k-익명성 알고리즘 관련 측도들에 대한 비교 분석[1)]

신 윤 경, 강 주 성[2)]
국민대학교 수학과

Comparisons of several measures related to k-anonymity algorithms

Youn-kyoung Shin, Ju-Sung Kang
Department of Mathematics, Kookmin University

요 약

개인정보 노출 위험을 최소화하면서 데이터 유용성을 최대화하기 위한 기법 중의 하나인 k-익명성 개념과 연관된 다양한 측도(measure)들을 비교 분석한다. 원본 데이터와 변형된 데이터가 주어졌을 때, 각각 다른 k-익명성 알고리즘들에서 제안된 높이(height), 정확도(precision), 손실측도(loss metric), 비용(cost), 점수(score) 등의 측도들이 데이터의 정확성(accuracy)을 측정하는 데에 대한 일관성과 개별성을 조사하고, 그 측도들의 특징에 따른 의미와 효율성을 비교분석한다.

I. 서론

개인정보 노출 위험성 문제를 해결하기 위해서 Sweeny-Samarati[1]는 k-익명성(k-anonymity) 개념을 제안하였다. k-익명성은 k-익명화된 데이터 집합에서의 각 기록들이 적어도 k-1개의 다른 레코드들과 구별되지 않는다는 성질을 의미한다. k의 값이 커질수록 개인이 식별될 확률은 $1/k$을 넘지 못하게 되어 개인정보의 노출 위험성(disclosure risk)은 줄어든다. 하지만 데이터의 이용 가치를 말하는 데이터 유용성(data utility)이 변화하게 되는 단점이 생긴다. 따라서 데이터 변형 알고리즘의 궁극적인 목표는 개인정보의 노출 위험을 최소화하면서 데이터 유용성은 최대화하는 것이다.

본 논문에서는 이러한 k-익명성 알고리즘들에 의해 변형된 마이크로 데이터의 정확성(accuracy)을 측정하기 위한 다양한 측도(measure)들을 비교 분석하고자 한다. 주목할 만한 기존 측도들 중 높이(height)[2], 정확도(precision)[3], 손실측도(loss metric)[4], 비용(cost)[5], 점수(score)[6] 등을 고려한다. 이 다섯 개의 측도들은 각각 다른 기준에서 정의되고, 일반화 기법을 사용하는 상향식 알고리즘과 특수화 기법을 사용하는 하향식 알고리즘들에서 서로 다르게 사용되므로 통합적 비교 분석에 어려움이 따른다. 그러나 우리는 가능한 균형 있고 공정한 시각으로 합리적인 비교 분석을 실시하고자 한다. 원본 데이터 테이블 PT(private table)과 변형된 데이터 테이블 GT(generalized table)이 동일하게 주어졌을 때, 각 측도들의 마이크로 데이터의 정확성에 대한 일관성과 개별성을 조사하고, 그 측도들의 특징에 따른 의미와 효율성을 비교분석한다.

1) 본 연구는 지식경제부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2005-Y001-04, 차세대 시큐리티 기술 개발]
2) 교신저자

II. k-익명성 개념

일반적으로 개인정보를 포함한 데이터는 행과 열로 이루어진 표(table)로 구성된다. 한 행의 데이터를 기록(record)이라 하며, 하나의 열은 속성(attribute)이라 부르고, n개의 속성을 갖는 데이터 표를 $T[A_1,\cdots,A_n]$와 같이 나타낸다. 표에 포함된 속성은 두 가지로 나뉜다. 식별자(identifier) $I_1, I_2, \cdots, I_l$은 이름이나 주민등록번호와 같이 개인을 식별할 수 있는 속성을 말한다. 성별이나 우편번호와 같은 속성들은 준식별자(quasi-identifier)라 하여 $K_1, K_2, \cdots, K_m$으로 표현한다.

정의 1. k-익명성(k-anonymity)
데이터 표 $T[A_1,\cdots,A_n]$와 준식별자 K_T가 주어졌을 때, T가 k-익명성을 만족한다는 의미는 $T[K_T]$의 모든 기록들이 k개 이상 $T[K_T]$에 존재한다는 것이다.

k-익명성을 만족하는 데이터 표가 배포될 경우, 이를 소유한 사람은 특정인이 배포된 데이터 내에 포함되어 있다는 사실은 알 수 있을지 모르지만, 정확히 어떤 기록이 특정인의 것인지 알 수 있는 확률은 $1/k$ 이하가 된다.

