(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0144350
(43) 공개일자 2022년10월26일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *G06F 16/25* (2019.01)
(52) CPC특허분류
G06F 21/6245 (2013.01)
G06F 16/254 (2019.01)
(21) 출원번호 10-2022-0131693(분할)
(22) 출원일자 2022년10월13일
심사청구일자 2022년10월13일
(62) 원출원 특허 10-2021-0003251
원출원일자 2021년01월11일
심사청구일자 2021년01월11일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
박유랑
서울시 송파구 잠실4동 파크리오아파트 310동 103호
성민동
서울시 서대문구 이화여대길 50-12
(74) 대리인
특허법인비엘티

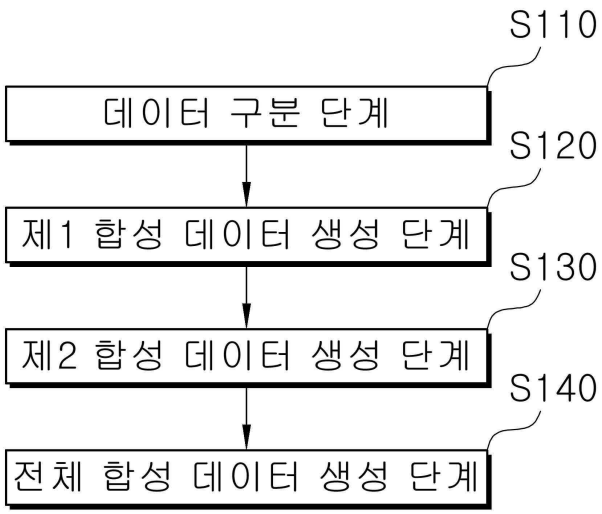
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용하여 합성 데이터를 생성하는 방법 및 장치**

(57) 요 약

차등 프라이버시를 이용한 합성 데이터 생성 방법 및 장치가 제공된다. 상기 방법은, 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분하는 데이터 구분 단계; 상기 연속형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계; 상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화 (post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계; 및 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계를 포함한다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

서버에 의해 수행되는 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용하여 합성 데이터를 생성하는 방법에 있어서,

상기 원본 데이터에 포함된 이산적 값이 아닌 값을 갖는 연속형 데이터에 대해 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계;

상기 원본 데이터에 포함된 이산적 값을 갖는 범주형 데이터에 대해 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계; 및

상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계를 포함하고,

상기 제1 합성 데이터 및 상기 제2 합성 데이터는,

상기 원본 데이터를 소유하는 엔티티에 의해 수행되는 것을 특징으로 하는, 방법.

청구항 2

제1항에 있어서,

상기 제1 합성 데이터 및 상기 제2 합성 데이터는,

상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 생성되는 것을 특징으로 하는, 방법.

청구항 3

제2항에 있어서,

상기 일정 값은 75 % 이상인 값을 갖고,

상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 갖는 것을 특징으로 하는, 방법.

청구항 4

제1항에 있어서,

상기 제1 합성 데이터 생성 단계는,

상기 연속형 데이터에 대해 경계 라플라시안 기법(bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하고,

상기 경계 라플라시안 기법은,

-1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함하는 것을 특징으로 하는, 방법.

청구항 5

제4항에 있어서,

상기 제2 합성 데이터 생성 단계는,

상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화(post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하고,

상기 후-처리 이산화 기법은,

상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함하는 것을 특징으로 하는, 방법.

청구항 6

제5항에 있어서,

상기 확률적으로 이산화하는 기법은,

베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함하는 것을 특징으로 하는, 방법.

청구항 7

원본 데이터를 저장하는 데이터 베이스; 및

상기 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용하여 합성 데이터를 생성하도록 설정된 프로세서를 포함하고,

상기 프로세서는,

상기 원본 데이터에 포함된 이산적 값이 아닌 값을 갖는 연속형 데이터에 대해 제1 합성 데이터를 생성하고,

상기 원본 데이터에 포함된 이산적 값을 갖는 주형 데이터에 대해 제2 합성 데이터를 생성하고,

상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하되,

상기 제1 합성 데이터 및 상기 제2 합성 데이터는,

상기 원본 데이터를 소유하는 엔티티에 의해 수행되는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 8

제7항에 있어서,

상기 프로세서는,

상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 생성하는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 9

제8항에 있어서,

상기 일정 값은 75 % 이상인 값을 갖고,

상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 갖는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 10

제7항에 있어서,

상기 프로세서는,

상기 연속형 데이터에 대해 경계 라플라시안 기법(bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하고,

상기 경계 라플라시안 기법은,

-1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함하는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 11

제10항에 있어서,

상기 프로세서는,

상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화(post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하고,

상기 후-처리 이산화 기법은,

상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함하는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 12

제11항에 있어서,

상기 확률적으로 이산화하는 기법은,

베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함하는 것을 특징으로 하는, 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용한 합성 데이터 생성 장치.

청구항 13

컴퓨터와 결합하여, 원본 데이터에 포함된 이산적 값이 아닌 값을 갖는 연속형 데이터에 대해 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계;

상기 원본 데이터에 포함된 이산적 값을 갖는 범주형 데이터에 대해 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계; 및

상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계를 포함하고,

상기 제1 합성 데이터 및 상기 제2 합성 데이터는,

상기 원본 데이터를 소유하는 엔티티에 의해 수행되는 것을 실행시키기 위하여 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

청구항 14

제13항에 있어서,

상기 제1 합성 데이터 및 상기 제2 합성 데이터는,

상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 생성되는 것을 특징으로 하는, 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

청구항 15

제14항에 있어서,

상기 일정 값은 75 % 이상인 값을 갖고,

상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 갖는 것을 특징으로 하는, 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

청구항 16

제13항에 있어서,

상기 제1 합성 데이터 생성 단계는,

상기 연속형 데이터에 대해 경계 라플라시안 기법(bounded Laplacian method)을 이용하여 제1 합성 데이터를 생

성하고,

상기 경계 라플라시안 기법은,

-1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함하는 것을 특징으로 하는, 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

청구항 17

제16항에 있어서,

상기 제2 합성 데이터 생성 단계는,

상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화(post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하고,

상기 후-처리 이산화 기법은,

상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함하는 것을 특징으로 하는, 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

청구항 18

제17항에 있어서,

상기 확률적으로 이산화하는 기법은,

베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함하는 것을 특징으로 하는, 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 원본 데이터에 포함된 연속형 데이터와 범주형 데이터를 이용하여 합성 데이터를 생성하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 빅데이터 및 인공지능 (artificial intelligence; AI) 학습 과정에서 개인 정보 보호의 중요성이 부각되며, 개인 정보 보호 기법으로써 차등 프라이버시 (differential privacy) 기법에 많은 관심이 쏠리고 있다.

[0003] 차등 프라이버시 기법은 특정 개인의 개인 정보가 노출되는 위험을 최소화하며 해당 데이터를 유의미하게 활용 가능하도록 데이터를 변형/구성하는 것을 목적으로 한다. 이를 위해, 차등 프라이버시 기법은 특정 데이터 세트에 임의의 노이즈를 삽입함으로써 개인 정보가 제3자에게 노출되지 않도록 보호하는 것을 특징으로 한다.

