랜덤대치 기반 프라이버시 보호 기법의 효율적인 구현 및 안전성 분석[1)]

안아론*, 강주성*[2)], 홍도원**
*국민대학교 수학과
**한국전자통신연구원 정보보호연구단
e-mail : jskang@kookmin.ac.kr

Efficient Implementation and Security Analysis of Privacy-Preserving Technique based on Random Substitutions

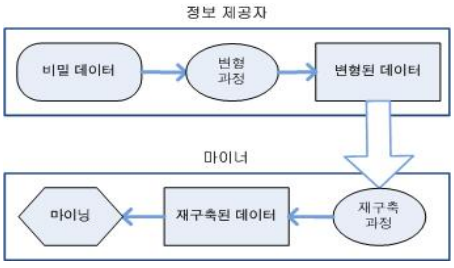
Aron An*, Ju-Sung Kang*, Dowon Hong**
*Dept. of Mathematics, Kookmin University
**Information Security Research Division, ETRI

요 약

본 논문에서는 랜덤대치(random substitution) 기법에 대하여 심도 있는 분석을 실시한다. 랜덤대치 기법의 효율적인 구현을 위하여 데이터 재구축(reconstruction) 과정에서 필요로 하는 역행렬을 구하는 공식을 제시한다. 또한, 랜덤대치에 사용되는 다양한 파라미터들의 의미를 실험적으로 밝혀내며, 정확도와 프라이버시를 합리적으로 측정할 수 있는 새로운 측도(measure)들을 제안한다.

1. 서론

실용적인 프라이버시 보호 기술은 대부분 다양한 랜덤화(randomization) 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(그림 1). 데이터의 변형은 프라이버시 관련 정보를 노출시키지 않기 위함이며, 데이터 변형의 가장 실용적인 방법이 랜덤화 기법이다. 최근에 발표된 랜덤대치(random substitutions)는 여러 가지 랜덤화 기법 중의 하나로 안전성과 효율성이 높고, 다양한 분야에 응용 가능한 방법으로 알려져 있다[1].



(그림 1) 랜덤 대치 방법의 구조

랜덤화 기법과 관련하여 최근에 발표된 의미 있는 연구 결과로 랜덤회전(random rotation)[2] 기법과 랜덤사영(random projection)[3] 기법을 들 수 있다. 이 두 가지 방법은 비교적 높은 정확도(accuracy)를 가지지만 계산량적인 효율성과 실제 데이터에 적용하는 관점에서 실용성이 높다고 할 수 없다. 특히, 랜덤회전 기법에서는 사용되는 변환행렬에 대한 안전성적인 문제점이 지적되기도 하였다[3]. 한편, 랜덤대치 기법은 위의 두 기법에 비해 구현이 용이하고 응용 가능성이 높다는 장점을 지니고 있으며, 사용되는 파라미터를 이용하여 프라이버시와 정확도의 취사선택(tradeoff)이 가능하므로 데이터를 취급하는 개체 사이의 적절한 보안성을 고려할 수 있다.

본 논문에서는 랜덤대치 기법에 대하여 심도 있는 분석을 실시한다. 랜덤대치 기법의 효율적인 구현을 위하여 데이터 재구축(reconstruction) 과정에서 필요로 하는 역행렬(inverse matrix)을 구하는 공식을 제시한다. 이 공식을 사용할 경우 가우스 소거법 등과 같은 일반적인 역행렬 계산 알고리즘의 사용이 필요 없게 된다. 또한, 랜덤대치에 사용되는 다양한 파라미터들의 의미를 실험적으로 밝혀낸다. 그리고 정확도와 프라이버시를 좀 더 합리적으로 측정할 수 있는 새로운 측도(measure)를 제안한다.

2. 랜덤대치 기법

정보제공자는 이산(discrete) 형태의 정의역을 갖는 단일 속성 A에 대한 데이터 레코드들을 가지고 있다고 가정한다. 연속(continuous) 형태 또는 다수의 속성을 갖는 데이터 집합에 대해서는 구간을 설정하거나 단일 속성에

1) 본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장 동력핵심기술개발사업의 일환으로 수행하였음.
[2005-Y001-04, 차세대 시큐리티 기술 개발]

2) 교신저자

대한 각 과정을 여러 번 반복함으로써 쉽게 확장이 가능하다. 본 절에서는 Dowd-Xu-Zhang[1]이 제안한 랜덤대치 기법에 대하여 간략히 소개한다.

