Asia-pacific Journal of Convergent Research Interchange
Vol.9, No.11, November 30 (2023), pp.23-33
http://dx.doi.org/10.47116/apjcri.2023.11.03

A Study on The Improvement of Information Loss Metrics in Real-Time Stream Data Anonymization

실시간 스트림 데이터 익명화에서 정보 손실 측도의 개선에 관한 연구

Ji Yeon Lee[1], Yong Wan Joo[2], He Young Yun[3], Soon Seok Kim[4]

이지연[1], 주용완[2], 윤혜영[3], 김순석[4]

[1] Professor, Department of AI Convergence Security, Halla University, Korea, jiyeon.lee@halla.ac.kr
[2] Professor, Industry-University Cooperation Foundation, Gangneung-Wonju National University, Korea, ywju@gwnu.ac.kr
[3] Professor, Department of Future Life Convergence, Halla University, Korea, sskim@halla.ac.kr
[4] Professor, Department of AI Convergence Security, Halla University, Korea, sskim@halla.ac.kr

Corresponding author: Soon Seok Kim

Abstract: Stream data refers to information collected in real time, such as crime report information, online sales transaction information, and information from patient monitoring devices in hospitals. This paper deals with the anonymization problem, which is the privacy issue of stream data. Usually, the main consideration in anonymization is how to process the data in such a way that it is secure and useful at the same time. Data usefulness refers to the quality of data and is measured by a so-called loss of information measure. In this paper, we review the current information loss metric in real-time stream data anonymization and propose a new metric that improves the disadvantages by applying Goldberger et al.'s scheme. The measure proposed by Goldberger et al. is characterized by considering the total equivalent class of the data table and the number of records in the equivalent class, which were not previously applied in the generalized dataset to which the k-anonymity model was applied. In the past, only one equivalent class within one cluster was assumed, so it was not included in the consideration. The reason is that records that do not satisfy k-anonymity within one cluster are allocated to other clusters that can be moved, that is, clustered with the next-rank quasi-identifier, or are deleted if there are no clusters that can be allocated. However, in terms of information loss, it is reasonable to assign them to other clusters because deleting those records leads to increased loss. Therefore, it is desirable to minimize information loss by allocating these records to another cluster that can be moved or, if there is no cluster to be assigned, allocating all of these records to a separate independent cluster instead of deleting them. The proposed idea focuses on the latter rather than the former. In this case, it is because several equivalent classes with quasi-identifiers can exist in an independent cluster. Therefore, by using the proposed metric of Goldberger et al., it is necessary to generalize to minimize the measured loss value after measuring the information loss by considering the equivalent class and the number of records in the equivalent class.

Keywords: Stream Data, Anonymization, Privacy, Information Loss

Received: July 03, 2023; 1st Review Result: August 08, 2023; 2nd Review Result: September 11, 2023
Accepted: October 25, 2023

요약: 스트림 데이터는 각종 범죄 신고 정보, 온라인 판매 거래 정보, 병원 내 환자 모니터링 기기 정보 등 실시간으로 수집되는 정보를 말한다. 본 논문은 스트림 데이터의 프라이버시 문제인 익명화 문제를 다룬다. 일반적으로 익명화의 주요 고려 사항은 안전한 동시에 유용한 방식으로 데이터를 처리하는 것이다. 데이터 유용성은 곧 데이터의 품질을 말하며, 소위 정보 손실 척도로 측정된다. 본 논문에서는 실시간 스트림 데이터 익명화에서 그동안의 정보 손실 지표들을 검토하고 단점을 개선한 새로운 지표를 제안한다. 우리는 실시간 스트림 데이터 익명화에서 기존 Goldberger 등이 제안한 기법을 적용하여 단점을 개선하였다. Goldberger 등이 제안한 측도는 k-익명성 모델이 적용되어 일반화된 데이터셋에서 기존에 적용되지 않았던 데이터 테이블의 전체 동질집합과 동질집합 내 레코드 수를 함께 고려한 것이 특징이다. 기존에는 하나의 클러스터 내 하나의 동질집합만을 가정하였기 때문에 고려사항에 포함되지 않았다. 그 이유는 하나의 클러스터 내에서 k-익명성을 만족하지 못하는 해당 레코드들은 이동 가능한 즉, 차순위 유사 식별자로 클러스터링이 가능한 다른 클러스터에 할당하거나 할당 가능한 클러스터가 없는 경우 삭제되기 때문이다. 그러나 정보 손실 측면에서 해당 레코드들의 삭제는 곧 손실 증가로 이어지기 때문에 다른 클러스터에 배정하는 것이 타당하다. 따라서 이들 레코드들을 이동 가능한 타 클러스터로 배정하거나 배정할 클러스터가 없을 경우 삭제 대신 이들 레코드 모두를 별도의 독립된 클러스터에 할당하여 정보 손실을 최소화하는 것이 바람직하다. 제안 아이디어는 전자의 경우보다는 후자에 주목한다. 이 경우 독립된 클러스터 내에는 유사 식별자들 갖는 여러 동질집합들이 존재할 수 있기 때문이다. 따라서 제안하는 Goldberger 등의 측도를 이용함으로서 동질집합과 동질집합 내 레코드 수를 감안하여 정보 손실을 측정한 후 손실 값이 최소화 되도록 값들을 일반화할 필요가 있다.

