



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0103494
(43) 공개일자 2020년09월02일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) G06N 3/08 (2006.01)
(52) CPC특허분류
G06N 20/00 (2019.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2019-0022080
(22) 출원일자 2019년02월25일
심사청구일자 2019년02월25일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
양지훈
서울특별시 은평구 백련산로 2길 19, 백련산힐스
테이트 1차 105동 603호
신동일
경기도 성남시 분당구 수내로 174 101동 2001호
(수내동, 푸른마을벽산아파트)
(74) 대리인
이지연

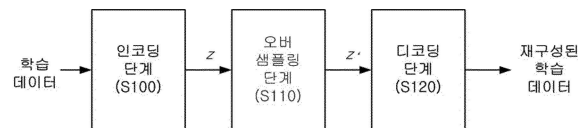
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 학습 데이터에 대한 오버샘플링 방법

(57) 요약

본 발명은 학습 데이터에 대한 데이터 불균형 처리를 위한 오버 샘플링 방법에 관한 것이다. 상기 오버샘플링 방법은, 학습 데이터에 대하여 적대적 오토인코더(AAE)를 적용하여 잠재 변수 공간에 대한 특징들을 추출하는 인코딩 단계; 상기 특징이 추출된 잠재 변수 공간에서의 소수 범주의 데이터들에 대하여 SMOTE를 적용하여 오버샘플링하는 오버샘플링 단계; 상기 오버 샘플링된 잠재 변수 공간에 대하여 디코딩하는 디코딩 단계; 를 구비하여, 학습 데이터를 재구성하여 소수 범주의 데이터와 다수 범주의 데이터의 균형을 맞춘다.

대표도 - 도3



이 발명을 지원한 국가연구개발사업

과제고유번호	2017-0-00498
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기술진흥센터
연구사업명	정보보호핵심원천기술개발
연구과제명	차분 프라이버시 기반 비식별화 기술 개발
기 여 율	1/2
과제수행기관명	서강대학교 산학협력단
연구기간	2017.04.01 ~ 2019.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	2018R1D1A1B07048790
부처명	교육부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구지원사업
연구과제명	GAN 기반 분산 딥러닝 프레임워크를 이용한 개인 정보 보호 데이터 마이닝 연구
기 여 율	1/2
과제수행기관명	서강대학교 산학협력단
연구기간	2018.06.01 ~ 2021.05.31

명세서

청구범위

청구항 1

학습 데이터에 대한 데이터 불균형 처리를 위한 오버 샘플링 방법에 있어서,

- (a) 학습 데이터에 대하여 잠재 변수 공간에 대한 특징들을 추출하는 인코딩 단계;
- (b) 상기 특징이 추출된 잠재 변수 공간에서의 소수 범주의 데이터들에 대하여 오버샘플링하는 오버샘플링 단계;
- (c) 상기 오버 샘플링된 잠재 변수 공간에 대하여 디코딩하는 디코딩 단계;

를 구비하여, 학습 데이터를 재구성하여 소수 범주의 데이터와 다수 범주의 데이터의 균형을 맞춘 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 2

제1항에 있어서, 상기 (a) 단계는

오토인코더의 구조를 갖는 생성 모델(Generative Model)을 이용하여 학습 데이터의 잠재 변수 공간에 대한 특징들을 추출하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 3

제2항에 있어서, 상기 오토인코더의 구조를 갖는 생성 모델(Generative Model)은 변분 인코더(Variational Autoencoder, VAE)와 적대적 생성 네트워크(Generative Adversarial Network, GAN)를 결합한 적대적 오토인코더(Adversarial Auto-Encoder, AAE)를 적용하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 4

제1항에 있어서, 상기 (b) 단계는,

잠재 변수 공간의 소수 범주의 각 샘플에 대하여 사전 설정된 개수(k)의 최근접 이웃을 합성하여 새로운 샘플들을 생성하여, 소수 범주의 샘플들에 대하여 오버 샘플링하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 5

제4항에 있어서, 상기 (b) 단계는,

소수 범주의 한 샘플(x_i)을 선택하고, 사전 설정된 개수(k)의 최근접 이웃으로부터 임의의 이웃($\hat{x}_i$)을 선택하고, 0과 1 사이의 난수(λ)를 생성하고, 아래의 수학적식에 따라 합성 샘플(x_{new})을 생성하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

$$x_{new} = x_i + \lambda (\hat{x}_i - x_i), \quad \lambda \in [0, 1]$$

청구항 6

제1항에 있어서, 상기 (b) 단계는 SMOTE((Synthetic Minority Over-sampling Technique)을 이용하여 오버샘플링하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 7

제1항에 있어서, 상기 (b) 단계는

소수 범주의 데이터와 다수 범주의 데이터가 동일 비율이 될 때까지 반복 수행되는 것을 특징으로 하는 학습 데

이터에 대한 오버 샘플링 방법.

