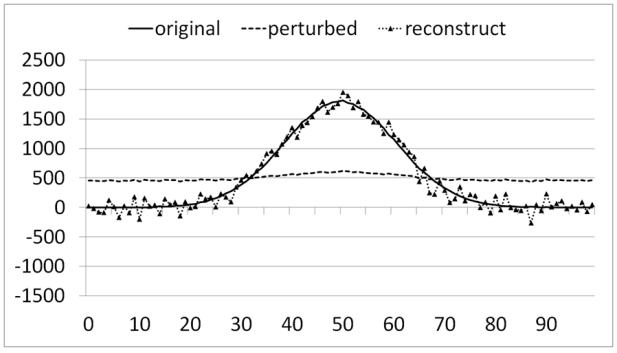
다음 표는 각 파라미터별 l값의 변화가 정확성에 끼치는 영향을 측정하기 위한 실험으로 5.1절에서 계산한 표준오차



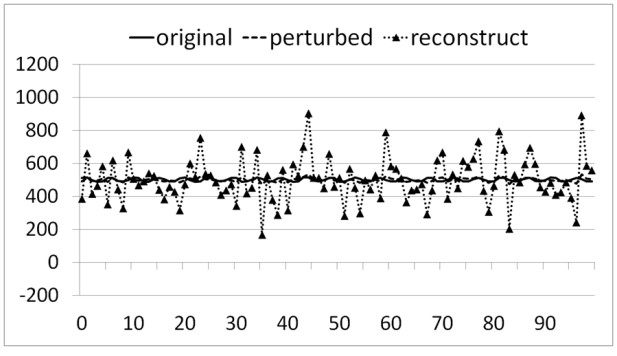
(그림 8) $N=50$, $n=10{,}000$일 때, 속성값 분포(정규)



(그림 9) $N=50$, $n=10{,}000$일 때, 속성값 분포(균등)



(그림 10) $N=100$, $n=50{,}000$일 때, 속성값 분포(정규)



(그림 11) $N=100$, $n=50{,}000$일 때, 속성값 분포(균등)

를 각 파라미터별로 100회 반복실험을 실시하여 계산한 표이며, 기존의 랜덤대치 알고리즘은 $l=1$일 때와 같다.

위의 결과에서는 실제 표준오차의 감소 비율이 $\sqrt{l}$에 약간 못 미치는 것을 확인 할 수 있다. 이는 l-확장 랜덤대치에서 임의의 데이터가 l개로 랜덤화 될 때, 랜덤대치와 동일한 프라이버시를 유지하기 위하여 재랜덤화를 시행하여 랜덤화가 다소 균등하게 일어나지 않은데 원인이 있다. 따라서 l-확장 랜덤대치의 실행 시에는 세 파라미터 (n, N, γ)에 따라 충돌이 되도록 최소가 되도록 l값을 결정하여야 할 것이다.

〈표 1〉 $n=5{,}000$ 일 때, 표준오차 비교(정규)

		$l=1$	$l=2$	$l=4$
$N=50$ $\gamma=5$	실제 표준오차	0.6812	0.4760	0.3401
	이론적 상계	1.3328	0.9424	0.6664
	계산량 (랜덤 횟수)	5,000	10,131	20,806
$N=100$ $\gamma=10$	실제 표준오차	0.8837	0.6082	0.4444
	이론적 상계	1.6984	1.2010	0.8492
	계산량 (랜덤 횟수)	5,000	10,088	20,529

〈표 2〉 $n=50{,}000$ 일 때, 표준오차 비교(정규)

		$l=1$	$l=2$	$l=4$
$N=50$ $\gamma=5$	실제 표준오차	0.2623	0.1862	0.1589
	이론적 상계	0.4214	0.2980	0.2107
	계산량 (랜덤 횟수)	50,000	101,326	208,105
$N=100$ $\gamma=10$	실제 표준오차	0.3358	0.2380	0.1817
	이론적 상계	0.5370	0.3800	0.2685
	계산량 (랜덤 횟수)	50,000	100,884	205,274

〈표 3〉 $n=5{,}000$ 일 때, 표준오차 비교(균등)

