(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0089592
(43) 공개일자 2022년06월28일

(51) 국제특허분류(Int. Cl.)
G06F 16/215 (2019.01) *G06F 16/22* (2019.01)
G06N 3/08 (2006.01) *G06N 5/02* (2006.01)
(52) CPC특허분류
G06F 16/215 (2019.01)
G06F 16/2228 (2019.01)
(21) 출원번호 10-2021-0014007
(22) 출원일자 2021년02월01일
심사청구일자 2021년02월01일
(30) 우선권주장
1020200179416 2020년12월21일 대한민국(KR)

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
유성준
서울특별시 광진구 능동로 209, 대양AI센터 719호(군자동)
구영현
서울특별시 광진구 능동로 209, 대양AI센터 717호(군자동)
박나리
서울특별시 광진구 능동로 209, 학술정보원 7층(군자동)
(74) 대리인
양성보

전체 청구항 수 : 총 16 항

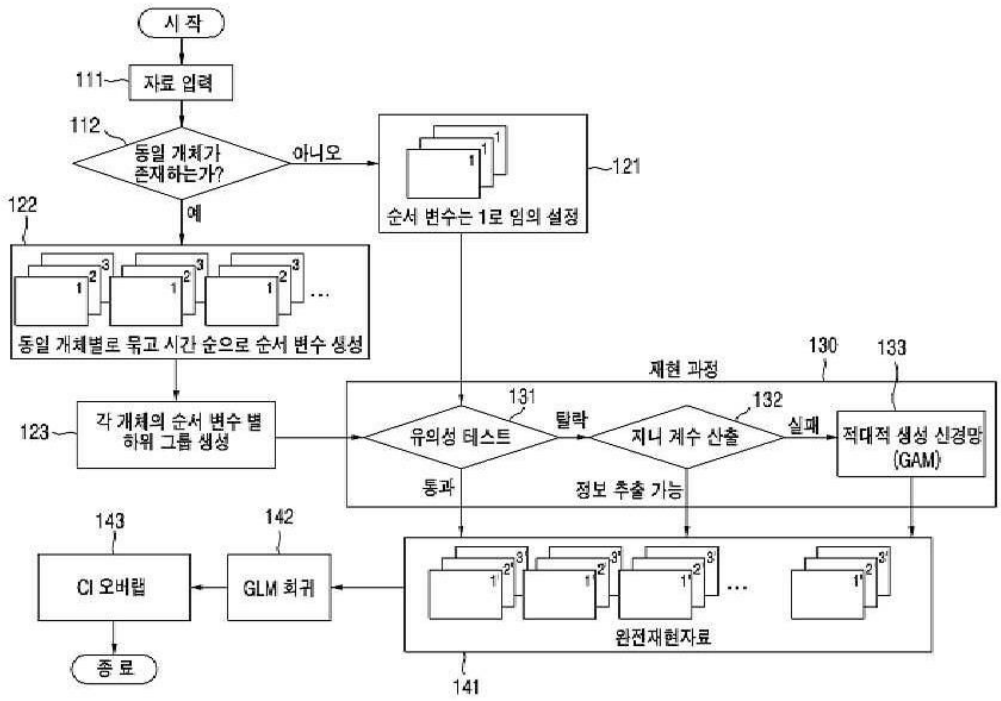
(54) 발명의 명칭 **재현자료 DB 생성 알고리즘**

(57) 요 약

재현자료 DB 생성 방법 및 장치가 제시된다. 본 발명에서 제안하는 재현자료 DB 생성 방법은 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 단계, 개체 별

(뒷면에 계속)

대 표 도



식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성하는 단계, 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계, 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행하는 단계 및 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계를 포함한다.

(52) CPC특허분류

G06N 3/08 (2013.01)

G06N 5/02 (2019.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711103318
과제번호	2017-0-00302-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	ICT융합산업원천기술개발(R&D)
연구과제명	자가진화형 인공지능 투자 기술 개발
기 여 율	1/1
과제수행기관명	세종대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 단계;

개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성하는 단계;

개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계;

상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행하는 단계; 및

모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계

를 포함하는 재현자료를 생성 방법.

청구항 2

제1항에 있어서,

개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계는,

개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 동일개체에 대한 기록들을 인식한 뒤, 기록들의 종류에 따라 선분이력형태로 변경하고 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 순서변수를 생성하는

재현자료를 생성 방법.

청구항 3

제1항에 있어서,

상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 단계는,

조건부 추론 방식(Conditional Inference Tree; CTREE)을 통해 유의성 테스트를 수행하고,

분류 및 회귀 방식(Classification and Regression Tree; CART)을 통해 지니 계수를 산출하고,

적대적 생성 신경망을 통해 학습을 수행하는

재현자료를 생성 방법.

청구항 4

제3항에 있어서,

과적합을 피하기 위한 순서변수 선택을 위해 조건부 추론 방식을 통한 유의성 테스트를 수행하고 공변량 선택 후 노드 분할의 과정을 재귀적으로 수행하는 조건부 추론 방식 알고리즘의 모든 시작 부분에서 순열을 검정(permutation test)하며, 과적합을 일으키는 트리기반 분류 및 회귀 방식을 보완하기 위해 조건부 추론 방식을 우선적으로 적용하는

재현자료를 생성 방법.

청구항 5

제3항에 있어서,

조건부 추론 방식을 우선 적용한 뒤 하위 그룹에 관한 데이터의 분포에 따라 또는 변수의 값 또는 범위가 결측값(NA)이거나 또는 검정통계량을 산출하지 못해 조건부 추론이 실패한 데이터에 대해 과적합 여부에 상관없이 바이너리 분할 방식인 분류 및 회귀 방식을 적용하는

재현자료를 생성 방법.

청구항 6

제3항에 있어서,

조건부 추론 방식 및 분류 및 회귀 방식을 통해 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 적대적 생성 신경망을 적용하는

재현자료를 생성 방법.

청구항 7

제6항에 있어서,

적대적 생성 신경망은 조건부 변수(y)가 있는 적대적 생성 신경망의 확장으로, 입력 레이어로 판별자와 생성자 모두에 제공되며, 순서변수를 조건부 변수(y)로 설정하여 시작되며, 각 부분집합에 대해, 조건(y), 랜덤 노이즈(z) 및 랜덤 생성된 조건(y')에 대해 원래 데이터(x)로 판별자를 훈련하고, 판별자에 의해 학습된 생성자는 랜덤 노이즈(z)와 랜덤 생성된 조건(y')을 입력으로 취하고, 판별자에 입력값이 되는 가짜 데이터(x') 및 가짜 조건(y')을 생성하고, 이후 판별자에 의해 실수로 예측된 레이블이 지정된 데이터는 다시 원래 데이터의 형태로 변환되는

재현자료를 생성 방법.

청구항 8

제1항에 있어서,

모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계는,

재현 결과를 검증하기 위해 특정 순서변수의 효율에 대한 신뢰구간(Confidence Interval; CI)의 오버랩 구간을 계산하여 평균으로 나타내고, 계산된 오버랩 구간이 넓을수록 유용성이 높은 것으로 해석하는

재현자료를 생성 방법.