2.1 원본 데이터의 변형 과정

랜덤대치 기법의 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 전환행렬(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.

$$\Pr[u_k \rightarrow u_h] = m_{h,k} \ .$$

이렇게 정의된 확률 값 $m_{h,k}$를 성분으로 하는 $N\times N$ 크기의 행렬을 M이라 놓는다. 각 속성 값은 자기 자신을 포함해서 반드시 U안에 있는 값으로 바뀌기 때문에 각 열의 합은 1이 된다. 그러므로 행렬 M에서 각 열은 확률함수로 정의 될 수 있고, 열의 누적 확률함수를 이용해서 속성 값을 변형한다.

랜덤대치 기법에서 데이터를 변형하는 방법을 알고리즘으로 표현하면 다음과 같다.

알고리즘 1. 랜덤대치 기법의 데이터 변형 알고리즘

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

결과 : 변환된 데이터 집합 P

방법 :

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 레코드 $p \in P$의 속성 값을 U상에서 h의 인덱스 값을 갖는 u_h로 결정한다.

데이터 변형 알고리즘의 복잡도는 $O(n \cdot N)$이다. 프라이버시와 정확도를 FRAPP 프레임워크[4]에서 제안한 γ를 사용하여 저자들이 [1]에서 제시한 최적의 변형행렬은 $M = xG$의 형태를 가지는 다음과 같은 γ-대각 행렬(γ-diagonal matrix])이다.

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

여기에서 행렬 G는 최소 상태수(condition number)를 갖는다. 행렬의 상태수란 임의의 두 벡터 u, v에 대해서 $u = Gv$의 수식이 주어졌을 때, v에서 발생하는 오차에 대해 G가 얼마나 민감하게 u에 영향을 주는지를 나타내는 수이다. 예를 들어, G가 큰 상태수를 갖는다면, v에서의 작은 오차가 곱셈 결과인 u에서 오차가 커지는 영향을 줄 수 있다. 즉, 행렬 G가 최소 상태수를 갖는다는 것은 재구축 과정에서 최대의 정확도를 갖는 행렬임을 의미한다. γ는 변형 과정의 프라이버시가 아무런 지식 없이 원본 데이터를 추측할 확률(prior probability)이 ρ_1, 변형 데이터를 얻고 난 후, 원본 데이터에 대한 추측 확률(posterior probability)이 ρ_2인 ρ_1-to-ρ_2 프라이버시 손상(privacy breach)[5]에 의해 관측 되었을 때, 다음 부등식으로 결정된다.

$$\gamma \leqq \frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)} \ .$$

재구축 과정은 변형된 데이터 집합으로부터 원본 데이터 레코드의 분포를 추정하기 위해 이루어진다[1,4]. 데이터 분포의 추정은 변형된 데이터 집합 P와 변형 행렬 M을 이용한다. 즉, M은 제공자와 마이너가 사전에 미리 설정하거나 공개된다. 각 $u_i \in U$에 대해서, Y_i를 변형된 데이터 집합에서 u_i의 개수라고 하고, X_i를 원본 데이터 집합의 u_i의 개수라고 하자. 즉, $X=(X_1,\cdots,X_N)^T$와 $Y=(Y_1,\cdots,Y_N)^T$를 각각 원본 그리고 변형된 데이터 집합에서 각 레코드들이 갖는 속성의 개수에 대한 열벡터라고 하자. 그러면 주어진 O에 대해서 Y에 대한 기대값은 다음과 같이 구할 수 있다.

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

만약 M이 가역이고 $E[Y]$가 알려져 있다면, 아래의 방정식을 풀어냄으로써 X를 구할 수 있다.

$$X = M^{-1}E[Y] \ .$$

한편, X의 분포는 공개되지 않으므로 $E[Y]$를 MX로부터 계산해낼 수 없다. 그러므로 정확한 X는 알아내기 어렵게 된다.

X를 추정하기 위해서 변형 데이터 집합의 u_i의 개수에 대한 벡터 Y의 관측값 $\boldsymbol{y} = (y_1,\cdots,y_N)$을 이용하면 X에 대한 추정량(estimator) $\widehat{X}$를 얻을 수 있다. 즉,

$$\widehat{X} = (\widehat{X_1},\cdots,\widehat{X_N})^T = M^{-1}\boldsymbol{y}$$

로 놓는다. 여기에서 $E[\widehat{X}] = M^{-1}E[Y] = X$이므로, 추정량 $\widehat{X}$은 추정값의 기댓값이 원래의 값 X와 일치하는 무편향 추정량(unbiased estimator)이 된다. 그러므로 $\widehat{X}$를 재구축된 데이터 분포로 보면, 원본 데이터의 분포와 유사한 분포를 얻게 될 것이다.