<표 1> PT를 일반화한 테이블 GT의 예

S_0	Z_0
남	48201
남	48275
남	41076
남	41088
남	41099
여	48201
여	48275
여	41076
여	41088
여	41099

PT

S_1	Z_0
*	48201
*	48275
*	41076
*	41088
*	41099
*	48201
*	48275
*	41076
*	41088
*	41099

$GT_{[1,0]}$

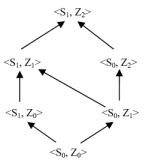
S_0	Z_1
남	482**
남	482**
남	410**
남	410**
남	410**
여	482**
여	482**
여	410**
여	410**
여	410**

$GT_{[0,1]}$

Ⅲ. 정확성 관련 측도

3.1. 높이(height)

높이(height) 개념은 Truta-Bindu[2]에 의하여 소개된 것으로 두 개 이상의 속성을 일반화시킬 때, 데이터 소유자가 (그림 1)과 같이 일반화된 도메인의 가능한 모든 조합을 나타내는 일반화 격자(GL)를 고려하는 것에서 출발한다. GL에서의 최소원 $\langle S_0, Z_0\rangle$에서 노드 X까지의 최단거리를 $height(X, GL)$로 나타낸다. (그림 1)에서 노드들의 높이는 다음과 같이 계산된다.



(그림 1) 일반화 격자(GL)

$height(\langle S_0, Z_0\rangle, GL)=0$, $height(\langle S_1, Z_0\rangle, GL)=1$,
$height(\langle S_0, Z_1\rangle, GL)=1$, $height(\langle S_1, Z_1\rangle, GL)=2$,
$height(\langle S_0, Z_2\rangle, GL)=2$, $height(\langle S_1, Z_2\rangle, GL)=2$.

3.2. 정확도(precision)

정확도(precision)는 일반화 기법을 소개하면서 데이터 유용성을 측정하기 위하여 Sweeny[3]가 제안한 측도이다. <표 1>은 (그림 1)의 일반화 격자에 의한 속성 단위로 시행되는 PT의 일반화를 보여준다. $GT_{[1,0]}$는 성별을 한 단계 일반화한 것이고, $GT_{[0,1]}$은 우편번호를 한 단계 일반화한 것이다. 3.1절에서 소개한 측도에 의하면 두 개의 데이터 표는 1이라는 같은 높이를 갖는다. 그러나 두 개의 데이터 표를 보면 $GT_{[0,1]}$가 $GT_{[1,0]}$에 비해 더 많은 정보를 포함하고 있다는 것을 직관적으로 알 수 있다. 이것은 각 속성에 대해서 똑같이 한 단계의 일반화가 진행된 데이터 표라 할지라도 $GT_{[1,0]}$의 경우 가장 일반화된 값으로 일반화하였다는 것을 감안하지 않았기 때문이다. 전체 일반화 단계에서 몇 단계의 일반화가 진행되었는지를 고려한 측도가 정확도이다.

정의 3. $PT(A_1, \ldots, A_{N_A})$와 $GT(A_1, \ldots, A_{N_A})$가 주어졌을 때, GT의 정확도 $Prec(GT)$는 다음과 같이 정의된다.

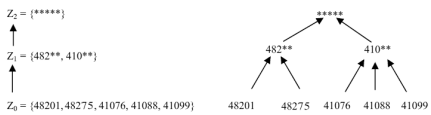
$$Prec(GT) = 1 - \frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N}\frac{h_{ji}}{|DGH_{A_i}|}}{N \bullet N_A}.$$

여기에서 h_{ji}는 j번째 기록에서 속성 A_i가 일반화된 계층의 수를 의미하고, $|DGH_{A_i}|$는 속성 A_i의 전체 계층의 수를 의미한다.

<표 1>의 예에서 일반화된 각 데이터 표에 대한 정확도를 구해보면 다음과 같다.