청구항 9

특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 자료 입력부;

개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성하고, 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 하위 그룹 생성부;

상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 재현 자료 생성부; 및

모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 재현 자료 검증부

를 포함하는 재현자료를 생성 장치.

청구항 10

제9항에 있어서,

하위 그룹 생성부는,

개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 동일개체에 대한 기록들을 인식한 뒤, 기록들의 종류에 따라 선분이력형태로 변경하고 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 순서변수를 생성하는

재현자료를 생성 장치.

청구항 11

제9항에 있어서,

재현 자료 생성부는,

조건부 추론 방식(Conditional Inference Tree; CTREE)을 통해 유의성 테스트를 수행하고,

분류 및 회귀 방식(Classification and Regression Tree; CART)을 통해 지니 계수를 산출하고,

적대적 생성 신경망을 통해 학습을 수행하는

재현자료를 생성 장치.

청구항 12

제11항에 있어서,

재현 자료 생성부는,

과적합을 피하기 위한 순서변수 선택을 위해 조건부 추론 방식을 통한 유의성 테스트를 수행하고 공변량 선택 후 노드 분할의 과정을 재귀적으로 수행하는 조건부 추론 방식 알고리즘의 모든 시작 부분에서 순열을 검정(permutation test)하며, 과적합을 일으키는 트리기반 분류 및 회귀 방식을 보완하기 위해 조건부 추론 방식을 우선적으로 적용하는

재현자료를 생성 장치.

청구항 13

제11항에 있어서,

재현 자료 생성부는,

조건부 추론 방식을 우선 적용한 뒤 하위 그룹에 관한 데이터의 분포에 따라 또는 변수의 값 또는 범위가 결측값(NA)이거나 또는 검정통계량을 산출하지 못해 조건부 추론이 실패한 데이터에 대해 과적합 여부에 상관없이 바이너리 분할 방식인 분류 및 회귀 방식을 적용하는

재현자료를 생성 장치.

청구항 14

제11항에 있어서,

재현 자료 생성부는,

조건부 추론 방식 및 분류 및 회귀 방식을 통해 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 적대적 생성 신경망을 적용하는

재현자료를 생성 장치.

청구항 15

제14항에 있어서,

적대적 생성 신경망은 조건부 변수(y)가있는 적대적 생성 신경망의 확장으로, 입력 레이어로 판별자와 생성자 모두에 제공되며, 순서변수를 조건부 변수(y)로 설정하여 시작되며, 각 부분집합에 대해, 조건(y), 랜덤 노이즈(z) 및 랜덤 생성된 조건(y')에 대해 원래 데이터(x)로 판별자를 훈련하고, 판별자에 의해 학습된 생성자는 랜덤 노이즈(z)와 랜덤 생성된 조건(y')을 입력으로 취하고, 판별자에 입력값이 되는 가짜 데이터(x') 및 가짜 조건(y')을 생성하고, 이후 판별자에 의해 실수로 예측된 레이블이 지정된 데이터는 다시 원래 데이터의 형태로 변환되는

재현자료를 생성 장치.

청구항 16

제9항에 있어서,

재현 자료 검증부는,

재현 결과를 검증하기 위해 특정 순서변수의 효율에 대한 신뢰구간(Confidence Interval; CI)의 오버랩 구간을 계산하여 평균으로 나타내고, 계산된 오버랩 구간이 넓을수록 유용성이 높은 것으로 해석하는

재현자료를 생성 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 개인정보보호법으로 인해 제한적으로 사용되는 금융, 고용 및 근로 정보 DB를 타 산업분야에서 융합하여 활용할 수 있도록 재현자료를 생성하는 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] AI 또는 IoT, 빅데이터 기반 연구 및 서비스 개발은 데이터 접근이나 활용이 가장 중요하지만 충분한 학습데이터 확보가 어렵다는 한계가 있다. 개인정보 침해 등 여러 가지 이유로 국내에서는 데이터 활용에 많은 제약이 있다. 예를 들어, 신용카드 사기거래, 사이버 공격, 사고 등 이상치(outlier)에 대한 학습데이터가 부족한 실정이다.

[0003] 각 산업분야에서의 학습 데이터 수집 또는 제공하는 데이터의 비용이 AI 도입으로 기대되는 비용절감 또는 생산성 향상에 비해 너무 크다는 단점이 있다. 부족한 학습데이터를 확보하기 위해 민감정보를 비식별화하여 활용하거나, 기존의 통계적 기법이나 뉴럴 네트워크를 사용하여 재현자료를 생성하고 사용하는 방법이 연구되고 있다.

[0004] 하지만 가명처리, 총계처리, 값 삭제, 범주화, 마스킹 등의 전통적인 익명화 기법은 변수들 간의 관계를 왜곡하는 등 분석에 한계가 있기 때문에 데이터의 유용성이 낮다는 단점이 있다. 이에 대한 대안으로 다른 변수들의 조건부 확률 분포를 추정하여 적절한 한 개의 대체값이 아니라 주로 여러 개를 제공하는 다중대체기법을 적용하여 재현자료를 생성하는 기법을 사용하기 시작했다.

[0005] 금융 분야에서의 재현자료에 대한 일반적인 요구 사항은 다음과 같다. 사기 활동, 경기 침체 또는 내부 또는 외부 요인에 의한 새로운 소비자 행동 경향과 같은 특정 사건에 대한 과거 데이터가 부족한데, 재현자료는 시뮬레이션 및 기계 학습 알고리즘 훈련을 위해 이렇게 드문 경우의 데이터를 생성할 수 있다.

[0006] 재현자료를 다른 산업과 공동으로 평가하는 경우 데이터의 가치가 높아지므로 노출위험 없이 유용성(utility)데이터를 공유하는 것이 필요하다. 클라우드 서비스 또는 컴퓨팅 파워와 같은 인프라가 준비되지 않았고, 재현자료를 훈련 모델에 사용하고 현장에서 실제 데이터에 적용할 수 있는 경우 방대한 양의 데이터를 공유할 수 없다는 문제점이 있다.

[0007] 기존의 재현데이터 연구 및 활용 사례를 보면 한 시점의 데이터셋을 기준으로 하거나, 랜덤(정규분포)하거나, 또는 특정 분포를 따르고 있다.

[0008] 종래기술에 사용된 예시 및 상용 또는 비상업용 프로그램들 대부분이 하나의 행이 하나의 개체를 나타내는 가구 또는 설문 조사 자료를 대상으로 하고 있으며(다시 말해, 횡단연구), 인구조사 같은 모집단 추정은 정규분포를 가정하는 경우가 대부분이다.

[0009] 종래기술에 따른 합성데이터, 재현데이터, 합성 데이터 생성에서, 이미지(영상) 또는 음성(음역대 또는 단어)

합성이 주를 이루고 있으며, 이들 데이터의 특징은 특정 분포를 따르고 있다는 것이다.