청구항 8

제1항에 있어서, 상기 (c) 단계는

오버샘플링된 소수 범주의 데이터만을 복원하고, 초기의 학습 데이터와 복원된 오버샘플들로 학습 데이터를 재구성하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 9

제1항에 있어서, 상기 (c) 단계는

초기의 학습 데이터와 오버샘플링된 소수 범주의 데이터를 모두 복원하고, 복원된 학습 데이터와 복원된 오버샘플들로 학습 데이터를 재구성하는 것을 특징으로 하는 학습 데이터에 대한 오버 샘플링 방법.

청구항 10

청구항 제1항에 따른 학습 데이터에 대한 오버 샘플링 방법을 구현한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록 매체.

발명의 설명

기술 분야

- [0001] 본 발명은 학습 데이터에 대한 오버샘플링 방법에 관한 것으로서, 더욱 구체적으로는 적대적 오토 인코더를 이용하여 잠재 변수에 대한 특징들을 추출하고, 데이터의 특징이 추출된 잠재 변수 공간에서 SMOTE를 적용하여 오버샘플링하고 다시 디코딩함으로써, 소수 범주 데이터를 오버 샘플링하는 방법에 관한 것이다.

배경 기술

- [0002] 일반적으로 기계학습 실험을 진행하기 위한 학습 데이터로는 범주 별로 데이터의 수가 비슷하게 구성되어 있는 것을 사용한다. 그러나 많은 실제 데이터들은 클래스 별로 데이터의 수가 비슷하지 않아 불균형 문제에 속하게 되고 이러한 경우 소수 범주에 속한 데이터들은 잘못 분류되어 성능이 낮게 나올 가능성이 높다. 이러한 부작용은 기계학습 알고리즘의 설계 특성상 각 범주의 상대적인 분포를 고려하는 대신 전반적인 성능을 최적화시키려 하기 때문에 발생하는 것이다.
- [0003] 불균형 데이터 문제를 해결하기 위한 방법으로는 크게 데이터 수준, 알고리즘 수준 및 앙상블 접근 방법이 있다. 데이터 수준의 접근 방법은 언더샘플링(Under-sampling), 오버샘플링(Over-sampling)이나 이들을 동시에 사용하여 데이터의 균형을 조정하는 방법이다. 알고리즘 수준의 접근 방법은 기존 기계학습 알고리즘의 오차함수를 수정 및 비용 개념을 사용해 소수 군집에 속한 데이터에 더 큰 중요도를 주는 방법이다. 앙상블 접근 방법은 낮은 성능의 분류기 여러 개를 종합하여 최종 분류기의 성능을 개선시키는 방법이다. 특히, 기계학습 알고리즘과 독립적인 데이터 수준의 접근 방법에서 많은 연구가 진행되어왔다. 언더샘플링 방법은 다수 범주에 속한 데이터를 임의적으로 표본 추출하여 소수 범주와 균형을 이루도록 하는 방법이며, 오버샘플링 방법은 소수 범주에 속한 데이터를 반복적으로 복사하여 다수 범주와 균형을 이루도록 하는 방법이다.
- [0004] 대표적인 오버샘플링 방법으로 SMOTE(Synthetic Minority Over-Sampling Technique)가 있다. SMOTE는 소수 범주의 샘플들을 중심으로 최 근접 이웃 k개(k-Nearest Neighbor)를 합성하여 새로운 샘플을 생성하는 방법이다. 도 1은 SMOTE를 활용하여 합성 샘플을 생성하는 방법을 설명하는 그래프들이다. 도 1의 (a)는 소수 범주의 샘플($\hat{x}_i$)에 대하여 최 근접 이웃($\hat{x}_i$)을 선택한 것이고, (b)는 $\hat{x}_i$ 와 $\hat{x}_i$ 사이의 임의의 난수를 생성하여 합성 샘플을 생성하는 것을 나타낸다. 이 방법을 소수 범주의 샘플이 다수 범주가 원하는 균형 비율에 도달할 때까지 무작위로 선택된 소수 범주에 대해 수행한다. 따라서 소수 범주를 복제하는 것이 아니라 새로운 샘플을 합성하기 때문에 과적합의 문제가 발생되지 않는데 도움이 된다.
- [0005] 한편, 오토인코더(Auto-Encoder)는 비지도 학습으로 입력 값과 출력 값이 같아지도록 학습을 한다. 도 2는 오토인코더의 구조를 개념적으로 도시한 모식도이다. 도 2를 참조하면, 오토인코더는 입력 데이터 X 로 구성된 입력 층과 변형된 특징을 나타내는 ‘code’를 가리키는 은닉 층, 그리고 입력 층을 복원한 데이터 X' 을 나타내는 출