[0004] 다만, 삽입되는 노이즈는 학습용 데이터로부터 구체적인 개인 정보가 노출되는 것을 방지하는 역할을 하기도 하지만, 노이즈의 정도에 따라 해당 데이터를 활용하는 인공지능 모델의 성능을 저하시키는 문제점이 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 등록특허공보 제10-1935528호, 2019.01.04

발명의 내용

해결하려는 과제

[0006] 본 발명이 해결하고자 하는 과제는 차등 프라이버시를 이용해 데이터를 충분히 익명화하면서 동시에 데이터의

유용성을 유지시킬 수 있는 합성 데이터 생성 방법 및 장치를 제공하는 것이다.

[0007] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0008] 상술한 과제를 해결하기 위한 본 발명의 일 면에 따른 차등 프라이버시를 이용한 합성 데이터를 생성하는 방법은, 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분하는 데이터 구분 단계; 상기 연속형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계; 상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화 (post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계; 및 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계를 포함할 수 있다.

[0009] 본 발명에 있어, 상기 제1 합성 데이터 및 상기 제2 합성 데이터는, 상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 생성될 수 있다.

[0010] 이때, 상기 일정 값은 75 % 이상인 값을 가질 수 있다. 또한, 상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 가질 수 있다.

[0011] 본 발명에 있어, 상기 경계 라플라시안 기법은, -1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함할 수 있다.

[0012] 본 발명에 있어, 상기 후-처리 이산화 기법은, 상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함할 수 있다.

[0013] 이때, 상기 확률적으로 이산화하는 기법은, 베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함할 수 있다.

[0014] 본 발명에 있어, 상기 차등 프라이버시를 이용한 상기 합성 데이터를 생성하는 방법은 상기 원본 데이터를 소유하는 엔티티에 의해 수행될 수 있다.

[0015] 상술한 과제를 해결하기 위한 본 발명의 다른 면에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는, 원본 데이터를 저장하는 데이터 베이스; 및 상기 원본 데이터로부터 차등 프라이버시를 이용한 합성 데이터를 생성하도록 설정된 프로세서를 포함할 수 있다. 이때, 상기 프로세서는, 상기 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분하고, 상기 연속형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하고, 상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화 (post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하고, 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하도록 설정될 수 있다.

[0016] 본 발명에 있어, 상기 프로세서는, 상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 생성하도록 설정될 수 있다.

[0017] 본 발명에 있어, 상기 제1 합성 데이터 및 상기 제2 합성 데이터는, 상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 생성될 수 있다.

[0018] 이때, 상기 일정 값은 75 % 이상인 값을 가질 수 있다. 또한, 상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 가질 수 있다.

[0019] 본 발명에 있어, 상기 경계 라플라시안 기법은, -1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함할 수 있다.

[0020] 본 발명에 있어, 상기 후-처리 이산화 기법은, 상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함할 수 있다.

[0021] 이때, 상기 확률적으로 이산화하는 기법은, 베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함할 수 있다.

[0022] 상술한 과제를 해결하기 위한 본 발명의 또 다른 면에 따른 컴퓨터 판독가능 기록매체에 저장된 컴퓨터 프로그램은, 컴퓨터와 결합하여, 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분하는 데이터 구분 단계; 상기 연속형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계; 상기 범주형 데이터에 대해 상기 경계 라플라시안 기법 및 후-처리 이산화 (post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계; 및 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계를 실행시키도록 설정될 수 있다.

[0023] 본 발명에 있어, 상기 제1 합성 데이터 및 상기 제2 합성 데이터는, 상기 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값을 이용하여 생성될 수 있다.

[0024] 이때, 상기 일정 값은 75 % 이상인 값을 가질 수 있다. 또한, 상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 가질 수 있다.

[0025] 본 발명에 있어, 상기 경계 라플라시안 기법은, -1 내지 1 범위로 모든 변수들을 정규화(normalization)하는 기법을 포함할 수 있다.

[0026] 본 발명에 있어, 상기 후-처리 이산화 기법은, 상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화하는 기법을 포함할 수 있다.

[0027] 이때, 상기 확률적으로 이산화하는 기법은, 베르누이 분포 함수에 따른 베르누이 확률에 기초하여 이산화하는 기법을 포함할 수 있다.

[0028] 본 발명의 기타 구체적인 사항들은 상세한 설명 및 도면들에 포함되어 있다.

발명의 효과

[0029] 본 발명에 따르면, 익명화를 위한 노이즈는 범위를 벗어난 (out-of-bounds) 데이터가 생성되지 않도록 하며, 생성된 합성 데이터는 익명성 및 데이터의 유용성 모두를 만족시킬 수 있는 효과가 있다.

[0030] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0031] 도 1은 본 발명의 일 예에 따른 차등 프라이버시를 이용한 합성 데이터 생성 방법을 나타낸 흐름도이다.

도 2는 본 발명에 따라 로컬 차등 프라이버시를 이용한 합성 데이터 생성 방법을 간단히 나타낸 도면이다.

도 3a 내지 도 3c는 본 발명에 따른 경계 라플라시안 기법의 성능을 나타내는 그래프이다.

도 4a 및 도 4b는 본 발명에 따른 엡실론 값 및 데이터 교란 정도의 관계를 간단히 나타낸 그래프이다.

도 5는 본 발명에 따른 엡실론 값에 대한 다른 기계 학습 모델 간의 분류 정확도를 나타낸 그래프이다.

도 6은 본 발명의 다른 예에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0032] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0033] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다

(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0034] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0035] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0036] 본 발명에 대해 상세히 설명하기에 앞서, 본 발명이 적용 가능한 기술의 특징에 대해 상세히 설명한다.

[0037] 빅 데이터는 여러 분야 (예: 의료 분야, 영상 처리 분야, 자연어 해석 분야 등)의 혁신을 위한 핵심 요소 기술로 평가되고 있다. 데이터 자체는 대부분 무용지물이지만 기계 학습 (machine learning; ML)과 같은 알고리즘을 적용하면 이러한 데이터 대부분은 유의미하게 활용될 수 있다. ML 알고리즘의 특성상, 규칙 기반 시스템과 달리 데이터 중심적이며 (data-driven) 대량의 데이터를 필요로 한다. 또한 기존의 ML 접근 방식에 따른 학습 시스템은 중앙 집중식 데이터를 필요로 한다. 이러한 많은 양의 데이터를 획득하여 강력한 ML 모델을 구축하기 위해, 서로 다른 조직 간 데이터 교환은 반드시 필요하게 된다.

[0038] 그러나, 서로 다른 당사자들 간 데이터 교환은 프라이버시 문제를 야기하고, 이에 따라 대기업들이 프라이버시 문제를 위반하는 것에 대한 우려 또한 증가하고 있다. 특히, 대부분 민감한 정보를 포함하는 의료 데이터는 제3자와 공유할 때 적절히 보호될 필요가 있다. 이에 따라, 유럽연합의 GDPR (General Data Protection Regulation) 및 미국의 HIPAA (Unitied State's Health Insurance Portabiliy and Accountability Act of 1996)은 이러한 문제점을 인식하고 사용자의 프라이버시를 강화할 것을 요구한다.