[0010] 반면에 개인이나 기업의 금융(신용)데이터 또는 기업의 종사자 정보는 인구 조사나 설문 조사 자료와는 성격이 다르며, 노출될 경우 개인이나 기업에 차별적으로 작용할 우려로 인해 데이터 공개 및 활용에 더욱 보수적이다.

[0011] 이러한 개인 및 기업의 금융 관련 데이터는 변수가 충분히 많지 않으며, 대부분이 민감 정보일 확률이 높거나, 시계열 특성을 가지며, 종종 샘플이나 변수 단위의 특징이 관찰된다는 특성이 있다.

[0012] 이는 금융 관련 데이터가 가구 또는 설문 조사와는 본질적으로 다르다는 것을 나타내며, 정보 손실을 최소화하면서 정보를 보호하는 것이 더욱 어렵다는 것을 의미한다.

발명의 내용

해결하려는 과제

[0013] 본 발명이 이루고자 하는 기술적 과제는 개인정보보호법으로 인해 제한적으로 사용되는 금융, 고용 및 근로 정보 DB를 타 산업분야에서 융합하여 활용할 수 있도록 재현자료를 생성하는 기술을 제공하는데 있다. 종래 기술의 인구통계학적 자료의 구조와 달리 개인의 금융, 고용 및 근로 정보를 시간 순으로 재현하여 특정 기간의 특징까지 원자료와 유사하도록 재현하고자 한다.

과제의 해결 수단

[0014] 일 측면에 있어서, 본 발명에서 제안하는 재현자료 DB 생성 방법은 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 단계, 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성하는 단계, 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계, 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행하는 단계 및 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계를 포함한다.

[0015] 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계는 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 동일개체에 대한 기록들을 인식한 뒤, 기록들의 종류에 따라 선분이력형태로 변경하고 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 순서변수를 생성한다.

[0016] 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 단계는 조건부 추론 방식(Conditional Inference Tree; CTREE)을 통해 유의성 테스트를 수행하고, 분류 및 회귀 방식(Classification and Regression Tree; CART)을 통해 지니 계수를 산출하고, 적대적 생성 신경망을 통해 학습을 수행한다.

[0017] 과적합을 피하기 위한 순서변수 선택을 위해 조건부 추론 방식을 통한 유의성 테스트를 수행하고 공변량 선택 후 노드 분할의 과정을 재귀적으로 수행하는 조건부 추론 방식 알고리즘의 모든 시작 부분에서 순열을 검정(permutation test)하며, 과적합을 일으키는 트리기반 분류 및 회귀 방식을 보완하기 위해 조건부 추론 방식을 우선적으로 적용한다.

[0018] 조건부 추론 방식을 우선 적용한 뒤 하위 그룹에 관한 데이터의 분포에 따라 또는 변수의 값 또는 범위가 결측값(NA)이거나 또는 검정통계량을 산출하지 못해 조건부 추론이 실패한 데이터에 대해 과적합 여부에 상관없이 바이너리 분할 방식인 분류 및 회귀 방식을 적용한다.

[0019] 조건부 추론 방식 및 분류 및 회귀 방식을 통해 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 적대적 생성 신경망을 적용한다.

[0020] 적대적 생성 신경망은 조건부 변수(y)가있는 적대적 생성 신경망의 확장으로, 입력 레이어로 판별자와 생성자 모두에 제공되며, 순서변수를 조건부 변수(y)로 설정하여 시작되며, 각 부분집합에 대해, 조건(y), 랜덤 노이즈(z) 및 랜덤 생성된 조건(y')에 대해 원래 데이터(x)로 판별자를 훈련하고, 판별자에 의해 학습된 생성자는 랜덤 노이즈(z)와 랜덤 생성된 조건(y')을 입력으로 취하고, 판별자에 입력값이 되는 가짜 데이터(x') 및 랜덤 생

성된 조건(y')을 생성하고, 이후 판별자에 의해 실수로 예측된 레이블이 지정된 데이터는 다시 원래 데이터의 형태로 변환된다.

[0021] 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계는 재현 결과를 검증하기 위해 특정 순서변수의 효율에 대한 신뢰구간(Confidence Interval; CI)의 오버랩 구간을 계산하여 평균으로 나타내고, 계산된 오버랩 구간이 넓을수록 유용성이 높은 것으로 해석한다.

[0022] 또 다른 일 측면에 있어서, 본 발명에서 제안하는 재현자료 DB 생성 장치는 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 자료 입력부, 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성하고, 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 하위 그룹 생성부, 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 재현 자료 생성부 및 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 재현 자료 검증부를 포함한다.

발명의 효과

[0023] 본 발명의 실시예들에 따르면 개인정보보호법으로 인해 제한적으로 사용되는 금융, 고용 및 근로 정보 DB를 타 산업분야에서 융합하여 활용할 수 있도록 재현자료를 생성할 수 있다. 종래 기술의 인구통계학적 자료의 구조와 달리 개인의 금융, 고용 및 근로 정보를 시간 순으로 재현하여 특정 기간의 특징까지 원자료와 유사하도록 재현할 수 있다. 트리기반 알고리즘과 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 사용하여 매스킹 위주의 재현자료 생성으로 원자료의 민감정보를 노출하지 않으면서도 자료의 유용성(utility)을 최대한 보존할 수 있다. 또한, 원자료와 동일한 개인정보를 포함하지 않는 완전재현금융자료를 생성하여 타 산업과 원활하게 융합하여 부가 가치를 창출하고, 금융전문가가 아닌 개발 또는 분석 목적으로 실데이터에 접속하기 전 교육용 자료로 활용 가능하다.

도면의 간단한 설명

[0024] 도 1은 본 발명의 일 실시예에 따른 재현자료 DB 생성 방법을 설명하기 위한 흐름도이다.

도 2는 본 발명의 일 실시예에 따른 재현자료 DB 생성 장치의 구성을 나타내는 도면이다.

도 3은 본 발명의 일 실시예에 따른 적대적 생성 신경망을 통한 재현 자료 생성 과정을 설명하기 위한 도면이다.

도 4는 본 발명의 일 실시예에 따른 적대적 생성 신경망의 생성자의 학습 과정을 설명하기 위한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0025] 본 발명의 실시 예에 따른 용어 "변수"는 테이블의 컬럼을 의미할 수 있다. 본 발명의 실시 예에 따른 용어 "클래스"는 범주형 변수의 데이터 범위(예를 들어, 성별 컬럼에서 1: 남, 2: 여인 경우, 1과 2는 클래스가 됨)를 의미할 수 있다. 본 발명의 실시 예에 따른 용어 "개체"는 예를 들어, 종사자 또는 경제활동을 하는 개인을 의미할 수 있다. 본 발명의 실시 예에 따른 용어 "과적합"은 학습이 과하게 진행되어 부분집합에는 잘 맞지만 실제 데이터와는 맞지 않게 되는 현상을 의미하고, 여기에서는 분류를 너무 꼼꼼하게 진행하여 원본과 동일한 분포가 나타나는 현상을 의미할 수 있다. 본 발명의 실시 예에 따른 용어 "선분이력형태"는 동일 개체 내 동일 기록(예를 들어, 신용계좌)이 여러 개일 때, 생성일과 종료일로 묶어 하나의 기록으로 만든 형태를 의미할 수 있다. 이하, 본 발명의 실시 예를 첨부된 도면을 참조하여 상세하게 설명한다.

