비식별 데이터의 유사성 보존에 관한 연구

강동현*, 오현석*, 용우석*, 이원석*
*연세대학교 컴퓨터과학과
e-mail: donghyun.kang@database.yonsei.ac.kr

A Study on the Preservation of Similarity of privated Data

Dong-Hyun Kang*, Hyun-Seok Oh*, Woo-Seok Yong*, Won-Seok Lee*
*Dept. of Computer Science, Yonsei University

요 약

비식별화 모델은 데이터 공유를 위한 모델로 원본데이터를 비식별화 변환 처리하여 개인정보를 보호함과 동시에 분석에 필요한 데이터를 외부에 제공하는 모델로 연구되어 왔다. 변환 방법으로는 삭제, 일반화, 범주화 기술 등이 주로 사용되며 변환 과정 중에는 재식별 가능성을 최소화하기 위해 k-익명성, l-다양성, t-근접성 혹은 differential privacy 등의 프라이버시 모델이 적용되고 있다. 하지만 변환된 비식별 데이터 세트는 필연적으로 원본 데이터 세트와 다른 값을 가지며 이는 결과적으로 최종 분석 결과에 영향을 주게 된다. 이를 위해 두 데이터 세트 간의 차이를 상이도(dissimilarity) 혹은 정보 손실율(information loss)이라는 지표로 측정 하고 있으며 본 지표는 비식별 데이터의 활용성을 평가 하는 데에 매우 중요한 역할을 한다. 본 연구에서는 비식별 데이터와 원본 데이터와 간의 차이를 도메인 기반의 절대적인 기준대비로 표현한 상이도 측정 방법을 제안하며, 그 유효성을 실데이터 기반의 실험을 통해 검증하였다.

1. 서론

최근 고도화된 분산 처리 시스템을 기반으로 하여 대용량 데이터의 수집 및 관리 기술과 분석 기술이 성숙해지고 있다. 이에 따라 빅데이터 기반의 데이터 마이닝과 기계 학습 모델들이 많은 주목을 받고 있으며, 학계는 물론 산업적으로도 다양한 활용 모델이 연구되고 있는 추세이다. 이러한 데이터 기반의 지식탐사 모델에서는 다양한 출처에서 수집된 데이터 간의 연계를 수행하고 이를 통한 새로운 지식 발견 및 학습 모델을 궁극적인 목표로 하고 있다. 하지만 각 기업이나 정부에 수집된 양질의 데이터에는 매우 직접적인 개인정보가 포함되어 있고, 원본 데이터 세트의 직접적인 공유는 개인정보 정보 유출에 대한 윤리적, 법적 문제 제기로 인해 현실적으로 불가한 상태이다. 이에 따라 데이터 세트를 공유하기 전 원본 데이터의 특정 속성들을 삭제, 범주화, 일반화 등의 방법으로 가공하여 개인정보를 모두 제거하는 비식별화 기법이 활용되고 있다 [1]. 더 나아가 그 과정 내에서 변환 후의 재식별 가능성을 최소화하기 위해 k-익명성, l-다양성, t-근접성 혹은 differential privacy 기반의 프라이버시 모델이 적용되는 것이 데이터 공유의 일반적인 절차이다 [5,4].

이렇게 변환된 비식별 데이터 세트는 필연적으로 원본 데이터 세트와 다른 값을 가지며 둘 사이에 차이는 상이도(dissimilarity) 혹은 정보 손실율(information loss)이라는 지표로 평가 되어왔다[2,3,4]. 하지만 기존의 상이도 평가 모델은 주어진 비식별 데이터 세트가 주어졌을 때, 각각의 비식별 레코드에 대응되는 원본 레코드의 속성값을 기준으로 상대적인 변화 폭이 얼마나 큰가를 계산하는 방식으로 측정되어 측정된 크기가 하나의 데이터 세트 내에서도 일관성을 유지하지 못한다는 문제를 안고 있다. 또한 기존 연구들에서는 연구목적으로 공개된 단순한 데이터를 기반으로 평가를 수행하여 실제 레코드에서의 그 효용성을 파악하기 힘들었다는 한계가 존재한다. 본 논문에서는 위의 한계를 보완한 새로운 도메인기반의 절대 크기를 갖는 상이도 측정 방법을 제안하며, 그 효용성을 실데이터 기반의 실험을 통해 검증한다.