력 층으로 구성되어 있다. 오토인코더의 목적은 입력 데이터와 가장 유사한 복원 데이터를 만드는 신경망을 구성하는 것이다. 즉, 오토인코더의 은닉 층을 입력 층의 인코더를 통해 뉴런수보다 작게 만들어 차원을 축소하고, 축소된 차원의 데이터를 디코더를 통해 복원함으로써, 복원 데이터와 원본 데이터의 차이를 최소화한다. 이렇게 학습을 진행하면 오토인코더는 은닉 층에서 입력 데이터의 압축된 정보를 해석할 수 있으며 차원의 축소가 가능하다. 즉, 은닉 층의 뉴런 수 조절을 통해 데이터 특징 추출할 수 있고 목적에 따라 특징을 변형시킬 수 있다. 또한, 축소된 공간에서 분류도 수행할 수 있다.

[0006] 오토인코더의 구조를 갖는 생성 모델(Generative Model)의 중 하나인 적대적 오토인코더(Adversarial Auto-Encoder, AAE)를 통해 잠재 변수 공간에서 특징을 추출한다. 오토인코더는 입력 변수와 잠재 변수 사이에 상관관계가 없다는 단점이 있어 이를 보완한 방법으로 인코더와 디코더를 확률 모델로 만들어 잠재 변수가 학습 데이터와 비슷한 확률 분포를 갖게 하여 더 나은 특징을 추출할 수 있다.

선행기술문헌

특허문헌

[0007] (특허문헌 0001) 한국등록특허공보 제 10-1563406호
(특허문헌 0002) 한국등록특허공보 제 10-1843074호
(특허문헌 0003) 미국공개특허공보 US 2016/0092789

발명의 내용

해결하려는 과제

[0008] 전술한 문제점을 해결하기 위한 본 발명의 목적은 학습 데이터에 대해 적대적 오토 인코더를 적용하여 잠재 변수 공간에서 특징을 추출하고, 특징이 추출된 잠재 변수 공간에서 소수 범주 데이터를 오버샘플링함으로써, 학습 데이터의 소수 범주 데이터와 다수 범주 데이터의 불균형을 해소할 수 있도록 하는 학습 데이터에 대한 오버샘플링 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 전술한 기술적 과제를 달성하기 위한 본 발명의 특징에 따른 학습 데이터에 대한 데이터 불균형 처리를 위한 오버 샘플링 방법은, 학습 데이터에 대하여 잠재 변수 공간에 대한 특징들을 추출하는 인코딩 단계; 상기 특징이 추출된 잠재 변수 공간에서의 소수 범주의 데이터들에 대하여 오버샘플링하는 오버샘플링 단계; 상기 오버 샘플링된 잠재 변수 공간에 대하여 디코딩하는 디코딩 단계; 를 구비하여, 학습 데이터를 재구성하여 소수 범주의 데이터와 다수 범주의 데이터의 균형을 맞춘다.

[0010] 전술한 특징에 따른 학습 데이터에 대한 오버 샘플링 방법에 있어서, 상기 (a) 단계는 오토인코더의 구조를 갖는 생성 모델(Generative Model)을 이용하여 학습 데이터의 잠재 변수 공간에 대한 특징들을 추출하는 것이 바람직하며, 상기 오토인코더의 구조를 갖는 생성 모델(Generative Model)은 변분 인코더(Variational Autoencoder, VAE)와 적대적 생성 네트워크(Generative Adversarial Network, GAN)를 결합한 적대적 오토인코더(Adversarial Auto-Encoder, AAE)를 적용하는 것이 더욱 바람직하다.

[0011] 전술한 특징에 따른 학습 데이터에 대한 오버 샘플링 방법에 있어서, 상기 (b) 단계는, SMOTE((Synthetic Minority Over-sampling Technique)을 이용하여 잠재 변수 공간의 소수 범주의 각 샘플에 대하여 사전 설정된 개수(k)의 최근접 이웃을 합성하여 새로운 샘플들을 생성하여, 소수 범주의 샘플들에 대하여 오버 샘플링하는 것이 바람직하며,

[0012] 상기 (b) 단계는, 소수 범주의 한 샘플(x_i)을 선택하고, 사전 설정된 개수(k)의 최근접 이웃으로부터 임의의 이웃($\hat{x}_i$)을 선택하고, 0과 1 사이의 난수(λ)를 생성하고, 이를 이용하여 합성 샘플(x_{new})을 생성하는 것이 바람직하다.