		$l=1$	$l=2$	$l=4$
$N=50$ $\gamma=5$	실제 표준오차	1.3351	0.9409	0.6341
	이론적 상계	1.3328	0.9424	0.6664
	계산량 (랜덤 횟수)	5,000	10,134	20,810
$N=100$ $\gamma=10$	실제 표준오차	1.6918	1.1773	0.8209
	이론적 상계	1.6984	1.2010	0.8492
	계산량 (랜덤 횟수)	5,000	10,087	20,525

〈표 4〉 $n=50{,}000$ 일 때, 표준오차 비교(균등)

		$l=1$	$l=2$	$l=4$
$N=50$ $\gamma=5$	실제 표준오차	0.4289	0.2926	0.2054
	이론적 상계	0.4214	0.2980	0.2107
	계산량 (랜덤 횟수)	50,000	101,331	208,130
$N=100$ $\gamma=10$	실제 표준오차	0.5374	0.3763	0.2617
	이론적 상계	0.5370	0.3800	0.2685
	계산량 (랜덤 횟수)	50,000	100,884	205,269

7. 결 론

정확성 개선을 위한 l-확장 랜덤대치는 앞서 제안된 랜덤대치의 프라이버시 수준을 유지하면서 보다 정확하게 원본 데이터의 분포를 재구축할 수가 있다. 이로 인해 연관규칙 마이닝, 의사결정나무 마이닝 등 여러 데이터 마이닝 기법에서 SMC등과 결합되어 보다 실용적으로 사용될 수 있다. 하지만 정확도가 증가하는 만큼의 계산량이 늘어나기 때문에 앞으로도 정확도와 안전성, 그리고 계산량에 대한 심도 있는 연구가 지속적으로 수행되어야 할 것으로 보인다.

참 고 문 헌

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

[2] Y. Lindell, B. Pinkas, “Privacy preserving data mining”, CRYPTO 2000, pp.36-54.

[3] J. Vaidya, C. Clifton, “Privacy-Preserving Data Mining: Why, How, and When”, IEEE Security & Privacy, 2004, www.computer.org/security/

[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] S. Agrawal and J. Haritsa, “A Framework for High-Accuracy Privacy-Preserving Mining”, Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), IEEE, 2005.

[6] S. Agrawal and J. Haritsa, “A framework for high-accuracy privacy-preserving mining”, Tech- nical Report TR-2004-02, *Database Systems Lab, Indian Institute of Science*, 2004.

[7] S. Agrawal, J. Haritsa, and B. Prakash, “FRAPP: a framework for high-accuracy privacy- preserving mining”, Data Mining and Knowledge Discovery, Springer, Vol.18, No.1, 2009, pp.101-139.

[8] 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.

[9] D. Agrawal and 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.

[10] J. Dowd, S. Xu, and W. Zhang, “Privacy- Preserving Decision Tree Mining Based on Random Substitutions”, ETRICS2006, LNCS 3995, Springer-Verlag, pp.145-159, 2006.

[11] 강주성, 안아론, 홍도원, “행렬기반 랜덤화를 적용한 프라이버시 보호 기술의 안전성 및 정확성 분석”, 한국정보보호학회논문지, 제18권 4호, pp.53-68, 2008.

강 주 성

e-mail : jskang@kookmin.ac.kr
1989년 고려대학교 수학과(학사)
1991년 고려대학교 수학과(이학석사)
1996년 고려대학교 수학과(이학박사)
1997년~2004년 한국전자통신연구원 선임연구원
2004년~현 재 국민대학교 수학과 부교수
관심분야 : 암호 이론, 정보보호 이론, 응용 확률론 등

이 창 우

e-mail : foolfor@naver.com
2007년 국민대학교 수학과(학사)
2009년 국민대학교 수학과(이학석사)
관심분야 : 암호 이론, 정보보호 이론 등

홍 도 원

e-mail : dwhong@etri.re.kr
1994년 고려대학교 수학과(학사)
1996년 고려대학교 수학과(이학석사)
2000년 고려대학교 수학과(이학박사)
2000년~현 재 한국전자통신연구원 지식정보보안연구부 선임연구원
관심분야 : 암호 이론, 정보보호 이론, 이동통신 정보보호 등