[0027] 도 1은 본 발명의 일 실시예에 따른 재현자료 DB 생성 방법을 설명하기 위한 흐름도이다.

[0028] 제안하는 재현자료 DB 생성 방법은 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출하는 단계, 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기

록에 관한 하위 그룹을 생성하는 단계, 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 단계, 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행하는 단계 및 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계를 포함한다.

[0029] 본 발명의 실시예에 따르면, 개인신용정보나 근로정보는 분석가가 지역, 나이, 성별 등의 조건을 설정하는지 여부와 상관없이 특정 집단의 특징을 반영하는 모집단으로 고려될 수 있다. 데이터가 커질수록 정규분포를 따르거나, 특정 분포를 따라가는 인구통계학적 데이터와 다르게, 지역이나 산업분류에 따른 종사자 특징 및 신용정보는 정규를 포함한 그 어떠한 특정분포를 따르지 않는 경우가 많다. 이것은 데이터의 특성 외에도 수집 당시의 오기, 오타 등 외부요인에 기인하기도 한다. 따라서, 제안하는 재현자료 DB 생성 방법은 이와 유사한 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단 자료에 적용될 수 있으며, 범주형 및 연속형 자료 모두를 포함할 수 있다.

[0030] 본 발명의 실시예에 따른 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료는 개인신용정보 또는 근로정보일 수 있다. 개인신용정보나 또는 근로정보는 실시예일뿐 이에 한정되지 않고, 데이터가 커질수록 정규분포를 따르거나, 특정 분포를 따라가는 인구통계학적 데이터 및 지역이나 산업분류에 따른 종사자 특징 및 신용정보와 같이 특정분포를 따르지 않는 자료 모두에 적용될 수 있다. 이하, 본 발명에서는 개인신용정보 또는 근로정보를 예시로서 설명한다.

[0031] 도 1을 참조하면, 먼저 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받을 수 있다(111). 예를 들어, 종사자 자료(다시 말해, 근무정보) 또는 개인신용정보 자료를 입력 받을 수 있다.

[0032] 이후, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출한다(112). 예를 들어, 개인 별 식별 번호에 따라 근무지 또는 신용계좌가 관찰된 횟수(n)를 산출할 수 있고, 이때 중복을 허용한다.

[0033] 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹(subgroup)을 생성하고, 이때 순서변수는 1로 설정할 수 있다(121). 예를 들어, 기록이 1개만 존재하는, 즉, n==1인 개체는 따로 하위 그룹을 생성하고 재현 과정을 수행한다.

[0034] 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고(122), 각각의 개체의 순서변수 별 하위 그룹을 생성하는 한다(123). 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 동일개체에 대한 기록들을 인식한 뒤, 기록들의 종류에 따라 선분이력형태로 변경하고 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 순서변수를 생성한다. 예를 들어, 기록이 2개 이상 존재하는 그룹(n>1)에 대해 각 개인으로 묶은 자료를 근무시작일 또는 신용계좌 생성일 순서에 따라 순서변수(order)를 생성하고, 이때 중복을 허용하지 않는다. 기록이 2개 이상 존재하는 그룹은 각 순서변수 별로 하위 그룹을 만들고 재현 과정을 수행한다.

[0035] 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행한다(130).

[0036] 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 단계에서는 조건부 추론 방식(Conditional Inference Tree; CTREE)을 통해 유의성 테스트를 수행하고(131), 분류 및 회귀 방식(Classification and Regression Tree; CART)을 통해 지니 계수를 산출하며(132), 적대적 생성 신경망을 통해 학습을 수행한다(133).

[0037] 과적합을 피하기 위한 순서변수 선택을 위해 조건부 추론 방식을 통한 유의성 테스트를 수행하고 공변량 선택 후 노드 분할의 과정을 재귀적으로 수행하는 조건부 추론 방식 알고리즘의 모든 시작 부분에서 순열을 검정(permutation test)하며, 과적합을 일으키는 트리기반 분류 및 회귀 방식을 보완하기 위해 조건부 추론 방식을 우선적으로 적용한다.

[0038] 조건부 추론 방식을 우선 적용한 뒤 하위 그룹에 관한 데이터의 분포에 따라 또는 변수의 값 또는 범위가 결측값(NA)이거나 또는 검정통계량을 산출하지 못해 조건부 추론이 실패한 데이터에 대해 과적합 여부에 상관없이 바이너리 분할 방식인 분류 및 회귀 방식을 적용한다.

[0039] 조건부 추론 방식 및 분류 및 회귀 방식을 통해 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 적대적 생성 신경망을 적용한다.

[0040] 모든 하위 그룹에 대한 재현 자료를 생성한 후(141), 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀(142)를 수행하여 재현 결과를 검증한다. 다시 말해, 모든 하위 그룹의 재현이 끝나면 각각의 데이터 프레임(dataframe)을 열(row)로 묶어 GLM 회귀를 수행하여 닷-위스커(dot-whisker) 그래프 및 CI(Confidence Interval) 오버랩 구간의 크기로 재현 결과를 검증한다(143).

[0042] 도 2는 본 발명의 일 실시예에 따른 재현자료 DB 생성 장치의 구성을 나타내는 도면이다.

[0043] 제안하는 재현자료 DB 생성 장치(200)는 자료 입력부(210), 하위 그룹 생성부(220), 재현 자료 생성부(230) 및 재현 자료 검증부(240)를 포함한다.

[0044] 자료 입력부(210)는 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받고, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출한다.

[0045] 하위 그룹 생성부(220)는 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹을 생성한다.

[0046] 하위 그룹 생성부(220)는 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성한다.

[0047] 재현 자료 생성부(230)는 생성된 각각의 하위 그룹에 대하여 유의성 테스트(231), 지니 계수 산출(232) 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)(233)을 통해 재현 과정을 수행한다.

[0048] 재현 자료 검증부(240)는 및 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증하는 단계를 포함한다.

[0049] 먼저 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료를 입력 받을 수 있다. 예를 들어, 종사자 자료(다시 말해, 근무정보) 또는 개인신용정보 자료를 입력 받을 수 있다.

[0050] 이후, 입력된 특정 분포를 따르지 않는 개체 별 시계열 특성을 갖는 모집단의 자료 내에서 개체 별 식별 번호에 따른 기록이 관찰되는 횟수를 산출한다. 예를 들어, 개인 별 식별 번호에 따라 근무지 또는 신용계좌가 관찰된 횟수(n)를 산출할 수 있고, 이때 중복을 허용한다.

[0051] 개체 별 식별 번호에 따른 하나의 기록이 존재하는 경우, 해당 기록에 관한 하위 그룹(subgroup)을 생성하고, 이때 순서변수는 1로 설정할 수 있다. 예를 들어, 기록이 1개만 존재하는, 즉, n==1인 개체는 따로 하위 그룹을 생성하고 재현 과정을 수행한다.