[0039] 의료 데이터는 민감한 속성뿐만 아니라 다양한 속성도 가질 수 있다. 예를 들어, 혈청 포도당 레벨 (serum glucose level)은 연속적인 값인 반면 의료 기록 (medical history)은 일반적으로 범주형 값이다. 또한, 상기 의료 기록은 다중 모드 값 (multi-modal values)을 포함할 수 있다: 일부 테스트 결과는 혈액 테스트로부터 획득될 수 있고, 다른 테스트 결과는 방사선 검사 및 신체 검사 테스트로부터 획득될 수 있다.

[0040] 따라서, 프라이버시 정보 (예: 의료 데이터 등)는 무분별하게 제3자에게 공유되어서는 안된다. 이를 위해, 제3자 사용자에게 개인 정보가 제공되기 전에 개인 정보를 보호하기 위한 기술로써 익명화 (anonmization) 기술이 필요하다. 이때, 익명화 이후 프라이버시 위험을 식별하기 위한 방법으로써 다음의 세가지 주요 척도에 기반한 평가가 수행될 수 있다: k-anonymity, l-diveristy, t-closeness. ARX와 같은 탈-식별 도구 (De-identification tool)는 특징 일반화 (feature generalization), 기록들의 압축 (suppresion of records)를 통해 완벽한 개인 정보 보호를 제공할 수 있다. 차등 프라이버시는 데이터 프라이버시에 대한 또 다른 접근 방식으로, 의미론적 모델 (semantic model)이다. 통사적 익명성 (syntactice anonymity)에 비해, 차등 프라이버시 방법은 보다 적은 도메인 지식을 요구하고, 본질적으로, 도메인 지식과 결합된 연계 공격 (linkage attack)에 강력한 특징을 갖는다.

[0041] 이러한 특징들에 기초하여, 본 발명에서는 합성 데이터 (synthetic data)의 구현 가능성 (feasibility), 데이터 프라이버시 및 활용도 (utility) 간 균형 (balance) 등을 고려하여 차등 프라이버시 기법을 위한 합성 데이터를 생성하는 방법에 대해 상세히 설명한다.

[0042] 이하의 도 1 내지 도 6에서 설명되는 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 방법은 서버와 같은 컴퓨터 장치를 통해 모든 동작이 수행될 수 있다.

[0043] 도 1은 본 발명의 일 예에 따른 차등 프라이버시를 이용한 합성 데이터 생성 방법을 나타낸 흐름도이다.

[0044] 도 1에 도시된 바와 같이, 본 발명에 따라 차등 프라이버시를 이용한 합성 데이터를 생성하는 방법은, 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분하는 데이터 구분 단계 (S110), 상기 연속형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 제1 합성 데이터를 생성하는 제1 합성 데이터 생성 단계 (S120), 상기 범주형 데이터에 대해 상기 경계 라플라시안 기법

및 후-처리 이산화 (post-processing discretization) 기법을 이용하여 제2 합성 데이터를 생성하는 제2 합성 데이터 생성 단계 (S130), 및 상기 제1 합성 데이터 및 상기 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성하는 전체 합성 데이터 생성 단계 (S140)를 포함할 수 있다.

[0045] 본 발명에 있어, 전체 합성 데이터는 제3자에게 프라이버시 정보가 노출되지 않도록 교란된 데이터 (perturbed data)를 포함할 수 있다. 따라서, 이하에서는 설명의 편의상 합성 데이터 (synthetic data) 및 교란 데이터 (perturbed data)가 동일한 의미를 갖는다고 가정한다.

[0046] 본 발명에 있어, 상술한 각 단계에 따라 차등 프라이버시를 이용한 합성 데이터를 생성하는 방법은 차등 프라이버시를 이용한 합성 데이터 생성 장치에 의해 수행될 수 있다. 상기 합성 데이터 생성 장치는, 실시예에 따라 각 원본 데이터를 소유하는 엔티티 또는 각 원본 데이터 소유 엔티티로부터 데이터를 수합하는 큐레이터 (curator) 등을 포함할 수 있다.

[0047] 도 2는 본 발명에 따라 로컬 차등 프라이버시를 이용한 합성 데이터 생성 방법을 간단히 나타낸 도면이다.

[0048] 일반적으로, 본 발명에 적용 가능한 차등 프라이버시 기법은 노이즈를 추가하는 방식에 따라 로컬 차등 프라이버시 (local differential privacy) 및 글로벌 차등 프라이버시 (global differential privacy)로 구분할 수 있다. 이때, 로컬 차등 프라이버시 방식은 각각의 개별 데이터에 노이즈가 추가되어 해당 데이터들을 애그리케이터 (aggregator)가 수합하는 방식을 의미하고, 글로벌 차등 프라이버시 방식은 별도의 큐레이터 (curator)가 개별 데이터를 수합하여 노이즈를 추가하는 방식을 의미할 수 있다. 상기 글로벌 차등 프라이버시 방식을 위해서, 데이터 베이스 소유자는 상기 큐레이터를 신뢰할 것을 필요로 할 수 있다.

[0049] 본 발명에 따른 합성 데이터 생성 방법은 상술한 모든 차등 프라이버시를 이용한 합성 데이터를 생성하는 방법에 적용될 수 있다. 특히, 본 발명에서는 최악의 시나리오 (예: 큐레이터 및 제3자 모두를 신뢰할 수 없는 상황 등)를 고려한 차등 프라이버시를 이용한 합성 데이터를 생성하는 방법에 적용될 수 있다.

[0050] 본 발명에 적용 가능한 일 예로, 데이터 세트는 질병, 의료 기록, 보험 현황과 같은 민감한 정보를 포함할 수 있기 때문에 의료 영역의 누출에 중요한 영향을 미칠 수 있다. 따라서, 본 발명에서는 네트워크 외부의 어느 누구도 신뢰하지 않음으로써 데이터 유출 위험을 최소화하는 방법을 중점적으로 한 합성 데이터 생성 방법을 개시한다.

[0051] 보다 구체적으로, 도 2에 도시된 바와 같이, 본 발명의 바람직한 일 예에 따른 원본 데이터 소유 엔티티는 로컬 프라이버시 기법에 따라 원본 데이터를 교란 데이터로 변환하여 제3자에게 제공할 수 있다. 이를 위해, 본 발명에 따른 상기 원본 데이터 소유 엔티티는 원본 데이터 (x)에 대해 경계 라플라시안 기법 (M1) 및 후-처리 이산화 (M2)를 적용하여 교란 데이터 (z)를 생성할 수 있다. 이를 통해, 상기 원본 데이터 소유자는 원본 데이터의 프라이버시를 유지하면서 높은 충실도(fidelity)의 데이터를 제3자에게 제공할 수 있다.

[0052] 본 발명에 있어, (로컬) 차등 프라이버시를 이용한 엡실론 (Epsilon, ε)이 설정될 수 있다. 연접하는 데이터 Y_1 및 Y_2를 위해, 하기 수학식 1을 만족하는 경우, 함수 κ는 (ε, δ)-differentially private 하다고 정의할 수 있다.