핵심어: 스트림 데이터, 익명화, 프라이버시, 정보 손실

1. 서론

실시간 스트림 데이터라 함은 실시간 범죄신고정보, 온라인 판매거래정보, 병원 내 환자 모니터링 장치 등 각종 현장 진료형 의료장비들로부터 실시간으로 수집되는 개인생체신호 등의 데이터를 말한다. 실시간 데이터들은 현재 다양한 응용들에서 활용되고 있다. 예를 들어 주민이 신고한 강력범죄(2020년 기준 우리나라 24,332건, 국가통계 포털) 등을 실시간으로 분석하여 짧은 시간에 범죄자나 용의자를 발견하거나 미래 범죄 혹은 재난 발생에 대한 예측을 가능하게 한다. 또한 캐나다의 Questrade.com과 같은 온라인 거래 회사는 고객이 발행한 수십만건의 온라인 입찰을 매일 수집하고 있는데, 이때 입찰자의 입찰정보는 매번 실시간 스트림 형태로 각 회사 주식에 대한 통계 정보를 찾아 실시간으로 게시하기 위하여 데이터 마이닝 알고리즘을 이용, 거래에서 생성된 데이터 스트림을 처리하고 있다. 그밖에 병원 내 환자 모니터링 장치 등 각종 현장진료형 의료장비들로부터 수집한 개인생체신호들을 HL7이나 FHIR 등 표준 메시징 형식으로 변환하여 실시간으로 질병 징후를 탐지 분석하는 등에 활용하고 있다. 이러한 여러 응용들에는 필수적으로 선결되어야 할 요건이 있는데 그것이 바로 예측이나 분석시 활용되는 원천 데이터에서의 프라이버시 보호 문제이다. 그동안 실시간 스트림 데이터에서의 익명화에 관한 문제는 여러 연구자들로부터 다양한 연구들[1-7]이 진행되어왔다.

한편 개인정보를 익명화할 경우 고려되어야 할 중요한 두 가지 사항이 있는데 그것이

바로 개인정보에 대한 안전성과 유용성이다. 즉, 개인정보는 마땅히 안전하게 보호되어야 하지만 그러면서도 활용이나 분석을 위해 쓸모있는 유용한 데이터를 생성해야 한다. 여기서 쓸모있고 유용하다는 의미는 익명화를 통하여 품질이 어느 정도 보증됨을 의미하는데 다른 말로 정보 손실(information loss)의 최소화를 의미한다. 그러나 항상 이 둘의 관계는 이율배반적인 성격을 갖는다. 안전성의 경우 소위 프라이버시 보호 모델이라 불리우는 k-익명성[8], l-다양성[9], t-근접성[10] 등의 측도들을 활용하여 확률적으로 얼마나 안전한가에 대해 측정하고 있다. 유용성의 경우는 프라이버시 보호 모델의 적용 여부에 따라 대개 두 가지로 분류가 되는데 먼저 프라이버시 보호 모델이 적용되지 않았을 경우는 대개 원본 데이터셋과 익명 처리된 데이터셋 간의 통계적 유사도나 상관관계 등을 기반으로 한 측도들을 활용하고 있다. 또한 그 반대의 경우는 k-익명성 등의 프라이버시 보호 모델이 적용된 경우로 정보 손실과 익명 처리된 데이터의 품질 관점에서 소위 데이터 품질 모델이라 불리우는 측도들이 이용되고 있다. 그러나 대개 익명화 시는 지난 2006년 행정안전부 등 정부부처합동으로 발간한 개인정보 비식별 조치 가이드라인[11]에서도 k-익명성 모델 적용을 필수 요건으로 언급할 만큼 프라이버시 보호 모델 적용이 필수적이다.