[0052] 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 해당 기록을 개체 별로 묶고 시계열 특성에 따라 순서변수를 생성하고, 각각의 개체의 순서변수 별 하위 그룹을 생성하는 한다. 개체 별 식별 번호에 따른 두 개 이상의 기록이 존재하는 경우, 동일개체에 대한 기록들을 인식한 뒤, 기록들의 종류에 따라 선분이력형태로 변경하고 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 순서변수를 생성한다. 예를 들어, 기록이 2개 이상 존재하는 그룹(n>1)에 대해 각 개인으로 묶은 자료를 근무시작일 또는 신용계좌 생성일 순서에 따라 순서변수(order)를 생성하고, 이때 중복을 허용하지 않는다. 기록이 2개 이상 존재하는 그룹은 각 순서변수 별로 하위 그룹을 만들고 재현 과정을 수행한다.

[0053] 더욱 상세하게는, 하위 그룹 분류 및 재현 방식은 먼저 동일 개체를 구분하고 시계열적 특성을 나타내는 조건변수를 생성한다.

[0054] 각 조건 변수의 분포가 유사하게 재현된 경우에도 개체 단위의 재현이 고려되지 않으면 특정 집단의 평균이나, 특정 기간 내 개체 수, 개체 별 데이터 간 평균 시간(다시 말해, 시계열적) 등의 정보가 보존되지 못한다.

[0055] 따라서 동일개체를 먼저 인식한 뒤, 근무지 또는 신용계좌 종류에 따라 선분이력형태로 변경하고, 각 개체 내 기록이 생성된 순서로 시계열적 특성을 반영하는 조건 변수, 다시 말해 순서변수(order)를 생성한다.

[0056] 순서변수에 따라 하위 그룹을 구분하고 재현자료를 생성하게 되면, 개체 별 최초 신용계좌 생성 또는 최초 근무

기록, 평균 이직 기간 등 함축된 정보까지 보존 가능하다.

[0057] 상기 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 통해 재현 과정을 수행한다.

[0058] 생성된 각각의 하위 그룹에 대하여 유의성 테스트, 지니 계수 산출 및 적대적 생성 신경망을 통해 재현 과정을 수행하는 단계에서는 조건부 추론 방식(Conditional Inference Tree; CTREE)을 통해 유의성 테스트를 수행하고, 분류 및 회귀 방식(Classification and Regression Tree; CART)을 통해 지니 계수를 산출하며, 적대적 생성 신경망을 통해 학습을 수행한다.

[0059] 과적합을 피하기 위한 순서변수 선택을 위해 조건부 추론 방식을 통한 유의성 테스트를 수행하고 공변량 선택 후 노드 분할의 과정을 재귀적으로 수행하는 조건부 추론 방식 알고리즘의 모든 시작 부분에서 순열을 검정(permutation test)한다. 그리디(greedy)하게 데이터를 분류하여 과적합을 일으키는 트리기반 분류 알고리즘(CART) 및 회귀 방식을 보완하기 위해 조건부 추론 방식을 우선적으로 적용한다.

[0060] 조건부 추론 방식을 우선 적용한 뒤 하위 그룹에 관한 데이터의 분포에 따라 또는 변수의 값 또는 범위가 결측값(NA)이거나 또는 검정통계량을 산출하지 못해 조건부 추론이 실패한 데이터에 대해 과적합 여부에 상관없이 바이너리 분할 방식인 분류 및 회귀 방식을 적용한다.

[0061] 트리기반 알고리즘은 바이너리 분할 방식이기 때문에 이상치(outlier) 등의 값을 과대계상하거나, 정보 측정값(Gini impurity)을 최대화하는 변수를 먼저 찾아 노드를 그리디(greedy)하게 분할하여 과적합을 일으킨다.

[0062] 특히, 트리의 노드 선택 순서를 정하려고 모든 특성의 정보획득량을 얻고 정렬하는 과정을 거치기 때문에 클래스의 개수가 많고 데이터가 적을 때 적용하기 힘들며, 훈련과정에서 계산량이 많다.

[0063] 비교적 재현의 성능은 좋으나 위와 같은 이유로 지역적, 시대적 특성이 두드러지는 자료의 재현에 적용하기 힘들다. 특정 그룹은 변수 내 클래스의 수가 개체 별로 모두 다르거나 또는 데이터의 수가 너무 적기 때문이다.

[0064] 따라서, 조건부 추론(CTREE)을 우선 적용한 뒤 데이터의 분포가 단조롭거나 변수의 값 또는 그 범위가 결측값(NA) 또는 하나의 값에 치우쳐져 있어 검정통계량(p-value)을 산출하지 못해 조건부 추론(CTREE)이 실패한 데이터에 한해 과적합에 상관없이 트리기반 분류 알고리즘, 다시 말해 바이너리 분할 방식(CART)을 적용한다.

[0065] 사용자에 의해 미리 정해진 최종 노드의 수에 대해 데이터를 바이너리 분할할 수 없어 지니 불순도 계산이 되지 않는 데이터 분포는 재현이 되지 못한다.

[0066] 조건부 추론 방식 및 분류 및 회귀 방식을 통해 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 적대적 생성 신경망을 적용한다.

[0067] 데이터가 특정 분포를 따르지 않거나, 클래스의 수가 적어도 데이터 분포를 있는 그대로 학습하여 재현 가능하지만, 종사자나 개인신용정보와 같은 다양한 분포를 모두 구분하여 재현하기는 어렵기 때문에 바로 적용할 수 없다.

[0068] 따라서, 조건부 추론(CTREE)과 바이너리 분할(CART) 모두 재현할 수 없어 분포가 한정적으로 정의된 데이터에 대해 제한적으로 적대적 생성 신경망을 사용하면 제한된 분포에 한해 적용 가능하다.

[0069] 조건부 추론 방식의 유의성 테스트에 있어서, 관찰값의 라벨(다시 말해, 분류된 결과)을 재배열하고, 산출한 검정 통계량을 가능한 모두 계산하여 귀무가설(null) 하에서 산출된 검정 통계량(p-value)의 분포를 산출한다.

[0070] 이 중 가장 작은 검정 통계량을 가지는 공변량을 선택하고 순열 검정하는 과정을 반복하여 가장 많은 변환을 제공하는 변수를 선택한다.

[0071] 만약 클래스가 단조롭거나(NA 포함), 또는 개체 별로 너무 다양하여 재배열 가능하지 않은 그룹의 경우 검정 통계량을 산출하지 못하거나, 모든 가능한 재배열에 대한 검정 통계량이 기각된 경우, 조건부 추론 방식을 적용하여 재현하지 못하고 바이너리 분할 방식으로 재현하게 된다.

[0072] 바이너리 분할 지니 불순도(Gini Impurity) 산출식은 다음과 같다:

[0073] $$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2$$

[0074] D: 데이터가 나뉘는 파티션

[0075] p_i: 전체에서 집단에 속하는 관측치의 비율(확률)

[0076] m: 분할되는 클래스의 수

[0077] 위의 식을 간단히 설명하면 1에서 '전체 데이터의 개수 중 각 클래스가 차지하는 개수의 비율'을 제곱하여 빼주는 것이다.