수학식 1

[0053] $$P[\kappa(Y_1) \in S] \leq \varepsilon \cdot P[\kappa(Y_2) \in S] + \delta$$

[0054] 여기서, $S \subset Range(\kappa)$ 이다.

[0055] 다시 말해, 차등 프라이버시를 이용한 엡실론 값은 차등 프라이버시를 위해 생성된 두 연접 데이터의 결과물 분포의 비율 차이를 의미할 수 있다. 이에 따라, 엡실론 값이 작은 케이스는 엡실론 값이 큰 케이스보다 강한 차등 정보 보호를 제공함을 의미할 수 있다.

[0056] 위와 같은 사항들에 기초하여, S110 단계에서 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는 원본 데이터에 포함된 복수의 데이터를 종류에 따라 연속형 데이터 및 범주형 데이터로 구분할 수 있다. 여기서, 상기 데이터의 종류에 따른 구분 동작은 연속형 데이터와 범주형 데이터를 각각 식별하는 동작을 포함할

수 있다. 다시 말해, 상기 데이터의 종류에 따른 구분 동작은, 연속형 데이터에 대해서는 경계 라플라시안 기법을 적용하고 범주형 데이터에 대해서는 경계 라플라시안 기법 및 후-처리 이산화 기법을 적용하기 위해 연속형 데이터와 범주형 데이터를 구별하여 식별할 수 있다.

[0057] 본 발명에 있어, 연속형 데이터는 이산적 값이 아닌 값을 갖는 모든 데이터를 포함할 수 있다. 반면, 범주형 (categorical) 데이터 (또는, 서수형 (ordinal) 또는 명목형 (norminal) 데이터)는 이산적 값을 갖는 데이터를 포함할 수 있다 (예: 심장 박동수 (heart rates), 약물 치료 이력 (medico-surgical history) 등).

[0058] S120 단계 및 S130 단계에서, 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는 연속형 데이터 및 범주형 데이터에 대해 경계 라플라시안 기법 (bounded Laplacian method)을 이용하여 모든 변수들을 -1 내지 1 범위로 정규화(normalization)할 수 있다. 본 발명에 적용 가능한 일 예로, 상기 S120 단계 및 S130 단계의 경계 라플라시안 기법은 동시에 또는 병렬적으로 수행될 수 있다.

[0059] 일반적인 라플라시안 분포 (Laplacian distribution)는 경계가 무한하기에 이를 그대로 활용하기에 비논리적인 단점 (illogical drawback)을 야기할 수 있다는 단점을 갖는다. 예를 들어, 음의 값을 가질 수 없는 호흡률 (respiratory rates)에 대해 일반적인 라플라시안 기법을 적용할 경우, 상기 호흡률이 음의 값을 갖는 케이스가 발생할 수도 있다. 따라서, 본 발명에서는 경계 라플라시안 기법을 활용하여 상기 문제점을 해결하고자 한다. 이를 통해, 본 발명은 상기 문제점을 해결함과 동시에 데이터 조작 가능성을 최소화할 수 있다.

[0060] 본 발명에 따른 경계 라플라시안 기법에 있어, 입력 변수는 출력 영역 내에 있다고 가정한다. $b > 0,\ W_q: \Omega \rightarrow D$ 가 주어졌을 때, $q \in D$ 인 각각의 q를 위한 확률 밀도 함수 ($f_{W_q}(x)$)는 하기 수학식 2와 같이 정의될 수 있다.

수학식 2

[0061]
$$f_{W_q}(x) = \frac{1}{C_q \cdot 2b} e^{-\frac{|x-q|}{b}}$$

[0062] 여기서, 본 발명에 적용 가능한 일 예로써, 각각의 변수들은

$$b = \frac{\Delta Q}{\varepsilon - \ln(1-\delta)}, \quad C_q = 1 - \frac{1}{2}\left(e^{-\frac{q-l}{b}} + e^{-\frac{u-q}{b}}\right),$$

$\delta = 0, l\,(\text{lower bound}) = -1,\ u(\text{upper bound}) = 1,\ \Delta Q = 2$ 와 같이 정의될 수 있다. 이때, 상위 경계 값 (upper bound) 및 하위 경계 값 (lower bound)은 실시예에 따라 다양하게 변경될 수 있다. 다시 말해, 본 발명에 적용 가능한 다른 예로써, 경계 라플라시안 기법에 적용되는 상위 경계 값 (upper bound) 및 하위 경계 값 (lower bound)은 각각 n 및 -n (여기서 n은 1보다 큰 자연수)로 설정될 수도 있다. 그리고, 엡실론 (ε) 값은 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값으로 설정될 수 있다. 상기 엡실론 값을 결정하는 방법에 대해서는 이후에 상세히 설명한다.

[0063] S130 단계에서, 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는 경계 라플라시안 기법 (bounded Laplacian method)이 적용된 범주형 데이터에 대해 추가적으로 후-처리 이산화 (post-processing discretization) 기법을 적용할 수 있다. 이를 통해, 상기 합성 데이터 생성 장치는 상기 경계 라플라시안 기법에 의해 변환된 (perturbed) 데이터를 확률적으로 이산화할 수 있다.

[0064] 보다 구체적으로, 앞서 상술한 바와 같이 원본 데이터를 교란시키기 위해 경계 라플라시안 기법을 적용하는 경우, 주어진 입력에 대해 무한한 가능성이 있을 수 있다. 반면, 많은 의료 영역의 변수들은 심장 박동수 및 의료-수술 이력 등과 같이 범주형 데이터일 수 있다.

[0065] 따라서, 본 발명에 따른 합성 데이터 생성 장치는 경계 라플라시안 기법이 적용된 범주형 데이터에 대해 추가적인 후-처리 이산화 기법을 적용하여, 해당 데이터를 이산화시킬 수 있다.

[0066] 이를 위한 구체적인 일 예로, 경계 라플라시안 기법이 적용된 범주형 데이터 (예: 도 2의 Intermediate data)에

대해 베르누이 분포 (Bernoulli distrubition) 기법이 적용될 수 있다. 이를 위해, 교란된 데이터 $y \in [-C, C]$ 는 m 조각으로 분리될 수 있다. 이때, m은 원본 입력 변수의 집합 개수 (cardinality)로써 양의 값을 가질 수 있다. 이어, [-C,C] 범위는 [0,m]로 전환되며 동일한 간격 (예: $\frac{2C}{m}$)으로 나뉘어질 수 있다. 주어진 교란된 데이터 y를 위한 k 값은 하기 수학식3을 만족하도록 설정될 수 있다.

수학식 3

$$\frac{2C}{m}k - C < y < \frac{2C}{m}(k+1) - C \quad , \quad k \in \{0,1,\dots,m\}$$

$$k < \frac{m(y+C)}{2C} < k+1$$

$$k := \left\lfloor \frac{m(y+C)}{2C} \right\rfloor$$

[0067]

[0068] k를 산출 결과에 따라, 베르누이 확률 p는 하기 수학식 4를 만족하도록 설정될 수 있다. 이때, 확률 p는 두 인접한 확률들 간의 거리를 의미할 수 있다.