본 본문은 실시간 스트림 데이터의 익명화 시 데이터 유용성의 관점에서 특히 k-익명성의 프라이버시 보호 모델이 적용된 경우 정보 손실을 얼마나 줄일 것인가에 대한 그동안의 측도들을 고찰해보고 새로운 개선점을 제안하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 관련 연구로 기존 CASTLE 기법에서 사용한 Iyengar[12]의 손실 측도와 UBDSA에서 사용한 정보 손실 측도에 대해 살펴보고자 한다. 그리고 3장에서 이를 개선한 새로운 방법을 제안하고 이를 기존 UBDSA 기법과 비교하여 실험한 결과를 제시한 다음 4장을 끝으로 결론을 맺고자 한다.

2. 관련 연구

실시간 스트림 데이터의 익명화시 정보 손실 측도에 대한 제안은 지난 2011년 Cao 등이 발표한 CASTLE 알고리즘[3]에서 처음 소개되었다. CASTLE 알고리즘은 여러 제안된 방법들 가운데 클러스터링 기반 알고리즘에 속한다. 즉, 실시간 입력 값으로서 관계형 테이블에서의 각 레코드들을 가정하고 있는데 이들을 유사 식별자들을 갖는 레코드들끼리 모아서 각각의 클러스터에 저장, 분류한 다음 클러스터 내 각 레코드들을 k-익명성 모델을 적용하여 익명처리한 후 실시간으로 처리된 결과값을 클러스터 단위로 출력하고 있다. 이 과정에서 주어진 클러스터 내에 각 레코드들이 하나 이상 존재할 경우 새로 입력받은 레코드를 유사 식별자들을 갖도록 어느 클러스터에 할당할 것인가가 주요한 이슈가 된다. 왜냐하면 각 클러스터 내에 얼마나 유사한 식별자들끼리 클러스터링 되는가에 따라 데이터에 대한 유용성이 좌우되기 때문이다. 즉, 유사 식별자들끼리 모일수록 k-익명성 모델 적용시 데이터가 일반화가 최소한으로 수행되며 이울러 정보 손실이 가장 최소화될 수 있기 때문이다.

예를 들어 [그림 1]을 우편번호를 담은 일반화 계층 트리라 하자. 만일 클러스터 1에 우편번호 62201이라는 값이 저장되어 있고 클러스터 2에 35751이란 값이 저장되어 있을 때 새로운 입력값으로 62208이 입력되었다고 가정하자. 이때 이 값을 클러스터 1과 2중 어느 곳으로 할당할 것인가에 관한 문제이다. 이는 직관적으로도 클러스터 1에 할당하는 것이 당연하다. 왜냐하면 k-익명성 적용을 위해 일반화시 622**로 일반화될 수 있기

때문이다. 그렇지 않고 만일 62208 값을 클러스터 2에 할당할 경우 일반화는 357**이 아닌 *****로 일반화되어 유용성이 매우 떨어지기 때문이다.

[그림 1] 우편번호에 대한 일반화 계층 트리 예시

[Fig. 1] An Example of a Generalization Tree for Zip Codes

2.1 Iyengar[12]가 제안한 손실 측도

Cao 등은 CASTLE 기법에서 정보 손실 측도로 Iyengar[12]가 제안한 손실 측도(loss metric)을 사용하였다. 이 측도는 하나의 속성을 일반화하는 경우 몇 개의 값과 구별되지 않느냐 즉, 몇 개의 값이 하나의 속성으로 그룹화되는지에 기초하고 있다. 상기 [그림 1]에서 62201이 622**으로 일반화되면 62208과 구별되지 않고 35751이 357**으로 일반화되면 35733과 35791과 구별되지 않음으로 각 일반화가 다른 손실을 갖게 된다. 이러한 점에서 제안된 것이 바로 Iyengar[12]가 제안한 손실측도이다.

주어진 테이블의 전체 레코드 수를 I, 전체 속성 수를 J, J개의 속성에 대한 값 일반화 계층에서 해당 속성값의 하위 계층들에 있는 모든 자식 노드들의 수를 M, 노드 P에서 뿌리내린 모든 자식 노드들의 수를 MP이라 할 때, 손실측도의 계산은 다음과 같다.

$$\boldsymbol{LM(GT) = \sum_{j=1}^{J} \frac{\sum_{i=1}^{I} \frac{Mp-1}{M-1}}{I}} \quad (1)$$

그러나 이 기법은 그동안 제안된 높이나 정확도[13]를 이용한 측도에 비해 일반화 트리 계층구조에서 각 노드에 자식 노드들의 개수를 고려하여 보다 정확도를 개선한 장점은 있지만 전체 일반화 테이블 내에 동질집합의 개수와 레코드 수를 고려하지 않아 정확도가 떨어진다는 단점이 있다.

2.2 Ugur와 Osman[7]이 제안한 손실 측도

한편 Ugur와 Osman은 실시간 스트림 데이터에서의 익명화 방법으로 기존 CASTLE 기법을 개선하여 평균지연과 정보손실을 모두 고려한 UBDSA 기법[7]을 제안한 바 있다.