3. 효율적인 분포 재구축 방법

랜덤대치 기법의 효율적인 구현을 위하여 데이터 재구축 과정에서 필요로 하는 역행렬을 구하는 공식을 본 절에서 제안한다.

정리 1. $\gamma > 1$, $N > 1$인 γ-대각 행렬 M에 대해서, M의 역행렬 $M^{-1} = \left(m_{ij}^{-1}\right)_{N\times N}$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 다음의 두 가지 형태의 성분들로 이루어진다.

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

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

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

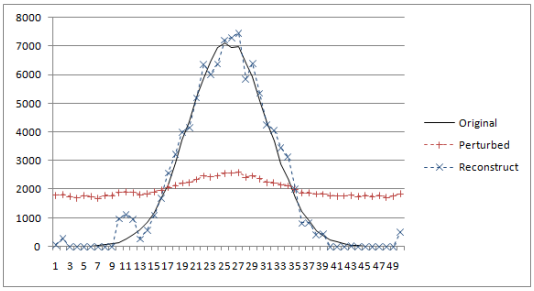
정리 1에서와 같이 M^{-1}를 계산할 경우, 실제로 데이터 분포 재구축 과정에서 가우스 소거법 등과 같은 복잡한 역행렬 계산 알고리즘 수행에 걸리는 시간을 크게 단축시켜 주므로 분포의 재구축 과정을 효율적으로 진행시킬 수 있다.

4. 랜덤대치 기법의 정확성 및 안전성

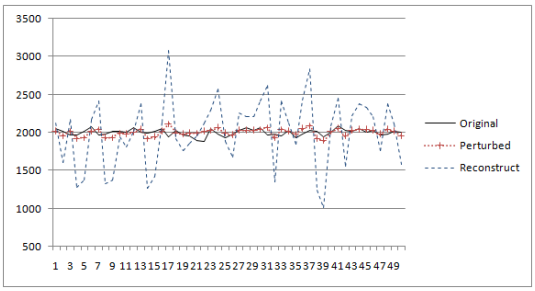
랜덤대치 기법의 정확성은 원본 데이터의 분포와 재구축된 분포와의 차이로 정의되는 에러를 통해 측정 될 수 있다. 재구축된 분포에 대한 벡터를 $R = M^{-1}\boldsymbol{y}$ 라고 하자. 3절에서 볼 수 있듯이 M^{-1}의 비대각성분이 모두 음수이기 때문에 재구축 과정에서 R의 성분이 음수 값을 가지게 되는 경우가 발생할 수 있다. 음수 값을 갖는 성분을 0으로 바꿔주면 에러를 좀 더 줄일 수 있다[1]. 이처럼 에러를 줄이고, 속성의 개수를 나타내는 정수 값을 부여하기 위하여 R을 보정한 추정량 $\hat{X}$ 을 다음과 같이 정의한다.

$$\forall i \in \{1,\cdots,N\}, \quad \hat{X}_i = \begin{cases} 0 & , R_i \leqq 0 \\ \lfloor R_i \rfloor & , R_i > 0 \end{cases} .$$

실험에서 데이터 레코드는 정규분포 또는 균등분포를 따르는 100,000개의 레코드로 γ를 2에서 21까지, N은 10, 20, 25, 50, 75, 90, 100으로 변화시키면서 실험하였으며, 각 실험을 10회 반복하여 측정된 에러의 평균을 관찰하였다. 그림 2는 원본(Original) 데이터의 속성 값이 정규분포를 따르는 경우에 변형된(Perturbed) 데이터의 속성 값 분포와 재구축(Reconstruct) 분포를 나타낸 것이다. 그림 3은 균등분포에 대한 각 데이터 속성값의 분포이다.



(그림 2) 각 데이터 속성값에 대한 분포 (정규)



(그림 3) 각 데이터 속성값에 대한 분포 (균등)

구체적인 에러 측정을 위해 [1]에 제안된 분포의 차이에 대한 방법 이외에 보다 합리적인 정확도 측정을 위하여 두 가지 측도(measure)를 추가로 제안한다.