수학식 4

$$p = \frac{m(y+C)}{2C} - k = \frac{m(y+C)}{2C} - \left\lfloor \frac{m(y+C)}{2C} \right\rfloor$$

[0069]

[0070] 끝으로, 베르누이 확률 p를 고려하여 교란된 데이터 y는 하기 수학식 5를 만족하도록 이산화될 수 있다.

수학식 5

$$z = \begin{cases} \frac{2C}{m}k - C, & if\ \mathcal{B}(p) = 0 \\ \frac{2C}{m}(k+1) - C, & if\ \mathcal{B}(p) = 1 \end{cases}$$

[0071]

[0072] 위 수학식에서, B는 베르누이 분포 함수를 의미한다.

[0073] S140 단계에서, 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는 제1 합성 데이터 및 제2 합성 데이터를 이용하여 전체 합성 데이터를 생성할 수 있다. 이를 위해, 상기 합성 데이터 생성 장치는 적정한 엡실론 값을 선택하여 상기 전체 합성 데이터를 생성할 수 있다. 일 예로, 상기 사용되는 엡실론 값으로는 원본 데이터 대비 상기 전체 합성 데이터의 정확도 (accuracy) 값이 일정 값 이상인 엡실론 값이 활용될 수 있다. 구체적인 일 예로, 상기 일정 값은 75 % 이상인 값을 가질 수 있고, 상기 엡실론 값은 상기 일정 값에 기초하여 10^3 내지 10^4 범위에 포함되는 하나의 값을 가질 수 있다. 다만, 상기 값들은 하나의 예시에 불과하며, 실시예에 따라 다양한 값으로 설정될 수 있다. 다시 말해, 상기 예시를 일반적으로 설명하면, 상기 일정 값은 X % 이상인 값으로 설정될 수 있고, 상기 X 값 및 데이터의 특성에 따라 엡실론 값은 일정 범위에 포함되는 값으로 설정될 수 있다.

[0074] 이하에서는 상술한 합성 데이터 생성 방법에 대한 시뮬레이션 검증 결과에 대해 상세히 설명한다.

[0075] 먼저, 본 발명에 대한 시뮬레이션 검증을 위해, eICU Collaborative Research Database의 데이터가 활용되었다. 보다 구체적으로, 첫째, 차등 프라이버시 알고리즘이 주어진 원래 데이터를 효과적으로 교란하는지 여부를 평가하기 위해, 두 데이터 사이의 유사성을 측정할 때 범주형 데이터에 대해서는 중첩 기법이, 연속형 데이터에 대해서는 평균 제곱 오차(MSE) 기법이 활용되었다. 둘째, 차등 프라이버시가 데이터 세트의 효용성에 어떤 악영향을 미치는지 평가하기 위해, 다양한 엡실론 값에 기반한 APACHE 점수 변수를 사용하여 중환자실(ICU) 입원 후 사망률을 예측하는 정확도가 서로 비교되었다. 데이터 세트에는 관입 (intubated), 환기 (ventilation), 투석 (dialysis), 약물 상태 (medication status, 이때, cardinality=2가 적용될 수 있음), 눈, 운동 (motor), 언어 상태 (verbal status, 이때, cardinality=6가 적용될 수 있음) 등의 범주형 데이터가 포함될 수 있다. 또한, 상기 데이터 세트에는 소변 생산량 (urine output), 온도, 호흡률, 나트륨 (sodium), 심박수, 평균 혈압, pH, 헤마토크리트 (hematocrit), 크레아티닌, 알부민, 산소 압력, CO2 압력, 혈당 질소, 포도당, 빌리루빈 및 FiO2 값 등의 연속형 데이터가 포함될 수 있다. 상기 데이터 세트에 있어, 초기에는 148,532명의 환자(행)가 있었지만, 결측값을 삭제한 후 총 4,740명의 환자(생존 3,597명, 유효 기간 1,143명)의 데이터가 포함되었다. 사망률 예측을 위해 다음과 같은 기계 학습 방법이 활용되었다: 의사 결정 트리 (Decision tree), K-Nearest Neighbor, 지원 벡터 머신 (Support Vector Machine), 로지스틱 회귀 분석 (Logistic Regression), Naive Bayes, 랜덤 포리스트 (random Forest). 주어진 데이터는 80 대 20의 비율로 설정된 트레인 (train)으로 나뉘었다. 모든 예측은 5배 교차 검증 방식 (five-fold cross-validation manner)을 사용하여 평균화되었으며, 이를 위해 파이썬 프로그래밍 언어를 사용하는 Scikit-learn 라이브러리가 사용되었다.

[0076] **[경계 라플라시안 함수의 검증을 위한 합성 데이터 (Synthetic data for validation of bounded Laplacian function)]**

[0077] 본 발명에 따르면, -1과 1 사이의 일정한 간격의 분포가 생성되어 경계 라플라시안 기법이 적용될 수 있다. 범위가 무한대인 기존의 라플라시안 기법과 대조적으로, 상기 경계 라플라시안 기법은 -1에서 1까지 범위를 가질 수 있다.

[0078] 도 3a 내지 도 3c는 본 발명에 따른 경계 라플라시안 기법의 성능을 나타내는 그래프이다. 보다 구체적으로, 도 3a는 -1 내지 1 범위로 임의적으로 생성된 연속형 데이터의 히스토그램을 나타낸 도면이고, 도 3b는 원래 0 내지 9 범위로 생성되어 -1 내지 1 범위로 정규화된 임의적으로 생성된 연속형 데이터의 히스토그램을 나타낸 도면이고, 도 3c는 도 3b에 후-처리 이산화가 적용된 이후 임의적으로 생성된 연속형 데이터의 히스토그램을 나타낸 도면이다. 상기 예시들에 있어, $\varepsilon=0.1$, $\delta=0$ 인 라플라시안 기법이 적용되었다고 가정한다.

[0079] 도 3a을 참고하면, 기존 라플라시안 기법에 따른 출력 값은 경계 라플라시안 기법에서는 존재하지 않는 범위 밖에 위치한 출력 값을 포함할 수 있다. 보다 구체적으로, 범주형 데이터와 후-처리 이산화 기법을 테스트하기 위해, 본 시뮬레이션에서 우리는 0에서 9 사이의 100개의 랜덤 정수를 생성한 이후 -1에서 1까지의 범위로 정규화했다. 이때, 기존 라플라시안 기법에 따르면, 범위를 벗어난 몇몇 출력 값들이 발생하였다. 반면, 경계 라플라시안 기법이 적용된 범주형 데이터는 연속형 데이터처럼 일정 데이터 범위 내에 존재하였다.

[0080] 다만, 도 3b을 참고하면, 일부 범주형 데이터는, 경계 밖의 값과 유사하게, 처음에 주어진 데이터에 존재하지 않음을 확인할 수 있다. 따라서, 도 3c를 참고하면, 추가적인 후-처리 이산화가 적용될 경우, 모든 범주형 데이터가 일정 데이터 범위 내에 존재하는 것이 보장됨을 확인할 수 있다.

[0081] **[실제 데이터를 사용한 검증 (Validation using real-world data)]**

[0082] 본 검증을 위해, 우리는 eICU Collaborative Research Database를 활용하였다. 또한, 연속형 데이터의 평균 제곱 오차(MSE)와 범주형 데이터의 오분류 비율 (misclassification rates)을 사용하여 원래 데이터와 교란된 데이터의 차이를 산출하였다.