주어진 클러스터 내 전체 테이블을 T, 시간 j에서 입력된 레코들 t들의 집합을 S_j={t_1, t_2, ..., t_j}, 출력 레코드 집합으로 k-익명성 적용 대상 레코드 집합을 S'_j를 = {t_1, t_2, ..., t*}, $|S'_j| \leq |S_j|$라 하자. 이때 nbrleaves()는 범주형 속성데이터를 대상으로 해당 트리 노드 아래에 위치한 잎들의 총 개수이며 MAX()는 숫자형 속성데이터를 대상으로 수치값이 일반화되어 구간으로 일반화될 경우의 구간의 최대값을, MIN()은 최소값을 의미한다. $|D_j|$는 해당 속성 컬럼 j 내 레코드들의 총 개수이다. 이때 주어진 T에서의 정보 손실 IL(T)는 아래와 같이 계산된다.

$$IL(T) = \frac{\Sigma_{t \in T} InfoLoss(t, t^*)}{|T|} \quad (2)$$

$$AvgInfoLoss(S_j, S'_j) = \frac{\Sigma_{t'_i \in S_j} InfoLoss(t_i, t'_i)}{|S_j|} \quad (3)$$

$$InfoLoss(t_i, t'_i) = \sum_{j=1}^{j=n} AttINfoLoss(t_i.a_j, t'_i.a_j) \quad (4)$$

$$AttInfoLoss(t_i.a_j, t'_i.a_j) = \frac{nbrleaves(t_i.a_j) - nbrleaves(t'_i.a_j)}{|D_j|} \quad (5)$$

$$AttInfoLoss(t_i.a_j, t'_i.a_j) = \frac{MAX(t'_i.a_j) - MAX(t_i.a_j)}{|D_j|} \quad (6)$$

제안한 정보손실 측도는 속성 유형을 수치형과 범주형으로 나누어 달리 적용하되 일반화 계층 트리 전체가 아닌 해당 값이 있는 노드에서의 부분트리에서의 정보손실만을 고려하여 제안하였다. 이 경우 손실 측정에 대한 계산 성능은 향상될 수 있으나 부분트리만을 고려함으로서 정확도는 오히려 감소하는 결과가 초래되었으며 수치형 속성의 경우라도 일반화 계층트리에서의 k-익명성 모델 적용을 범주형과 동일하게 가정하고 있기 때문에 굳이 수치형과 범주형 속성을 달리 구분할 필요가 없다는 단점이 있다. 그럼에도 불구하고 데이터 유용성 측도에 있어 정보 손실뿐만 아니라 평균 또는 최대 지연시간을 고려하였다는 점에서는 CASTLE 기법보다 우수하다.

본 논문은 상기 제시한 CASTLE 기법과 UBDSA 기법에서의 정보 손실 측도를 개선한 새로운 방법을 제안하고자 한다.

3. 제안하는 정보 손실 측도

본 논문에서 제안하는 정보 손실 측도의 기본 아이디어는 CASTLE 기법에서 사용한 Iyengar[12]의 손실 측도를 기반으로 단점을 보완한 Goldberger 등[14]이 제안한 측도를 실시간 스트림에서의 익명처리 손실 측도로 새롭게 제안하는 것이다. Goldberger 등[14]이 제안한 측도는 k-익명성 모델이 적용되어 일반화된 데이터셋 즉, 클러스터 내에 기존에 적용되지 않았던 데이터 테이블의 전체 동질집합과 동질집합 내 레코드 수를 함께 고려한 것이 특징이다. 기존에는 하나의 클러스터 내 하나의 동질집합만을 가정하였기 때문에 고려사항에 포함되지 않았다. 그 이유로 하나의 클러스터 내에서 k-익명성을 만족하지 못하는 해당 레코드들은 이동 가능한 즉, 차순위 유사 식별자로 클러스터링이 가능한 다른 클러스터에 할당하거나 할당 가능한 클러스터가 없는 경우 삭제되기 때문이다. 그러나 정보 손실 측면에서 해당 레코드들의 삭제는 곧 손실 증가로 이어지기 때문에 다른 클러스터에 배정하는 것이 타당하다. 따라서 이들 레코드들을 이동 가능한 타 클러스터로 배정하거나 배정할 클러스터가 없을 경우 삭제 대신 이들 레코드 모두를 별도의 독립된 클러스터에 할당하여 정보 손실을 최소화하는 것이 바람직하다. 제안 아이디어는 전자의 경우보다는 후자에 주목한다. 이 경우 독립된 클러스터 내에는 유사 식별자들 갖는 여러 동질집합들이 존재할 수 있기 때문이다. 따라서 제안하는 Goldberger 등[14]의 측도를 이용함으로서 동질집합과 동질집합 내 레코드 수를 감안하여 정보