$Prec(GT_{[1,0]})=0.5$, $Prec(GT_{[0,1]})=0.75$,
$Prec(GT_{[1,1]})=0.25$, $Prec(GT_{[0,2]})=0.5$.

네 가지 일반화된 데이터 표 모두 2-익명성 조건을 만족하므로, 이 중에서 가장 정확도가 높은 $GT_{[0,1]}$을 선택하는 것이 현명하다고 볼 수 있다.



(그림 2) 도메인 일반화와 값 일반화 계급

3.3. 손실측도(loss metric)

손실측도(loss metric)는 Iyengar[4]에 의해서 제안된 측도이다. (그림 2)에서의 값 일반화 계급과 <표 1>에서의 $GT_{[0,1]}$을 살펴보자. 값 48201이 482**으로 일반화된 것과 41076이 410**으로 일반화된 것은 앞에서 제시했던 두 가지 측도에서 모두 같은 값으로 간주된다. 하지만 손실측도는 하나의 속성을 일반화하는 경우 몇 개의 값과 구별되지 않느냐에 기초한 것으로 몇 개의 값이 하나의 값으로 그룹화 되는지에 주목한다. 48201이 482**으로 일반화되면 48275와 구별되지 않고 41076이 410**으로 일반화되면 41088과 41099와 구별되지 않으므로 각 일반화가 다른 손실을 갖게 된다.

정의 4. 값 일반화 계급에서 모든 하위 노드들의 수를 M이라 하고 노드 P에서 뿌리내린 하위 노드들의 수를 M_P라고 할 때, 손실측도 LM은 다음과 같이 정

의된다.

$$LM(GT)=\sum_{i=1}^{N_A}\frac{\sum_{j=1}^{N}\frac{M_P-1}{M-1}}{N}$$

정확도는 데이터 표에서 열 단위로 모든 값들이 같은 손실을 갖는 반면, 손실측도는 하나의 열에서도 값들에 따라 다른 손실을 갖게 되는 차이점이 있다. <표 1>의 예에서 일반화된 각 데이터 표에 대한 손실 측도를 구해보면 다음과 같다.

$LM(GT_{[1,0]})=1$, $LM(GT_{[0,1]})=0.4$,
$LM(GT_{[1,1]})=1.4$, $LM(GT_{[0,2]})=1$.

3.4. 비용(cost)

앞에서 소개된 세 가지 측도는 상향식인 일반화 기법에서 사용되는 반면, 앞으로 소개될 두 가지 측도는 하향식인 특수화 기법에 사용된다. 비용(cost) 개념은 Roberto-Bayardo-Agrawal[5]에 제안되어 있는 측도이다. 비용을 계산하기 위해서는 각 속성들이 가지는 모든 값들에 전체적으로 순서를 정한 후, 그 값들을 원소로 갖는 모든 가능한 부분집합을 고려한다. 속성들의 모든 값들이 n개라면, 2^n개의 익명화된 데이터 표를 만들 수 있다. 이 데이터 표 중에서 최적의 k-익명화를 찾기 위해서 비용을 계산한다. 비용은 익명화된 데이터 표에서 하나의 기록이 다른 기록들과 얼마나 구별되지 않느냐에 기초한 측도이다. 비용의 정의는 다음과 같다.

정의 5. k-익명화 관점에서 일반화된 데이터 GT의 비용 $C(GT,k)$는 다음과 같이 정의된다.

$$C(GT,k)=\sum_{|E|\geq k}|E|^2+\sum_{|E|<k}|D||E|\ .$$

여기에서 $|E|$는 동치클래스 E에 속하는 기록들의 개수를 나타낸다.

앞에서의 측도들은 원래의 데이터 표에서 익명화된 데이터 표로 변환되었을 때의 일반화 정도를 가지고 높이, 정확도, 손실을 나타낸다. 반면 비용은 원래의 데이터가 얼마나 변환되었는지는 상관없이 단순히 익명화된 데이터 표가 얼마만큼의 익명성을 갖고 있는지 나타낸다는 특징이 있다. <표 1>의 예에서 일반화된 각 데이터 표에 대한 손실 측도를 구해보면 다음과 같다.

$C(GT_{[1,0]},2)=20$, $C(GT_{[0,1]},2)=26$,
$C(GT_{[1,1]},2)=50$, $C(GT_{[0,2]},2)=50$.