본 논문은 다음과같은 구성으로 작성된다. 2절에서 관련 연구들에 대해 소개하며 3절에서 제안하고자 하는 유사도 산정 방법에 대해 소개한다. 마지막으로 4절에서는 새로이 모델링 된 유사도 지표에 대한 실데이터를 통한 유효성 검증을 소개하고 5절에서 그 결론과 함께 이후 추가적으로 필요한 연구에 대해 제안한다.

2. 관련 연구

상이도는 원본레코드 세트와 결과 레코드 세트 간의

차이를 수치적으로 측정한 지표이다. 생성된 비식별 레코드와 원본 레코드 간의 상이도를 기준 삼아 유통용 데이터의 활용성을 평가하고 더 나아가 레코드 수준에서 원본 레코드와 비식별 레코드 간의 재식별 가능성을 판단하는 간접적인 지표로도 사용될 수 있다.

일반적으로 결과 레코드와 원본 레코드 간의 상이도를 측정하는 방법은 (1)단일 속성을 수준에서 오차를 정의한 대한 접근방법과 (2)다차원 속성을 갖는 레코드 간의 거리를 기반으로 한 측정 방법으로 두 가지가 주로 사용되고 있다.

오차의 정의를 활용한 지표는 다시 절대오차, 상대 오차로 나뉜다. 절대 오차는 |원본 레코드 속성값 − 대표 레코드 속성값| 으로 표현되며 단순히 원본 레코드 속성값 대비 대표레코드 속성값의 크기를 오차로 표현하여 그 크기를 표현하는데 그친다. 상대 오차는 |(원본 레코드 속성값 − 비식별 레코드 속성값) / 원본 레코드 속성값| 으로 표현된어 원본 레코드 속성값 대비 변화된 크기를 비율로 나타내어 상대적인 변화양을 알 수 있는 지표로 사용이 된다. 이와 같이 오차의 정의는 단일 속성값 기준으로 측정되며 레코드 내의 속성 별 오차를 모두 구한 후 그 값을 평균내거나 총합하여 레코드 별 오차를 모두 구하고, 다시 한 번 모든 레코드의 오차를 평균내거나 총합하여 비식별 데이터 세트의 상이도로 정의하게 된다.

오차의 정의를 활용한 상이도는 [4]와 같이 상대 오차가 주로 사용된다. 본 방식의 문제점으로는 상대오차값은 절대적인 상한치가 정해지지 않으며, 원본레코드의 크기에 많은 영향을 받기 때문에 레코드의 상이함을 판단하기 위해서는 분석가에게 데이터에 대한 이해도를 요구한다. 예를 들어 비식별화된 2개의 결과레코드에 대해 절대 오차를 구할 때, 절대오차가 동일하게 10이더라도 원본 레코드 속성값이 2인 경우 상대오차 값은 5로 표현되고 원본레코드 속성값이 상대적으로 큰 10일 경우 상대오차 값이 1로 표현되어 도메인 내에서 절대적으로 비교하기는 힘들다는 문제가 있다.

두 번째로 원본 레코드와 비식별 레코드 간의 거리를 정의하여 상이도를 측정하는 방법이 사용된다. 일반적으로 다차원 레코드의 거리를 정의하는 방식은 다양하지만 상이도 측정 시에는 크게 두 접근법으로는 맨하탄 거리나 유클리드 거리를 이용하는 방법과 편차 제곱합 기반의 상이도 측정 방식[2,3]이 주로 사용된다. 하지만 본 방법은 레코드 세트의 도메인을 고려하지 않아 속성별로 상이도에 주는 영향력이 크게 차이날 수 있다는 단점이 있어 표준적인 측정 방법으로 활용하기 위하여 표준화가 필요하다.

살펴본 것처럼 기존의 상이도 계산 방법들은 데이터 세트 내의 속성별 값들에 따라 그 값이 판이하게 바뀌며 절대적인 범위가 주어지지 못한다. 이에 따라 분석가가 데이터에 대한 이해나 상이도 지표에 대한 이해를 잘 하지 못하고 있다면 비식별화 결과의 활용성을 판단하기 어렵다는 단점을 지니고 있다.

이를 해결하기 위해 본 연구에서는 상이도가 정해진 범위 [0,1] 내에서 표현되어져 데이터나 속성별 도메인에 상관없이 수치만으로도 그 활용성을 판단할 수 있는 지표를 제안하고자 한다.