[0078] 연속형 변수에서는 지니불순도가 노이즈 측정값으로써 손실함수로 사용되며, 범주형 변수에서는 지니 불순도의 가중치의 합인 지니 계수(Gini Index)로 산출한다.

[0079] 바이너리 분할 방식에서 지니 불순도가 산출되지 않아 실패한다는 것은 사용자가 미리 정의한 최종 노드의 최소값(minbucket)에 대해 모든 노드 순서에 따라 지니 불순도를 계산하였을 때, 클래스가 너무 다양하여 각 클래스가 차지하는 비율이 너무 작아 지니 불순도의 값이 1보다 더 이상 줄어들지 않거나, 데이터의 수가 너무 작아 노드 분할에 영향력이 없다는 것을 뜻한다.

[0081] 도 3은 본 발명의 일 실시예에 따른 적대적 생성 신경망을 통한 재현 자료 생성 과정을 설명하기 위한 도면이다.

[0082] 본 발명의 일 실시예에 따른 조건부 적대적 생성 신경망(GAN)은 조건부 변수(y)가있는 GAN의 확장으로, 입력 레이어로 판별자(320)와 생성자(310) 모두에 제공되는 추가 정보이다. 이 단계에서는 계정 발생 순서변수(ORDER)를 조건(y)으로 설정하여 시작되며, 각 부분집합에 대해, 조건(y), 랜덤 노이즈(z) 및 랜덤 생성된 조건(y')에 대해 원래 데이터(x)로 판별자(320)를 훈련한다.

[0083] 판별자(320)에 의해 학습된 생성자(310)는 랜덤 노이즈(z)와 그 랜덤 생성된 조건(y')을 입력으로 취하고 이를 처리하여 판별자(320)에 입력값이 되는 가짜 데이터(x' 및 y')를 생성한다. 이후, 판별자(320)에 의해 실수로 예측된 레이블이 지정된 데이터는 다시 원래 데이터의 형태로 변환된다.

[0085] 도 4는 본 발명의 일 실시예에 따른 적대적 생성 신경망의 생성자의 학습 과정을 설명하기 위한 도면이다.

[0086] 도 4와 같이 랜덤 노이즈(z)와 랜덤 생성된 조건 변수(y')는 생성자의 은닉 계층에서 연결되어 가짜 데이터(x')를 생성한다. 생성된 가짜 데이터(x')와 가짜 조건(y')은 판별자의 입력값으로 사용된다.

[0087] 손실함수(Loss function)는 판별자와 생성자 간의 최소-최대 동시 최적화 작업을 실행한다. 범주형과 클러스터링된 연속형 변수가 섞여 있으므로 일반적으로 사용되는 크로스-에트로피(Cross-Entropy) 대신에 KL(Kullback-Leibler) 발산(Divergence)를 사용하여 다음과 같이 상대적인 거리를 반영하였다:

[0088] $$min_G max_D = E_{x\sim p_{X(x)}}[log D(x|y)] + E_{z\sim p_{Z(z)}}[log(1 - D(G(z|y)))] + \sum KL(x', x)$$

[0089] 모든 하위 그룹에 대한 재현 자료를 생성한 후, 각각의 재현 자료에 대한 GLM(General Linear Regression Model) 회귀를 수행하여 재현 결과를 검증한다. 다시 말해, 모든 하위 그룹의 재현이 끝나면 각각의 데이터 프레임(dataframe)을 열(row)로 묶어 GLM 회귀를 수행하여 닷-위스커(dot-whisker) 그래프 및 CI(Confidence Interval) 오버랩 구간의 크기로 재현 결과를 검증한다.

[0090] 제안하는 알고리즘의 성능을 확인하기 위한 방법으로 특정 변수의 효율성(coefficient)에 대한 신뢰구간(CI)의 오버랩 구간의 측정값을 계산하고 평균으로 제공한다. 하기 산출식에 따른 오버랩 구간(Interval-Overlap; IO)이 넓을수록 유용성이 높다고 해석할 수 있다:

[0091] $$IO = 0.5\left[\frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_o - l_o} + \frac{\min(u_o, u_s) - \max(l_o, l_s)}{u_s - l_s}\right]$$

[0092] u_o: 해당변수 원자료의 최대값(upper limit)

[0093] u_o: 해당변수 재현자료의 최대값(upper limit)

[0094] l_o: 해당변수 원자료의 최소값(lower limit)

[0095] l_s: 해당변수 재현자료의 최소값(lower limit)

[0096] 이 측정값은 겹침이 없을 때 음수이며 간격이 더 멀어짐에 따라 감소한다.

[0097] 각 변수의 효율성 신뢰구간의 오버랩 구간의 비율 평균값이 양수이면 두 자료는 통계적인 차이를 추론할 수 없다는 결론을 내릴 수 있고, 만약 음수가 나오면 통계적인 차이를 추론할 수 있으므로, 알고리즘의 성능을 판단하는 척도로 사용한다. 예를 들어, 오버랩 구간은 일반적으로 사용되는 오차범위에 맞추어 95%로 설정할 수 있다. 본 발명의 실시예에 따르면, 상기 산출식을 사용하여 개인신용정보의 각 변수 신뢰구간의 평균값은 .7000921, 즉, 약 70%의 평균 오버랩 비율을 나타내므로 본 재현자료 생성 알고리즘으로 재현한 자료는 원자료와 통계적으로 유의한 차이를 추론할 수 없다는 의미를 가진다.

[0098] 검증 결과는 변수의 재현 순서를 제어하는 방식으로 조절할 수 있으며, 결과의 수용여부는 사용자가 미리 세팅할 수 있으나 통상적으로 양수인 경우 허용하는 편이다.

[0099] 본 발명의 실시예들에 따르면 개인정보보호법으로 인해 제한적으로 사용되는 금융, 고용 및 근로 정보 DB를 타 산업분야에서 융합하여 활용할 수 있도록 재현자료를 생성할 수 있다. 종래 기술의 인구통계학적 자료의 구조와 달리 개인의 금융, 고용 및 근로 정보를 시간 순으로 재현하여 특정 기간의 특징까지 원자료와 유사하도록 재현할 수 있다. 트리기반 알고리즘과 적대적 생성 신경망(Generative Adversarial Networks; GAN)을 사용하여 매스킹 위주의 재현자료 생성으로 원자료의 민감정보를 노출하지 않으면서도 자료의 유용성(utility)을 최대한 보존할 수 있다. 또한, 원자료와 동일한 개인정보를 포함하지 않는 완전재현금융자료를 생성하여 타 산업과 원활하게 융합하여 부가 가치를 창출하고, 금융전문가가 아닌 개발 또는 분석 목적으로 실데이터에 접속하기 전 교육용 자료로 활용 가능하다.

[0101] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0102] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0103] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를

포함한다.