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

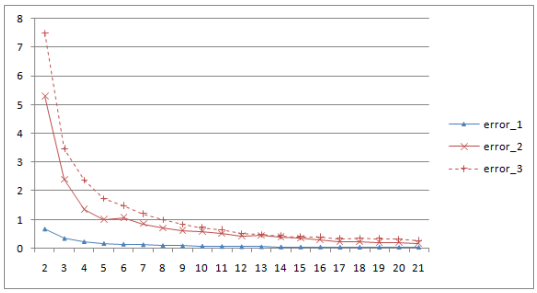
$$error_1 = \frac{1}{n}\sum_{j=1}^{N} |\hat{X}_i - X_i| \quad .$$

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

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

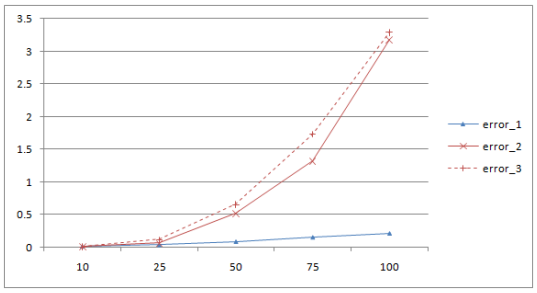
(3) $error_3$: 원본 데이터의 각 속성 값에 대한 표준편차 σ와 재구축된 각 속성 값에 대한 표준편차 $\hat{\sigma}$의 차이.

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



(그림 4) 정규분포에서 γ에 대한 에러 (N=50)

그림 4에서 정규분포 상의 고정된 N에 대해 γ가 감소할수록 에러가 증가하는 것을 볼 수 있다. 에러의 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.

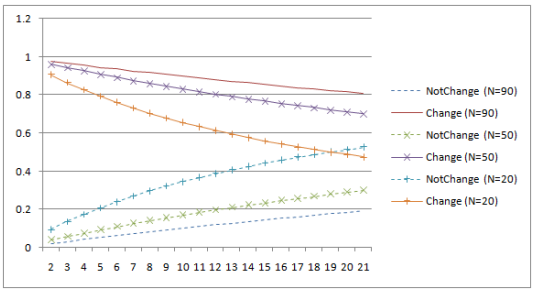


(그림 5) 정규분포에서 N에 대한 에러 (γ=11)

그림 5에서 정규분포상의 고정된 γ에 대해 N이 증가할수록 에러는 증가하는 것을 볼 수 있는데, 그 이유는 역행렬의 비대각성분이 N에 관계없이 일정하기 때문이다. N이 크면 각 속성값의 개수는 전체적으로 작은 값이 많이 지게 되고, 주변값보다 상대적으로 작은 값에 대한 재구축은 음수값을 갖는 값을 도출해낼 가능성이 커진다. 그러므로 에러 역시 증가하게 되고 정확도가 감소하게 된다.

$error_1$으로는 분포 사이의 전체적인 에러에 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 에러는 원본 분포와 재구축 분포의 구체적인 차이를 나타내는 파라미터가 된다. 위의 에러실험은 균등분포에서도 비슷한 결과를 가진다.

프라이버시 보호 정도는 γ와 N의 변화에 관련된 에러 측정방법 외에 또 다른 관점에서도 측정해볼 수 있다. 각 원본 레코드 속성값이 변형될 때, 자기 자신이 아닌 다른 속성값으로 변한 레코드가 많다면 프라이버시 보증이 높다고 볼 수 있다. 그림 6에서 보면 γ가 증가할수록 속성값이 바뀐 비율이 점점 감소하고 그에 따라 바뀌지 않은 비율이 점점 증가함을 알 수 있다.



(그림 6) N과 γ에 대한 속성값 변화비율

5. 결론

랜덤대치는 앞서 제안한 효율적인 γ-대각행렬의 역행렬을 구하는 방법을 이용하면 계산적인 관점에서 부하를 많이 줄일 수 있다. 그리고 단순한 분포값의 차이가 아닌 원본 데이터와 재구축된 분포의 평균과 편차로서 에러를 보다 구체적으로 측정하였다. 프라이버시 측정에서는 속성값 변화에 대한 실질적 비율을 새로운 측도로 제안하였다.

랜덤대치는 원본 데이터 분포를 재구축한다는 측면에서 연관규칙 마이닝, 의사결정나무 마이닝 등 여러 데이터 마이닝 기법에 프라이버시 보호 기술로 활용될 수 있기 때문에 정확도와 안전성에 대한 심도 있는 연구가 지속적으로 수행되어야 할 것으로 보인다.

참고문헌

[1] 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.

[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] Shipra Agrawal, and Jayant R. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining", Proc. of ICDE 2005, 2005.

[5] 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.