3. 도메인 기반 상이도 측정 기법

제안하는 유사도 측정 방식은 상대오차와 비슷한 정의를 취하되 그 비율을 원본 레코드의 속성값이 아닌 전체 도메인의 크기 대비로 계산하여 오차의 크기가 항상 [0,1] 사이의 값을 갖게 한다. 따라서 전체 상이도를 구하는 프로세스는 다음과 같다. 가장 기본이 되는 비식별화 레코드의 각 속성 상이도를 정의하고, 이를 평균 내어 각각의 레코드 상이도를 정의한다. 이후 모든 레코드 상이도 값을 평균 내어 전체 비식별 데이터 세트의 상이도 값을 정의하는 방법을 취한다.

3.1 속성 상이도

다음과 같이 n개의 레코드, m개의 속성을 갖는 결과 레코드 세트에 대해 각각의 i번째 레코드의 j번 째 속성의 값을 Res_{ij}로 표기하고 이에 매핑 되는 원본 레코드 속성값을 Ori_{ij}로 표기할 때, 원본-결과 레코드 쌍에서 각각의 속성들에 대한 속성 상이도 $RCol_{ij}(0 \le RCol_{ij} \le 1, 1 \le i \le n,\ 1 \le j \le m)$ 는 속성의 타입에 따라 다음과 같이 계산한다.

수치형 속성 상이도 계산법은 원본레코드의 수치형 속성값을 특정 로직에 의하여 유사한 수치값으로 변환 시킨 속성들에 대해 사용한다. 이 때, 원본-결과 레코드의 속성값 차를 활용하여 두 레코드의 상이도를 정량적으로 측정할 수 있는데 원본-결과 레코드의 차이의 최댓값은 대상 속성의 원본 도메인 크기이다. 이를 활용하여 원본 도메인 크기 대비 원본-결과 레코드의 속성 값 차를 계산하여 속성상이도를 측정한다. 따라서 선택된 수치형 속성에 대해 원본 도메인의 최댓값, 최솟값을 각각 $\max(Ori_i)$, $\min(Ori_i)$로 표기할 때 수치형 속성의 속성 상이도는 다음과 같다.

$$RCol_{ij} = \frac{|Ori_{ij} - Res_{ij}|}{\max(Ori_i) - \min(Ori_i)}$$

수치형 속성 상이도 계산법은 문자열이나 날짜 등의 특정 속성들을 묶어서 범주화, 일반화 등의 비식별화를 취한 경우에 대해 사용될 수 있다. 결과 레코드는 변환 속성값은 계층 구조 상으로 하위 노드들을 포함하는 집합으로 볼 수 있게 된다. 이때 원본-결과 레코드의 해당 속성 값 상이도 비교는 집합 간의 오차 비교로 볼 수 있으므로 이때의 수식은 $count(Set(Res_{ij} \cup Ori) - Set(Res_{ij} \cap Ori))$ 로 나타낸다. 이때 일반화, 범주화 등의 비식별화 기법이 적용되는 경우에는 원본 레코드 속성값은 1개이고 이 때 원

본은 항상 결과 레코드의 속성값에 포함되므로 $count(Set(Res_{ij})-1)$로 나타낼 수 있다. 이를 해당 속성의 전체 도메인 집합과 원본 레코드 간의 오차 대비로 나타내어 상이도를 측정한다.

따라서 선택된 컬럼 Ori_i에 대해 원본 도메인의 distinct value의 count 값을 $count(distinct\ Ori_i)$로 표기하고 결과 레코드의 속성 값 Res_{ij}을 만든 집합을 $Set(Res_{ij})$로 표기할 때 범주형 속성의 속성 상이도는 다음과 같다.

$$RCol_{ij}=\begin{cases}\dfrac{count(Set(Res_{ij}-1))-1}{count(distinct\ Ori_j)} & ,Ori_{ij}\in Set(Res_{ij})\text{ 일 때}\\ 0 & ,Ori_{ij}\notin Set(Res_{ij})\text{ 일 때}\end{cases}$$

3.2 레코드 상이도

레코드 상이도는 특정 원본레코드와 결과레코드 쌍 사이의 유사성을 의미하며 결과 레코드가 가지고 있는 각각의 속성에 대해 원본 속성과의 속성 유사도를 계산하고, 그 값을 평균을 내어 계산한다. 따라서 레코드 유사도 $(RRec_i)$는 다음과 같이 정의된다.

$$RRec_i=\frac{\sum_{j=1}^{m}RCol_{ij}}{m}(0\le RRec_i\le 1)$$

3.3 테이블 상이도

테이블 유사도는 원본레코드 세트와 결과레코드 세트 사이의 상이도를 의미하며 결과 레코드 세트의 각각의 레코드에 대해 연결된 원본 레코드와의 레코드 상이도를 계산 하고, 그 값을 평균 내어 계산한다. 따라서 테이블 유사도$(RTab)$는 다음과 같이 정의된다.