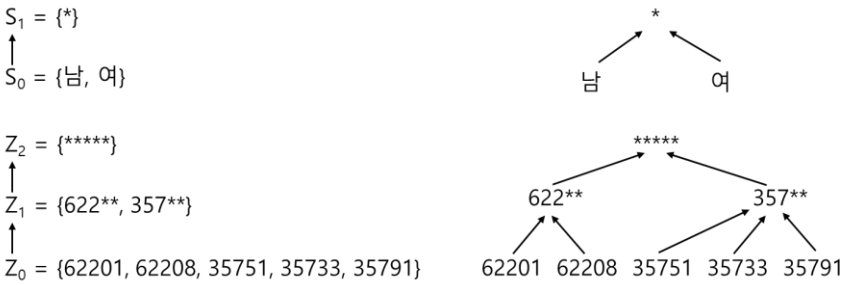
손실을 측정한 후 손실 값이 최소화 되도록 값들을 일반화할 필요가 있다.

한편 주어진 테이블의 전체 레코드 수를 I, 전체 속성 수를 J, J개의 속성에 대한 값 일반화 계층에서 해당 속성값의 하위 계층들에 있는 모든 자식 노드들의 수를 M, 노드 P에서 뿌리내린 모든 자식 노드들의 수를 MP, 클러스터 내 일반화된 동질집합의 개수를 t라하고 각 동질집합 내 레코드 수를 Cj 라 할 때, Goldberger 등[14]이 제안한 손실측도의 계산은 다음과 같다.

$$\boldsymbol{InfoLoss\ of\ Goldberger\ et.al}[\mathbf{14}] = \frac{\mathbf{1}}{\boldsymbol{I}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{t}}\left(\frac{\mathbf{1}}{\boldsymbol{J}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{J}}\frac{\boldsymbol{M_p}-\mathbf{1}}{\boldsymbol{M}-\mathbf{1}}\times\left|\boldsymbol{C_j}\right|\right) \quad (7)$$

3.1 제안하는 정보 손실 측도에 대한 분석

본 절에서는 먼저 예시를 통해 제안하는 손실 측도를 기존 Iyengar[12], Ugur와 Osman[7]이 제안한 손실 측도와 비교해 보고자 한다. 예시는 앞서 [그림 1]의 예시를 보다 확장하여 아래와 같이 2개의 속성을 갖는 일반화 트리를 가정한다.



[그림 2] 성별(남, 여) 및 우편번호에 대한 일반화 계층 트리 예시

[Fig. 2] An Example of a Generalization Tree for Gender(male, female) and Zip Codes

아울러 클러스터 내 저장된 일반화된 원본 테이블은 아래 [표 1]과 같이 5개의 레코드로 이루어진 것으로 가정한다.

[표 1] [그림 2]에 대한 일반화된 데이터셋 예시

[Table 1] An Example of Generalized Dataset for [Fig. 2]

S_0	Z_1
남	357**
남	357**
남	357**
여	357**
여	622**

[표 1]을 기준으로 Iyengar[12], Ugur와 Osman[7], 그리고 제안하는 정보 손실 측도를 계산해 보면 아래 [표 2]와 같다. [표 2]의 계산 결과 예시에서도 나타난 바와 같이 제안하는 방법은 기존 Iyengar[12], Ugur와 Osman[7]의 측도에 비해 정보손실 결과가 가장 낮음을 알수 있다.

[표 2] [표 1]의 일반화된 데이터셋에 대한 정보손실 값 비교

[Table 2] Comparison of Information Loss Values for the Generalized Dataset in [Table 1]

정보손실 측도	[표 1]에 대한 정보손실값	
	계산결과	계산식
Iyengar[12]	9/20	= 1/5(0+((3-1)/(5-1)+((3-1)/(5-1)+((3-1)/(5-1)+((3-1)/(5-1)+ ((2-1)/(5-1))
Ugur와 Osman[7]	18/20	= 1/2(0+(3+3+3)/5))
제안하는 방법	3/20	= 1/5(1/2(0+(3-1)/(5-1)×3))

제안하는 방법은 기존 Iyengar[12]에 비해 전체 일반화 테이블 내에 동질집합의 개수와 레코드 수를 고려하지 않아 정확도가 떨어진다는 단점을 개선하였으며 Ugur와 Osman[7]에 비해 부분트리가 아닌 전체 일반화 트리를 모두 고려함으로서 정확도를 개선하였다. 특히 제안 방법은 실시간 스트림 데이터의 익명화에 있어 기존 방법에 비해 아래와 같은 장점이 있다. 즉, 입력 버퍼에 실시간으로 수집되는 각각의 레코드들을 각각의 클러스터에 배정한 다음 각각의 클러스터별로 입력된 레코들을 일반화할 경우 갖게 되는 정보손실들을 보다 정확하게 측정해야하는데 이때 Iyengar[12]와 Ugur와 Osman[7]에 비해 보다 정확도 있는 계산 결과를 도출하여 가장 최적의 클러스터링을 수행할 수 있다.