[0083] 도 4a 및 도 4b는 본 발명에 따른 엡실론 값 및 데이터 교란 정도의 관계를 간단히 나타낸 그래프이다. 보다 구체적으로, 도 4a는 연속형 데이터들에 엡실론 값 및 데이터 교란 정도의 관계를 나타낸 그래프이고, 도 4b는 범주형 데이터들에 엡실론 값 및 데이터 교란 정도의 관계를 나타낸 그래프이다.

[0084] 도 4a를 참고하면, 원본 데이터의 값 간 분산으로 인해 eICU의 연속형 데이터의 MSE가 다양함을 확인할 수 있다. 예를 들어, pH와 알부민은 서로 다른 개인들 간에 유사하지만, 심박수와 포도당은 광범위한 차이를 가지는 것을 확인할 수 있다.

[0085] 도 4b를 참고하면, 범주형 데이터에 있어, 삽관, 환기, 투석 상태는 0 또는 1이고 기회 수준 (chance level)은

0.5이다. 눈은 0에서 4까지, 언어 범위는 0에서 5까지, 그리고 모터는 0에서 6까지 범위를 가질 수 있다. 따라서, 특히 ε가 작을 경우 오분류율에 차이가 있음을 확인할 수 있다.

[0086] 반면, 도 4a 및 도 4b를 참고하면, ε가 증가하면 연속형 데이터 및 범주형 데이터에서 모든 교란 값이 원래 값에 근접하게 되는 것을 확인할 수 있다.

[0087] 도 5는 본 발명에 따른 엡실론 값에 대한 다른 기계 학습 모델 간의 분류 정확도를 나타낸 그래프이다. 도 5에 있어, 중앙 집중식 모델 (centralized model)의 성능은 파선 (dashed lines)으로 표시된다.

[0088] 엡실론 값에 대한 데이터 효용을 시뮬레이션하기 위해, 우리는 eICU 데이터 세트의 사망률을 예측하기 위한 예측 분류기를 구성했다. 특히 4,740명의 환자로부터 3,597명의 살아있는 환자가 있는 바, 76% 정확도의 기회 수준 (chance level)을 제공하였다. 도 5를 참고하면, 엡실론 값이 낮을수록 데이터 교란이 심해져 기회 수준 (chance level)에 가까운 정확도가 나타났다. 반면, 엡실론 값이 커질수록 분류 성능이 향상되어 원본 데이터를 이용한 정확도로 수렴함을 확인할 수 있다 (도 5의 파선). 이러한 경향은 다른 모델들 사이에서 일관되게 나타났고, 특히 랜덤 포레스트가 최고의 성능을 발휘함을 확인할 수 있다.

[0089] 이처럼, 본 발명에 따른 합성 데이터 생성 방법에 따르면, 차등 프라이버시에 따른 개인 정보 보호 효과를 높이며 동시에 해당 데이터의 활용도 또한 높일 수 있다.

[0090] 추가적으로, 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치는 상술한 방법에 따라 생성된 전체 합성 데이터를 제3자에게 제공 (예: publish 등) 할 때 적용된 엡실론 값 정보를 함께 제공할 수 있다.

[0091] 도 6은 본 발명의 다른 예에 따른 차등 프라이버시를 이용한 합성 데이터 생성 장치를 나타낸 도면이다.

[0092] 도 6에 도시된 바와 같이, 본 발명에 따른 차등 프라이버시를 이용한 합성 데이터 생성 방법을 수행하는 합성 데이터 생성 장치 (600)는 데이터 베이스 (610) 및 프로세서 (620)를 포함할 수 있다.

[0093] 이때, 데이터 베이스 (610)는 원본 데이터뿐만 아니라 상술한 합성 데이터 생성 방법에 따라 생성된 결과물 (예: 제1 합성 데이터, 제2 합성 데이터, 전체 합성 데이터 등)을 포함할 수 있다. 이어, 프로세서 (620)는 상기 데이터 베이스 (610)와 연결되어 상술한 합성 데이터 생성 방법을 수행하도록 설정될 수 있다.

[0094] 추가적으로, 본 발명에 따른 컴퓨터 프로그램은, 컴퓨터와 결합하여, 앞서 상술한 합성 데이터 생성 방법을 실행시키기 위하여 컴퓨터 판독가능 기록매체에 저장될 수 있다.

[0095] 전술한 프로그램은, 컴퓨터가 프로그램을 읽어 들여 프로그램으로 구현된 상기 방법들을 실행시키기 위하여, 상기 컴퓨터의 프로세서(CPU)가 상기 컴퓨터의 장치 인터페이스를 통해 읽힐 수 있는 C, C++, JAVA, 기계어 등의 컴퓨터 언어로 코드화된 코드(Code)를 포함할 수 있다. 이러한 코드는 상기 방법들을 실행하는 필요한 기능들을 정의한 함수 등과 관련된 기능적인 코드(Functional Code)를 포함할 수 있고, 상기 기능들을 상기 컴퓨터의 프로세서가 소정의 절차대로 실행시키는데 필요한 실행 절차 관련 제어 코드를 포함할 수 있다. 또한, 이러한 코드는 상기 기능들을 상기 컴퓨터의 프로세서가 실행시키는데 필요한 추가 정보나 미디어가 상기 컴퓨터의 내부 또는 외부 메모리의 어느 위치(주소 번지)에서 참조되어야 하는지에 대한 메모리 참조관련 코드를 더 포함할 수 있다. 또한, 상기 컴퓨터의 프로세서가 상기 기능들을 실행시키기 위하여 원격(Remote)에 있는 어떠한 다른 컴퓨터나 서버 등과 통신이 필요한 경우, 코드는 상기 컴퓨터의 통신 모듈을 이용하여 원격에 있는 어떠한 다른 컴퓨터나 서버 등과 어떻게 통신해야 하는지, 통신 시 어떠한 정보나 미디어를 송수신해야 하는지 등에 대한 통신 관련 코드를 더 포함할 수 있다.