3.5. 점수(score)

점수(score)는 Fung-Wang-Yu[6]에 의해서 제안된 측도로 일반화에 비해 매우 효율적인 것으로 알려진 Top-Down 알고리즘에 사용된 것이다. 저자들은 (그림 2)의 분류나무에서 각 노드들을 연결한 경로들 중의 하나를 컷(cut)이라 부르고, 가장 일반화된 상태의 노드를 지나는 컷을 초기 컷으로 놓아 점수를 계산하여 속성별로 한 단계씩 특수화를 진행하는 과정이 Top-Down 알고리즘이다.

정의 6. 분류나무의 한 노드 v의 점수 $Score(v)$는 다음과 같이 정의된다.

$$Score(v)=\begin{cases}\frac{InfoGain(v)}{AnonyLoss(v)}, & AnonyLoss(v)\neq 0\\ InfoGain(v), & AnonyLoss(v)=0.\end{cases}$$

점수를 계산할 때 사용되는 양은 정보이론의 엔트로피와 연관되어 있다. <표1>의 데이터에서 각 노드별 점수를 계산한 예는 다음과 같다.

$Score$(ANY_성별)=0.529,
$Score$(ANY_우편번호)=0.026,
$Score$(482**)=0.156, $Score$(410**)=0.155.

한편, Fung-Wang-Yu[6]에 의해 정의된 점수(score)는 어떤 속성을 특수화시킬 때 얻어지는 정보량과 손실되는 익명성을 의미한다. 본 논문은 다양한 측도들을 가능한 공정한 시각으로 비교하고자 하므로 점수(score)를 분류나무의 노드별 측도가 아닌 데이터 표의 정보량과 익명성을 나타내는 측도로 사용한다. 따라서 먼저 데이터 표의 일반화 단계를 나타내는 분류나무의 컷을 찾아 그 컷을 지나는 노드에서 부분나무(subtree)를 만들어 부분나무 안에 속하는 모든 노드에서의 점수(score)를 합산한 값으로 정의한다. 이러한 테이블 점수는 데이터 표가 가장 특수화된 데이터가 되기까지 얻어야 할 정보량을 의미한다.

정의 7. 데이터 표 GT의 점수 $Score(GT)$는 다음과 같이 정의된다.

$$Score(GT)=\sum_{v}\sum_{c\in S_v}Score(c)\ .$$

여기에서 S_v는 노드 v에서 뻗어 내린 부분나무이고, c는 S_v의 한 노드이다.

<표 1>의 예에서 각 데이터 표 GT의 테이블 점수를 계산하면 다음과 같다.

$TS(GT_{[1,0]})=0.529$, $TS(GT_{[0,1]})=0.311$,
$TS(GT_{[1,1]})=0.84$, $TS(GT_{[0,2]})=0.337$.

Ⅳ. 측도들 사이의 상호관계 분석

4.1. 데이터 표에 대한 측도 값

데이터 표가 주어졌을 때 계산되는 다섯 가지 측

도들 사이의 일관성과 개별성을 조사해보자. (그림 1)의 일반화 격자에 나타난 여섯 가지 데이터 표에 대한 측도 값들을 비교분석한 결과가 <표 2>에 나타나 있다. <표 2>에서 대소 관계는 데이터 표의 측도들에 대한 순서를 나타내며, 최소 일반화를 우선순위로 한다.

<표 2> 데이터 표에 대한 정확성 측도 비교

측도	데이터 표의 정확성(accuracy)
height	[0, 0]>[1, 0]=[0, 1]>[1, 1]=[0, 2]>[1, 2]
Prec	[0, 0]>[0, 1]>[1, 0]=[0, 2]>[1, 1]>[1, 2]
LM	[0, 0]>[0, 1]>[1, 0]=[0, 2]>[1, 1]>[1, 2]
$C(GT)$	[0, 0]>[1, 0]>[0, 1]>[1, 1]=[0, 2]>[1, 2]
$TS(GT)$	[0, 0]>[0, 1]>[0, 2]>[1, 0]>[1, 1]>[1, 2]