[표 3] 데이터 세트 ADULT[10]에서 이용된 14개의 속성

[Table 3] 14 Attributes Used in the Dataset ADULT[15]

ADULT 속성	**도메인 사이즈**
Age	100
Work class	8
Fnlwgt	1500000
Education	16
Education-num	16
Marital-Status	7
Occupation	14
Relationship	6
Race	5
Gender	2
Capital-gain	100000
Capital-loss	10000
Hours-per-week	100
Native-country	41

3.2 실험 결과

우리는 제안하는 방법을 실제 실험을 통해 기존 Ugur & Osman[7]의 방법과 비교하여 그 성능을 측정하였다. 알고리즘은 Python으로 구현되었으며, 실험은 윈도우즈 10 환경에서 64GB 메모리가 있는 12th Gen Intel(R) Core(TM) i9-12900KS 3.40GHz 개인용 컴퓨터에서

수행되었다. 실험에 사용한 데이터 세트는 기존 Ugur & Osman[7]에서 사용된 것과 동일하게 UCI machine learning repository에서 제공하는 Census Income으로 알려진 ADULT[15]를 사용하였다. 데이터 셋 ADULT[15]는 데이터 개인정보보호 연구에 널리 사용되는 데이터 세트로 총 32,561개의 레코드와 14개의 속성([표 3] 참조)으로 이루어져 있다. [표 3]의 속성 중 3가지(Education, Occupation, Native-country) 속성은 범주형 속성으로 k-익명성 모델 적용을 위하여 일반화 트리로 구성하였다. 한편 트랜잭션 레코드들을 실시간으로 입력 받기 위한 데이터 수신은 UDP(User Datagram Protocol)을 이용하여 구현하였다. 데이터 수신 기능은 일반화 클러스터를 객체화한 이후에 호출된다. 데이터 수신 기능이 호출되면 일반화 클러스터가 객체화되면서 입력된 설정값에 따라 최대 지연시간 만큼 반복되는 반복 루프가 시작되고, 이 반복 루프 안에서 데이터 수신을 대기한다. 데이터가 수신되면 일반화 클러스터에 할당하다가 입력된 데이터의 수가 최대 지연시간에 도달하면, 반복 루프를 벗어나 k-anonymity를 만족하기 위한 일반화 과정으로 진입한다. 일반화가 끝나면 일반화 데이터의 종류 수를 기준으로 판단해서 최대 지연시간을 조정하고, 일반화가 완료된 데이터를 출력한 뒤 다시 데이터 수신 대기 상태로 들어간다. 만일 모든 데이터가 입력되었다는 신호가 통신을 통해 전해지면 모든 처리가 종료된다.