[0104] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0105] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

[0107] <참고 자료>

[0108] -재현자료를 생성하기 위한 또는 재현자료 생성을 목적으로 하는 프로그램 및 알고리즘:

[0109] PopGen (http://simtravel.wikispaces.asu.edu/Home): IPU 알고리즘을 구현한 오픈소스에 관한 것으로, Arizona State University의 SimTRAVEL Research Initiative에서 개발.

[0110] VirtualBelgium (https://sourceforge.net/projects/virtualbelgium/): 인구 통계, 주거 선택, 활동 패턴, 이동성 및 기타 정보를 시뮬레이션하여 벨기에 인구 변화를 관찰하는 프로그램으로, 모집단은 반복 비례 적합 방법(Iterative Proportional Fitting, IPF) 알고리즘을 사용하여 생성되며, 가구 정보는 가구 구성원으로부터 수집하여 만들어짐.

[0111] R 패키지 sms (https://cran.r-project.org/web/packages/sms/index.html): 주어진 영역 내 매크로 데이터로부터 마이크로 데이터를 시뮬레이션하는 기능을 제공하며, 가구를 구성하는 구성원 정보 등 계층적 구조를 이루는 데이터를 다룰 수는 없지만, SA (Simulated Annealing)를 단순화하여 제한된 영역에 대한 설명을 최적화 하는 기능을 제공함.

[0112] MoSeS(https://royalsocietypublishing.org/doi/10.1098/rsta.2009.0041): 영국의 특정 도시 및 지역 시스템을 위한 인구 정보를 재현하여 향후 25년 이후의 인구 정보를 예측할 수 있으며, 유전 알고리즘을 기반으로 구현됨.

[0113] R 패키지 synthpop(https://cran.r-project.org/web/packages/synthpop/index.html): 분류회귀모형을 사용하여 재현데이터에 대한 변수를 생성하며, 가구 내 구성원 정보 등의 계층적 또는 클러스터 같은 복잡한 데이터 구조는 처리할 수 없지만, 데이터 형태에 구애받지 않고 재현자료를 생성할 수 있다는 장점이 있음.

[0114] TRANSIMS(https://sourceforge.net/projects/transimsstudio/): 미국 Los Alamos National Laboratory의 연구원이 개발한 운송 분석 시뮬레이션 시스템으로, 인구조사 마이크로 데이터를 기반으로 IPF기법을 사용하여 재현데이터를 생성함.

[0115] Synthia(http://synthpopviewer.rti.org/): RTI에서 개발한 웹 기반 응용 프로그램으로, 사용자 정의 변수를 사용하여 사용자 정의 학습 영역에 대한 재현데이터를 생성함.

[0116] SMILE(Simulated Model of the Irish Local Economy, O'Donoghue 2014): 아일랜드 전체 인구를 대상으로 미시적 수준의 재현데이터를 생성하는 정적 공간 마이크로 시뮬레이션(static spatial microsimulation)모델로써, 아일랜드의 농업 인구조사(Irish Census of Agriculture)와 아일랜드 국립 농장 조사(NFS)를 결합하여, 농장 단위의 정적 재현데이터를 생산하기 위해 사용된 통계적 결합 기법을 사용함.

[0117] R 패키지 simPop(https://cran.r-project.org/web/packages/simPop/index.html): 주체가 가진 속성 값에 따라 다르게 적용되는 정책의 거시적인 효과를 예측하는 데에 주로 활용되는 마이크로 시뮬레이션을 위한 복잡한 구조의 데이터 재현에 매우 유용함.

[0118] 이 밖에도 온라인에서 재현데이터를 생성할 수 있는 python 웹 기반 응용 프로그램인 Data Synthesizer(https://github.com/DataResponsibly/DataSynthesizer)와 DPSynthesizer(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496798/) 등이 있음.

[0119] -관련 논문 또는 문서:

[0120] Raab, G.M., Nowok, B. and Dibben, C. (2017). Practical data synthesis for large samples. Journal of

Privacy and Confidentiality, 7(3), 67-97.

[0121] Reiter, J.P. (2005). Using CART to generate partially synthetic, public use microdata. Journal of Official Statistics, 21(3), 441-462.

[0122] Gillian M. Raab et al. (2017). Guidelines for Producing Useful Synthetic Data. arVix:1712.04078v1 [stat.AP].

[0123] Drechsler, J. and J. P. Reiter (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, computational statistics and data analysis. Computational Statistics and Data Analysis 55, 3232-3243.

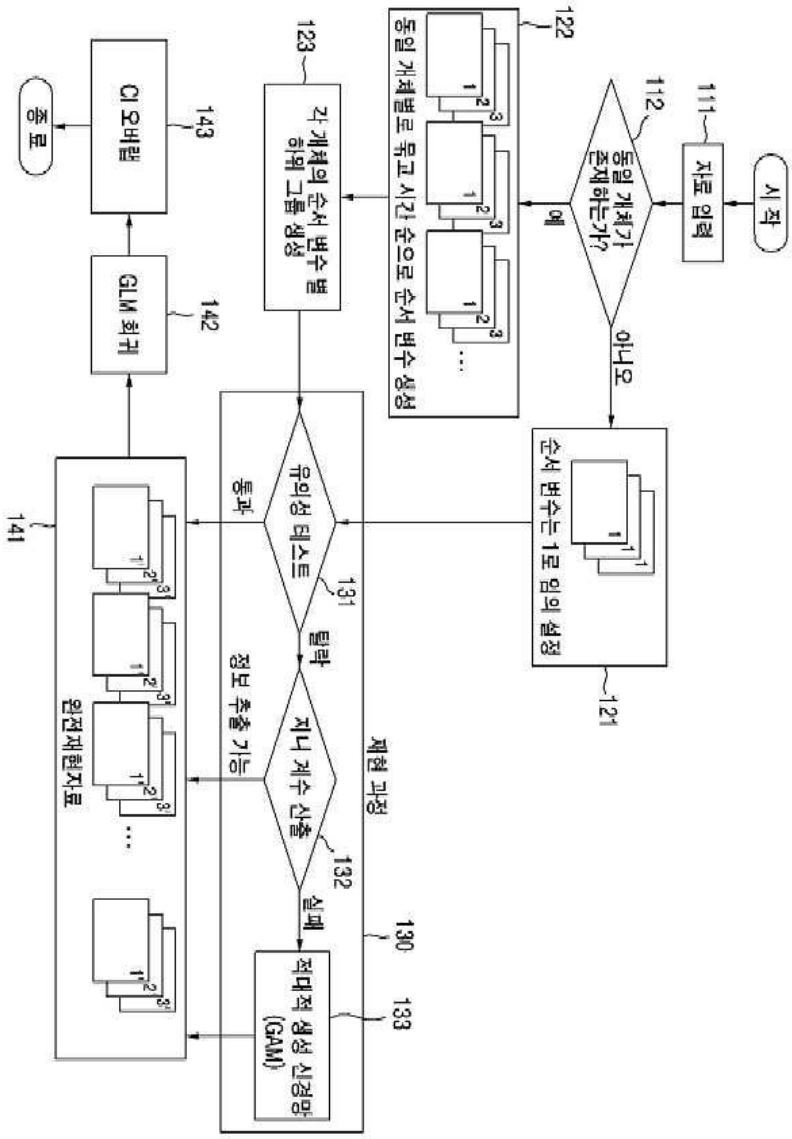
[0124] Joshua Snoke et al. (2018). General and Specific Utility Measures for Synthetic Data. Journal of the Royal Statistical Society, 181, Part3, pp.663-688.