높이는 단순히 일반화 단계 수를 의미하므로 다른 측도들에 비해 데이터 표의 측도 값이 다양하지 못함을 볼 수 있다. 대부분의 경우 이 예와 같이 정확도와 손실 측도는 같은 순서를 갖지만, 종종 다른 결과를 갖기도 한다. 정확도는 하나의 열에 속하는 모든 값들이 같은 손실을 갖지만 손실 측도는 셀 단위로 다른 손실을 갖는 차이점이 있기 때문이다. 비용은 정확도 및 손실측도와 상이한 결과를 갖는데, 이는 원본 데이터에서의 손실을 계산하는 것이 아니라 일반화된 데이터 표의 익명성만을 고려하기 때문이다. 테이블 점수는 정보 이론적 관점에서 엔트로피를 사용하고 익명성 관련 수치로 나누어 줌으로써 데이터 왜곡 정도와 익명성 손실 정도를 동시에 고려한 측도이므로 좀 더 종합적인 의미를 담고 있다.

4.2. 측도들에 대한 특성 분석

아래의 <표 3>은 다양한 측도들에 대한 특성을 종합적으로 비교 분석한 결과이다.

<표 3> 측도들의 특성 비교

측도	k-익명성 알고리즘 방식	측도 대상	측도의 기준
height	상향식 (bottom-up)	속성	속성의 일반화 수
Prec	상향식 (bottom-up)	속성	속성의 일반화 비율
LM	상향식 (bottom-up)	값	속성 값들의 그룹화 비율
$C(GT)$	하향식 (top-down)	기록	데이터의 익명성
$TS(GT)$	하향식 (top-down)	기록	정보량 및 익명성

<표 3>에서 보는 바와 같이 각 측도들은 그것들의 다양한 특성 때문에 <표 2>에서와 같이 일관성을 갖지 않는다. 이는 각각 값 일반화 계급과 일반화된 데이터 표에서의 기록들의 분포에 따라 측도 값이 다르게 계산되기 때문이다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 개인정보 노출 위험을 최소화하고 데이터 유용성을 최대화하기 위해 k-익명성을 갖도록 변형된 데이터 표의 정확성을 측정하기 위한 다양한 측도들을 비교 분석하였다. 기존 측도들 중 높이, 정확도, 손실측도, 비용, 점수 등을 고려 대상으로 하였다. 이들을 통합적 비교 분석하기 위하여 우리는 가능한 균형 있고 공정한 시각으로 합리적인 비교분석을 실시하고자 노력하였다. 원본 데이터와 변형된 데이터가 주어졌을 때, 각 측도들 사이의 마이크로 데이터의 정확성에 대한 일관성과 개별성을 조사하고, 그 측도들의 특성에 따른 의미와 효율성을 비교분석하였다.

향후 좀 더 면밀한 연구를 통하여 각 측도들의 장점을 통합하고 단점을 보완할 수 있는 복합적인 측도를 개발하는 것이 남아있는 과제라고 생각한다.

[참고문헌]

[1] L. Sweeney, "k−anonymity : A model for protection privacy," International Journal of Uncertainty, Fuzziness and Knowledge-based systems, pp. 557-570, 2002.

[2] T.M. Truta, V. Bindu, "Privacy Protection: P-Sensitive K-Anonymity Property", Proceedings of the Workshop on Privacy Data Management, In Conjunction with 22th IEEE International Conference of Data Engineering (ICDE), Atlanta, Georgia, 2006.

[3] L. Sweeny, "Achieving k-anonymity privacy protection using Generalization and usppression." International Journal of Uncertainty, Fuzziness and Knowledge-based systems, pp.571-588, 2002.

[4] V. Iyengar, "Transforming data to satisfy privacy constraints", In Proc. of the Eight ACM SIGKDD Int Conf. on Knowledge Discovery and Data Mining , 279-288, 2002

[5] R. J. Bayardo, R. Agrawal, "Data Privacy through Optimal k-Anonymization," icde, pp. 217-228, 21st International Conference on Data Engineering (ICDE'05), 2005.

[6] B. Fung, K. Wang, P. Yu, "Top-Down Specialization for information and privacy preservation," in Proceedings of the 21st IEEE international Conference on Data Engineering. pp.205-216, 2005.