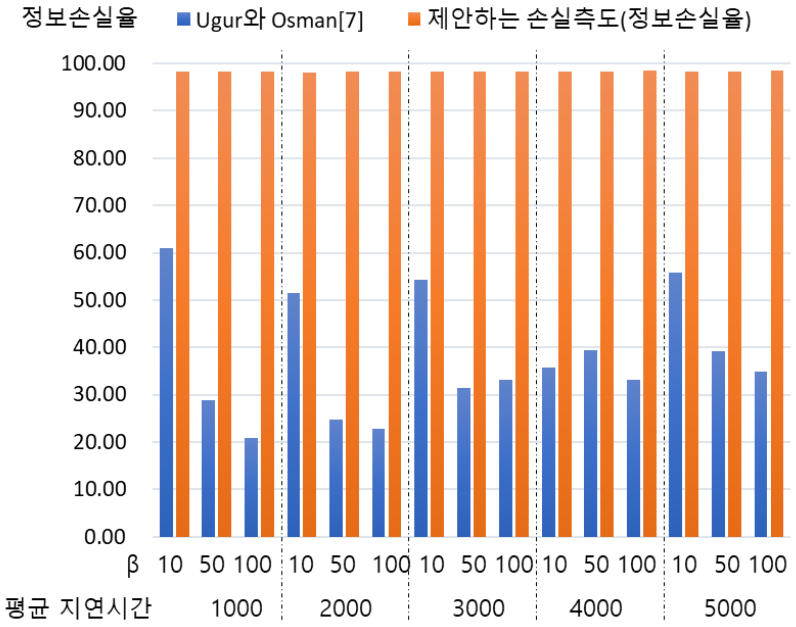
기존 Ugur & Osman[7]에서 제안된 방법과 우리의 제안 알고리즘에 대한 정보 손실에 대한 실험 결과는 [그림 3], [그림 4]와 같다. 제안하는 정보 손실 측도 계산을 위한 알고리즘의 모델 크기는 할당된 익명화 클러스터의 최대 개수(이를 각각 β와 μ라 부른다))에 의해 결정된다. 우리는 기존 [7]에서의 방법과 동일하게 β와 μ를 각각 낮음(β = μ = 10), 보통(β = μ = 50) 및 높음(β = μ = 100)으로 설정한 다음, k-익명성 모델에서의 각 k값에 대해 정보손실을 측정하였다. [그림 3]에서 볼 때 k값이 3일 경우, β가 10, 지연시간이 1,000일 때 최소 37.27%에서 β가 100, 지연시간이 1,000일 때 최대 77.58% 만큼 정보 손실의 차이가 발생하는 것을 확인하였다. 그리고 [그림 4]에서 볼 때 k값이 20일 경우에는 β가 10, 지연시간이 2,000일 때 최소 18.31%에서 β가 100, 지연시간이 2,000일 때 최대 63.99% 만큼 정보 손실의 차이가 발생하는 것을 확인하였다.

4. 결론

본 논문에서는 실시간 스트림 데이터의 익명화에서 그 성능을 측정하기 위한 주요 고려 요소 중 하나인 유용성 즉, 정보 손실 측도에 대해 기존 Cao 등이 제안한 Castle 기법[3]과 Ugur와 Osman이 제안한 UBDSA 기법[7]을 비교하여 이들의 단점을 개선한 새로운 측도를 제안하였다.

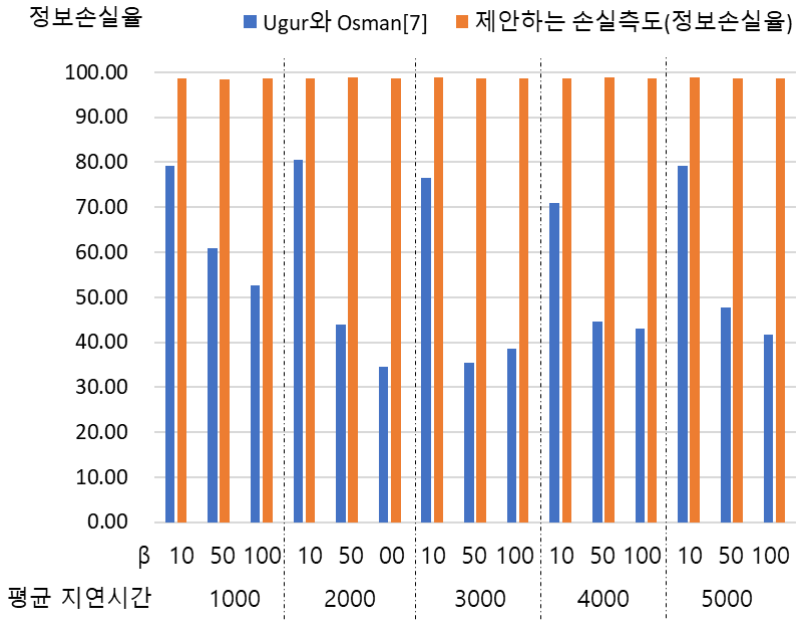
제안하는 방법은 기존 Iyengar[12]에 비해 전체 일반화 테이블 내에 동질집합의 개수와 레코드 수를 고려하지 않아 정확도가 떨어진다는 단점을 개선하였으며 Ugur와 Osman[7]에 비해 부분트리가 아닌 전체 일반화 트리를 모두 고려함으로서 정확도를 개선하였다. 특히 제안 방법은 실시간 스트림 데이터의 익명화에 있어 입력 버퍼에 실시간으로 수집되는 각각의 레코드들을 각각의 클러스터에 배정한 다음 각각의 클러스터별로 입력된 레코들을 일반화할 경우 갖게 되는 정보손실들을 보다 정확하게 측정해야하는데, 이때 Iyengar[12]와 Ugur와 Osman[7]에 비해 보다 정확도 있는 계산 결과를 도출하여 가장 최적의 클러스터링을 수행할 수 있다. 또한 제안한 방법을 Ugur와 Osman[7]의 방법과 비교하여 실험한 결과 k값을 3으로 할 경우 최소 최소 37.27% ~ 최대

77.58% 만큼, 그리고 k값을 20으로 하였을 경우 최소 18.31% ~ 최대 63.99% 만큼 정보 손실의 차이가 발생하는 것을 확인할 수 있었다. 향후 연구 방향으로는 실시간 스트림 데이터 익명화에서 안전성을 보다 개선하기 위한 새로운 방법들(예를 들어, k-익명성 기법에 더하여 l-다양성 기법 등을 적용하는 등)에 대해 제안해 보고자 한다.