[0125] Ashwini Venkatasubramaniam, et al. (2017). Decision Trees in Epidemiological Research. Emerg Themes Epidemiol.

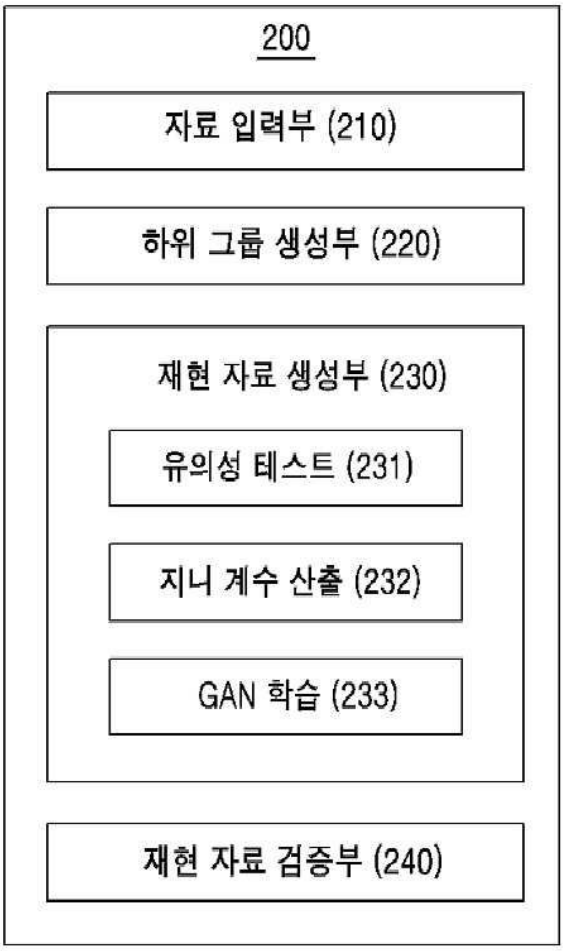
[0126] Beata Nowok, Gillian M. Raab, Chris Dibben (2016). synthpop: Bespoke Creation of Synthetic Data in R. Journal of Statistical Software, 74(11), 1-26. doi:10.18637/jss.v074.i11

도면

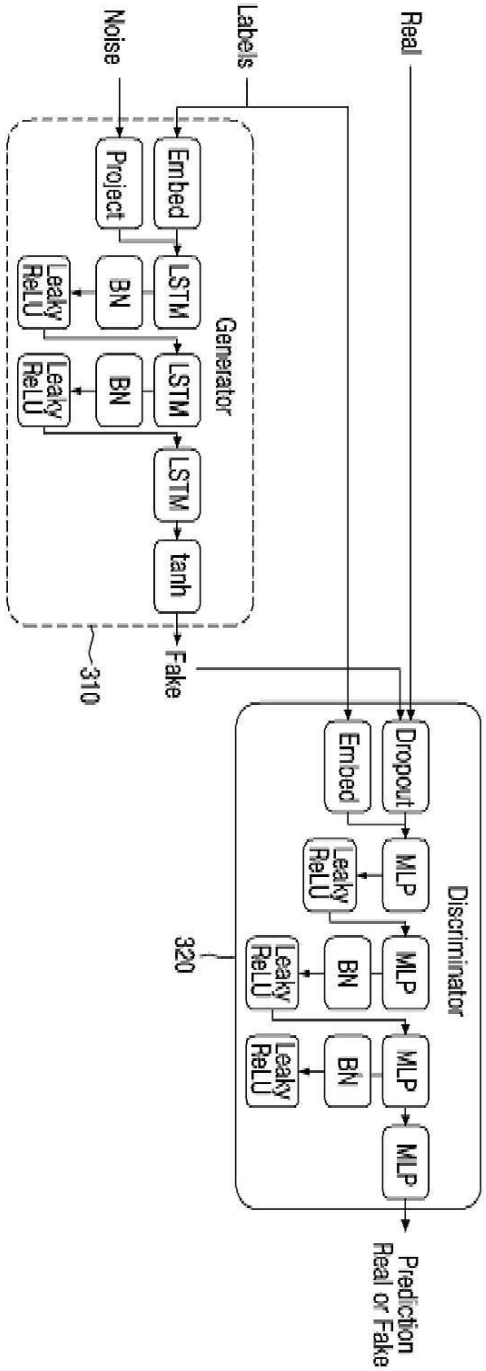
도면1


시 작
111
자료 입력
112
동일 개체가 존재하는가?
예
아니오
122
동일 개체별로 묶고 시간 순으로 순서 변수 생성
121
순서 변수는 1로 임의 설정
123
각 개체의 순서 변수별 하위 그룹 생성
131
유의성 테스트
통과
탈락
132
지니 계수 산출
실패
130
제한 과정
133
적대적 생성 신경망 (GAN)
정보 추출 기능
141
원전제한자료
142
GLM 회귀
143
CI 오버랩
종 료

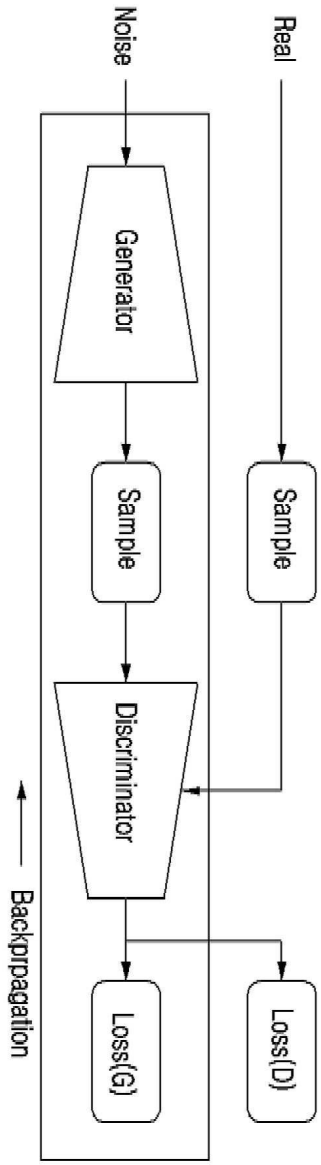
도면2


200
자료 입력부 (210)
하위 그룹 생성부 (220)
재현 자료 생성부 (230)
유의성 테스트 (231)
지니 계수 산출 (232)
GAN 학습 (233)
재현 자료 검증부 (240)

도면3


Noise
Labels
Real
Project
Embed
LSTM
BN
Leaky ReLU
Generator
tanh
310
Fake
Embed
Dropout
MLP
Discriminator
320
Prediction
Real or Fake

도면4


Noise
Real
Generator
Sample
Sample
Discriminator
Backprpagation
Loss(G)
Loss(D)