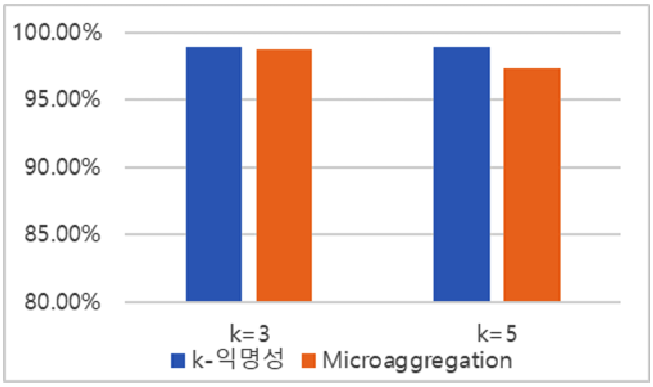
$$RTab=\frac{\sum_{i=1}^{n}RRec_i}{n}(0\le RTab\le 1)$$

4. 실증

본 논문에서는 국내 이동통신사가 보유한 개인정보 데이터를 대상으로 k-익명성과 마이크로어그리게이션(microaggregation) 프라이버시 모델로 비식별화된 데이터 세트의 활용성을 판단하기 위해 앞서 정의한 상이도를 계산하고 분석하였다.

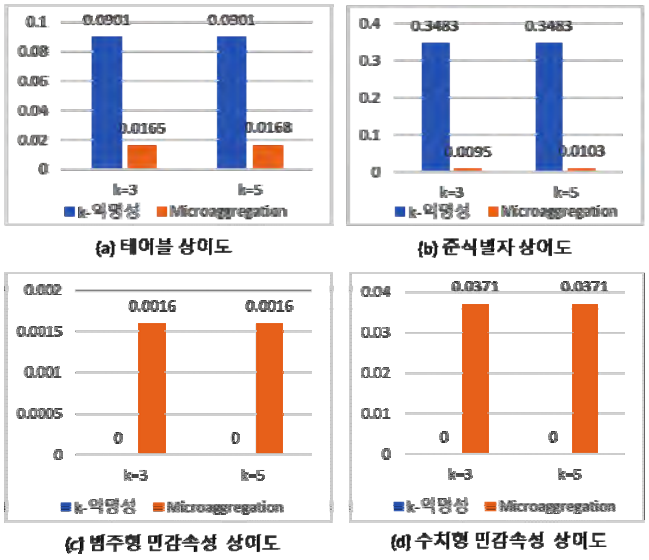
원본 데이터는 국내 이동통신사가 직접 수집한 신용도 데이터세트로 원본 레코드 수는 800만 건이며, 6개의 준식별자 속성, 7개의 명목형 민감속성, 11개의 수치형민감 속성을 가져 총 24개의 속성을 갖는 데이터 세트이다.
위의 동일한 데이터 세트에 대해 k-익명성 모델은 k값을 변환 시켜가며 준식별자 속성에 대해서만 일반화/범주화 비식별화를 수행하여 프라이버시 모델을 달성하였고, 마이크로어그리게이션 모델은 그룹사이즈를 k와 동일한 크기로 변환 시켜가며 준식별자 속성과 민감속성 모두에 대해 총계처리 비식별화를 수행하여 프라이버시 모델을 달성하였다.

실험에는 두 프라이버시 모델의 비식별화 결과에 대해 활용성 측면에서 두 가지 지표가 측정하였다. 첫 번째는 레코드의 잔존율이다. 두 프라이버시 모델은 변환 과정에서 재식별가능성이 존재하는 레코드는 완전히 삭제하여 결과 데이터 세트에 포함시키지 않는 과정이 존재하며, 이에 따라 원본 레코드 수보다 적은 수의 결과 레코드를 보유하고 있다. 이렇게 제거되는 레코드의 수가 많을수록 데이터 세트의 활용성은 떨어지게 되므로 원본레코드 수 대비 결과레코드 수를 비율로 나타내어 잔존율로 정의하고 각 결과 데이터 세트에 대해 잔존율을 평가하였다. (그림 1)은 이에 대한 실험 결과로 전반적으로 준식별자만을 대상으로 프라이버시 모델이 적용된 k-익명성 모델이 좋은 잔존율을 보임을 알 수 있었다. 상대적으로 민감속성까지 프라이버시 모델이 적용된 마이크로어그리게이션의 결과 세트는 미세하게 낮은 잔존율을 보이고 있다. 또한 k값이 커질수록 프라이버시의 엄격함이 커지므로 k값에 따라 잔존율이 감소하는 것도 확인할 수 있다.