[그림 3] Ugur & Osman[7]과 제안 방법과의 정보 손실 측도(정보 손실율) 비교(k=3)

[Fig. 3] Comparison of Information Loss Metric (Information Loss Rate) with Ugur & Osman [7] and the Proposed Method(k=3)



[그림 4] Ugur & Osman[7]과 제안 방법과의 정보 손실 측도(정보 손실율) 비교(k=20)

[Fig. 4] Comparison of Information Loss Metric (Information Loss Rate) with Ugur & Osman [7] and the Proposed Method(k=20)

5. 감사의 글

이 논문은 2023년도 정부(개인정보보호위원회)의 재원으로 한국인터넷진흥원의 지원을 받아 수행된 연구임 (No. 1781000005, 반정형 트랜잭션 및 실시간 수집 정형 데이터에서의 개인정보 가명,익명처리 기술 개발)

이 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화 혁신인재양성 사업의 연구결과로 수행되었음 (IITP-2023-RS-2023-00260267)

References

[1] W. Wang, J. Li, C. Ai, Y. Li, Privacy protection on sliding window of data streams, International conference on collaborative computing: networking, applications and worksharing, IEEE, (2007)
DOI: https://doi.org/10.1109/COLCOM.2007.4553832

[2] J. Li, B. C. Ooi, W. Wang, Anonymizing streaming data for privacy protection, IEEE 24th International conference on data engineering, IEEE, pp.1367-1369, (2008)
DOI: https://doi.org/10.1109/ICDE.2008.4497558

[3] J. Cao, B. Carminati, E. Ferrari, K. L.Tan, Castle: continuously anonymizing data streams, IEEE Transactions on Dependable and Secure Computing, (2011), Vol.8, No.3, pp.337-352.
DOI: https://doi.org/10.1109/TDSC.2009.47

[4] H. Zakerzadeh, S. L. Osborn, Faanst: fast anonymizing algorithm for numerical streaming data, Data privacy management and autonomous spontaneous security, Springer, pp.36-50, (2011)
DOI: https://doi.org/10.1007/978-3-642-19348-4_4

[5] K. Guo, Q. Zhang, Fast clustering-based anonymization approaches with time constraints for data streams, Knowledge-Based Systems, (2013), Vol.46, pp.95-108.
DOI: https://doi.org/10.1016/j.knosys.2013.03.007

[6] S. Kim, M. K. Sung, Y. D. Chung, A framework to preserve the privacy of electronic health data streams, Journal of Biomedical Informatics, (2014), Vol.50, pp.95-106.
DOI: https://doi.org/10.1016/j.jbi.2014.03.015

[7] Ugur Sopaoglu, Osman Abul, A utility based approach for data stream anonymization, Journal of Intelligent Information Systems, (2020), Vol.54, pp.605-631.
DOI: https://doi.org/10.1007/s10844-019-00577-6

[8] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, (2002), Vol.10, No.5, pp.557-570.
DOI: https://doi.org/10.1142/S0218488502001648

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), (2007), Vol.1, No.1, pp.3-14.
DOI: https://doi.org/10.1145/1217299.1217302

[10] N. Li, T. Li, S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity, 2007 IEEE 23rd International Conference on Data Engineering, IEEE, pp.106-115, (2007)
DOI: https://doi.org/10.1109/ICDE.2007.367856

[11] Joint Government Ministries, Guidelines for De-identification of Personal Information, (2012)
Available from: https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000827059&fileSn=0

[12] Vijay S. Iyengar, Transforming data to satisfy privacy constraints, Proceedings of the Eight ACM SIGKDD international Confonference on Knowledge Discovery and Data Mining, pp.279-288, (2002)
DOI: https://doi.org/10.1145/775047.775089

[13] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, (2002), Vol.10, No.5, pp.571-588.
DOI: https://doi.org/10.1142/S021848850200165X

[14] Jacob Goldberger, Tamir Tassa, Efficient Anonymizations with Enhanced Utility, IEEE Transactions on Data Privacy, (2010), Vol.3, pp.149-175.
DOI: https://doi.org/10.1109/ICDMW.2009.15

[15] https://archive.ics.uci.edu/dataset/2/adult, Aug 01 (2023)