[0096] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

[0097] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

부호의 설명

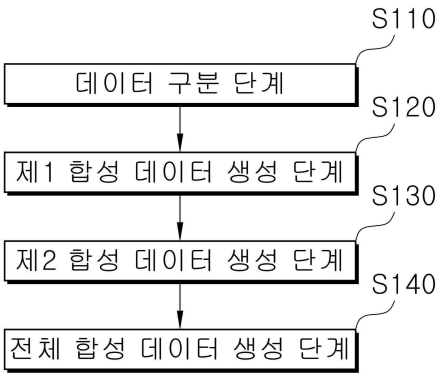
[0098] 400 : 합성 데이터 생성 장치

410: 데이터 베이스

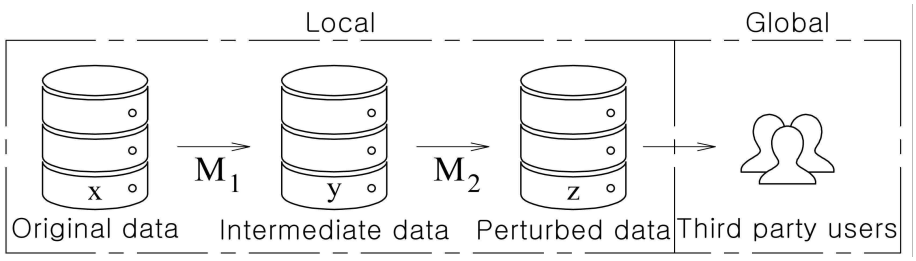
420: 프로세서

도면

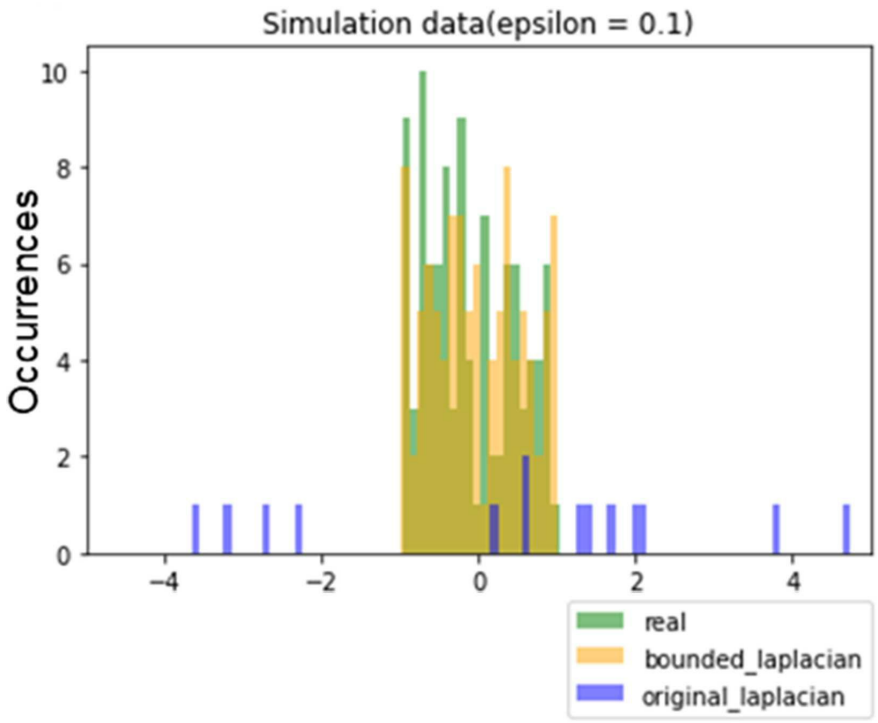
도면1



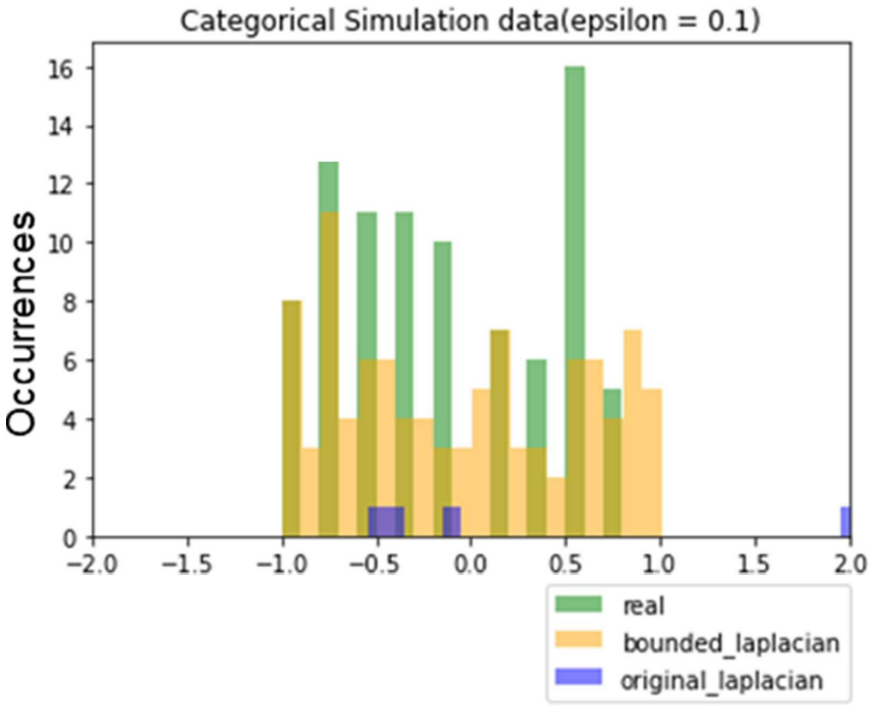
도면2



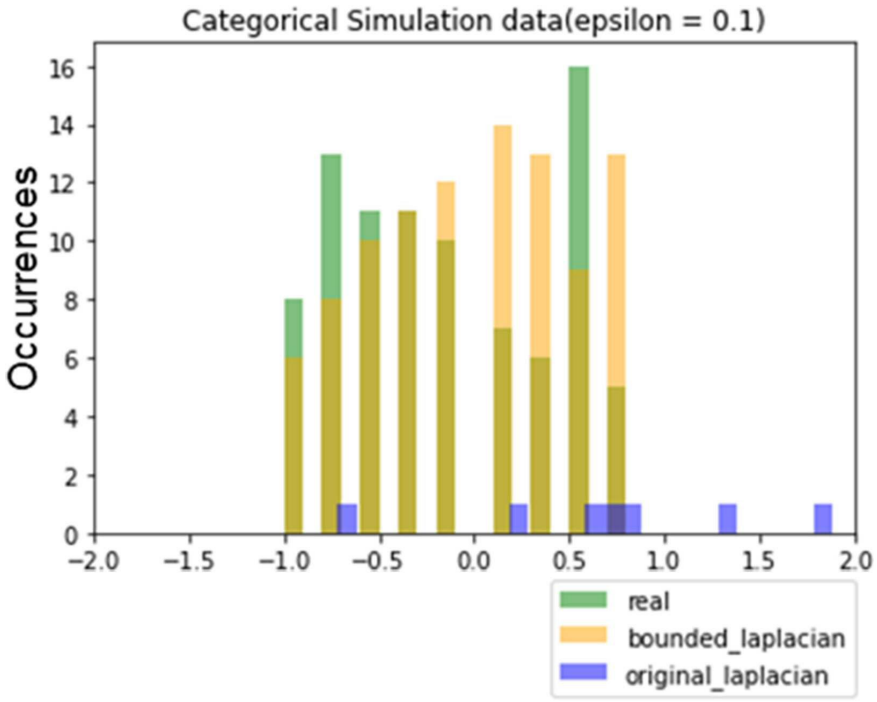
도면3a


Simulation data(epsilon = 0.1)
Occurrences
real
bounded_laplacian
original_laplacian