(그림 1) k값 변화에 따른 프라이버시 모델의 잔존율

두 번째 지표는 앞서 정의한 상이도 지표로 완성된 비식별 데이터 세트와 원본 데이터 세트간의 차이를 수치화하여 비교하였다. (그림2-a)는 프라이버시 모델과 k(그룹 사이즈) 값의 변화에 따른 테이블 상이도를 보이고 있다. 전체 상이도는 마이크로어그리게이션 기법이 높은 상이도를 보이고 있음을 알 수 있다. 좀 더 자세한 상이도 차이는 (그림 2-b,c,d)를 통해 확인할 수 있는데, k-익명성 모델의 경우 민감속성에 대한 비식별화를 수행하지 않으므로 완전히 동일한 값으로 속성 상이도가 0이 됨을 확인할 수 있다. 그에 반해 준식별자 속성만으로 프라이버시 모델을 달성하여야 하므로 준식별자에 대한 일반화/범주화가 심하게 이뤄져 그 변화가 수치적으로 매우 크게 나타남을 확인 할 수 있었다. 따라서 평균적인 테이블 상이도 역시 커질 수밖에 없었다. 그에 반해 마이크로어그리게이션 모델에서는 준식별자, 범주형 민감속성, 수치형 민감속성에 대해 모두 비식별화 변환을 수행하였으므로 모든 부분에서 그 상이도가 다름을 확인할 수 있다. 그럼에도 불구하고 각각의 속성의 변화 폭은 작았기 때문에 전체적인 상



(그림 2) k값 변화에 따른 프라이버시 모델의 상이도

이도는 작게 표현됨을 확인 할 수 있었다. 또한 마이크로어그리게이션 모델에서는 그루핑 크기가 커짐에 따라 상이도가 미세하게 커지는 것을 확인할 수 있었다.

결과적으로는 레코드 활용성의 측면에서 마이크로어그리게이션 모델이 잔존율은 조금 떨어지지만 상이도가 훨씬 높게 측정되어 그 사용가치가 높다고 판단되어진다. 특히 준식별자 속성들은 성별, 거주지, 연령대 등의 속성으로 분석 시에도 중요한 역할을 하는 속성들이므로 그 왜곡의 정도가 훨씬 적은 마이크로어그리게이션 모델의 활용성이 더 높다고 보인다.

5. 결론

본 논문에서는 비식별화 변환 기법의 판단지표중 하나인 상이도에 대해서 기존의 상대적인 비율이나 단순 크기 비교로 이뤄졌던 한계를 극복하기 위해 데이터 세트의 도메인에 기반하여 [0,1] 스케일로 표현가능한 상이도 모델을 제안했다. 또한 과거 프라이버시 모델들의 평가가 제한된 수 와 속성을 가진 샘플데이터로 이뤄졌던 것들에 비해 본 논문에서는 실제 기업에서 수집된 대용량의 데이터를 기반으로 비식별화 모델들에 대한 평가를 수행하였고 실제 새로이 제시된 상이도 지표의 활용가능성을 확인하였다.

향후 연구에서는 수치형 속성 상이도 모델에서 아웃라이어의 영향을 최소화하는 방법에 대한 연구와 일반화, 범주화가 아닌 명목형 속성에 대해서도 적용가능한 상이도 모델에 대한 연구가 필요할 것으로 보인다.

Acknowledgment

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2017R1A2B4005344).

참고문헌

[1] Vatsalan, Dinusha, Peter Christen, and Vassilios S. Verykios. "A taxonomy of privacy-preserving record linkage techniques." Information Systems 38.6 (2013): 946-969.

[2] Rebollo-Monedero, David, et al. "k-Anonymous microaggregation with preservation of statistical dependence." Information Sciences 342 (2016): 1-23.

[3] Soria-Comas, Jordi, et al. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity." The VLDB Journal 23.5 (2014): 771-794.

[4] Sánchez, David, et al. "Utility-preserving differentially private data releases via individual ranking microaggregation." Information Fusion 30 (2016): 1-14.

[5] Garfinkel, Simson L. "De-identification of personal information." National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. IR-8053 (2015).