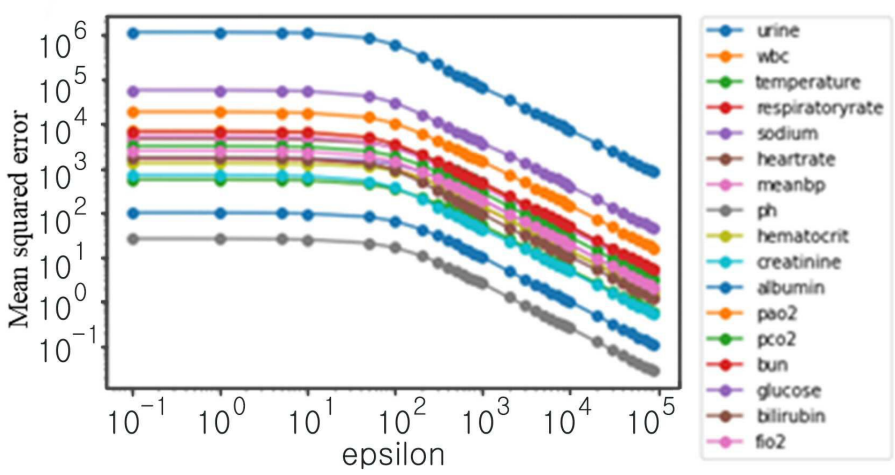
도면3b


Categorical Simulation data(epsilon = 0.1)
Occurrences
real
bounded_laplacian
original_laplacian

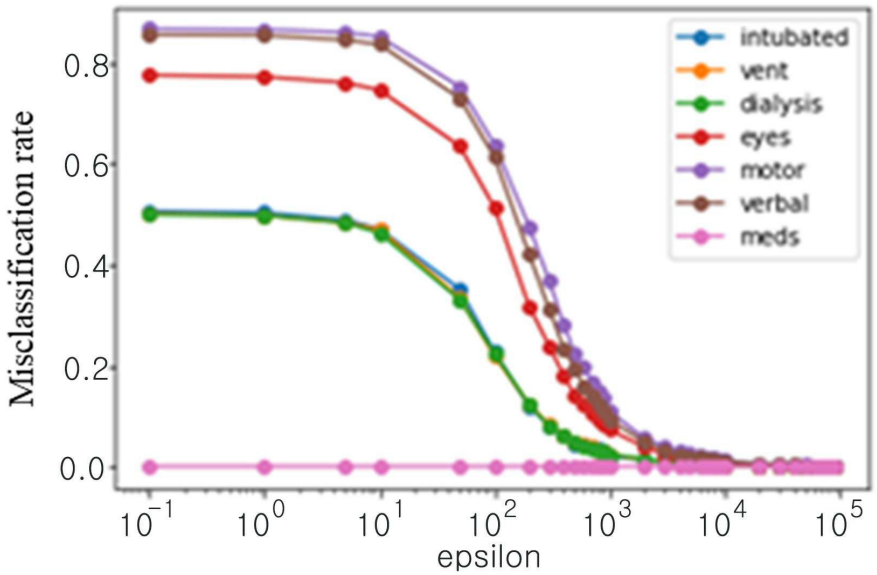
도면3c


Categorical Simulation data(epsilon = 0.1)
Occurrences
real
bounded_laplacian
original_laplacian

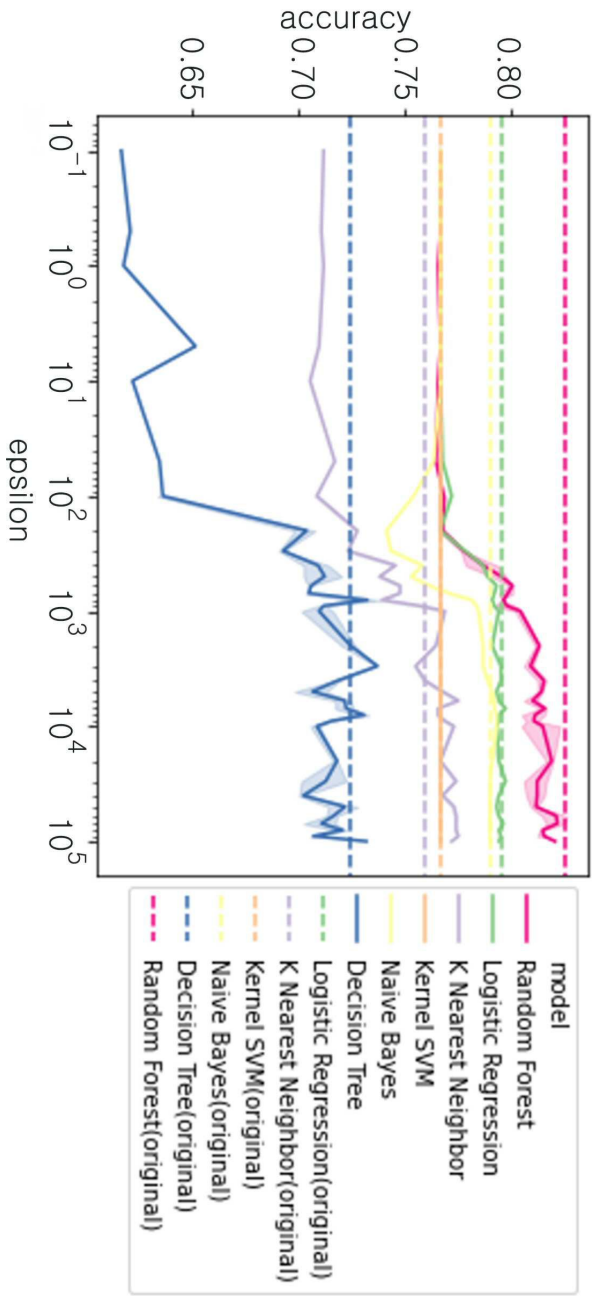
도면4a


Mean squared error
epsilon
urine
wbc
temperature
respiratoryrate
sodium
heartrate
meanbp
ph
hematocrit
creatinine
albumin
pao2
pco2
bun
glucose
bilirubin
fio2

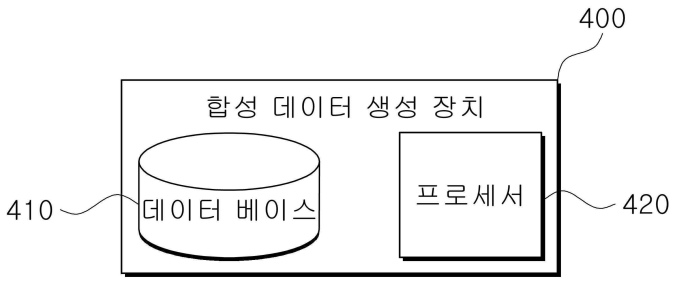
도면4b


intubated
vent
dialysis
eyes
motor
verbal
meds
Misclassification rate
0.0
0.2
0.4
0.6
0.8
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
epsilon

도면5


accuracy
0.65
0.70
0.75
0.80
epsilon
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
model
Random Forest
Logistic Regression
K Nearest Neighbor
Kernel SVM
Naive Bayes
Decision Tree
Logistic Regression(original)
K Nearest Neighbor(original)
Kernel SVM(original)
Naive Bayes(original)
Decision Tree(original)
Random Forest(original)

도면6


400
합성 데이터 생성 장치
410
데이터 베이스
프로세서
420