[0013] 전술한 특징에 따른 학습 데이터에 대한 오버 샘플링 방법에 있어서, 상기 (b) 단계는 소수 범주의 데이터와 다

수 범주의 데이터가 동일 비율이 될 때까지 반복 수행되는 것이 바람직하다.

[0014] 전술한 특징에 따른 학습 데이터에 대한 오버 샘플링 방법에 있어서, 상기 (c) 단계는 오버샘플링된 소수 범주의 데이터만을 복원하고, 초기의 학습 데이터와 복원된 오버샘플들로 학습 데이터를 재구성하거나, 초기의 학습 데이터와 오버샘플링된 소수 범주의 데이터를 모두 복원하고, 복원된 학습 데이터와 복원된 오버 샘플들로 학습 데이터를 재구성하는 것이 바람직하다.

발명의 효과

[0015] 전술한 구성을 갖는 본 발명에 따른 오버 샘플링 방법은 적대적 오토 인코더(AAE)와 SMOTE를 결합하여 데이터 불균형 문제를 해결할 수 있게 된다.

[0016] 본 발명에 따른 오버 샘플링 방법은 적대적 오토인코더 학습을 통해 데이터의 잠재 변수 공간에서 특징 추출을 잘 해낸 경우, 클래스별로 데이터의 분포가 확실하게 나뉠 수 있게 된다. 또한, 본 발명에 따라 데이터가 확실하게 분류된 잠재 변수 공간에서 SMOTE를 적용하여 소수 범주 데이터를 오버샘플링하여 소수 범주와 다수 범주의 데이터를 동일한 비율로 맞추고 SVM 기계학습 모델에 적용함으로써, 종래의 다른 오버샘플링 기법들보다 더 우수한 성능을 보여준다.

[0017] 또한, 본 발명에 따라 오버샘플링하여 학습 데이터를 재구성하고, 재구성된 학습 데이터를 기계 학습 모델에 적용함으로써, 종래의 방법들에 비하여 정확도 및 정확도에 대한 신뢰도를 향상시킬 수 있게 된다.

[0018] 본 발명에 따른 방법과 기존의 오버샘플링 방법의 성능을 확인하기 위한 비교 실험을 표 1과 같은 조건으로 진행하였다. 아홉 가지의 데이터를 사용하였으며, 기계학습 모델은 SVM을 사용하여 정확도를 도출하였으며, 5-fold cross validation을 사용하여 정확도에 대한 신뢰도를 높였다. 그리고, 다양한 성능 평가 척도를 통해 비교 실험을 진행하였는데, 불균형 데이터의 성능 측정을 할 경우 예측 정확도(Accuracy)의 경우 불균형의 비율이 크다면 다수 범주 데이터만을 예측해도 정확도가 높게 나오기 때문에 신뢰도가 낮은 편이다. 그러나 AUC와 F1-Score의 경우 불균형 여부에 관계없이 성능 측정이 가능해 이 두 가지의 성능 평가 척도의 정확도가 더 신뢰성을 갖는다.

표 1

오버샘플링 기법	기계학습 모델	성능 검증 방법	성능 평가 척도
제안 기법	SVM	5-fold cross validation	Accuracy
Random			AUC
SMOTE			
Borderline-SMOTE			F1-Score
ADASYN			

[0019]

[0020] 도 11은 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 사용된 데이터 셋(Data Set)을 도시한 도표이다. 도 11을 참조하면, UCI repository 데이터 총 아홉가지를 사용하였으며, 데이터의 설명은 다음과 같다. (1) 위스콘신 유방암 데이터: 유방암 진단 관련 속성과 양성, 음성에 대한 클래스를 갖는다. (2) 전리층 데이터: 전리층에 대한 레이더 관측치를 모은 데이터로 전리층의 패턴 유무에 대한 클래스를 갖는다. (3) 피마족 인디언 데이터: 피마족 인디언 당뇨 발병에 대한 데이터로 당뇨 발병에 대한 속성과 당뇨병 발병 유무의 클래스를 갖는다. (4) 심장 질환 발병 데이터: 심장 질환 발병에 대한 데이터로 심장 질환 발병에 대한 속성과 심장 질환 발병 유무의 클래스를 갖는다. (5) 유리 데이터 : 유리의 종류를 분별하기 위한 데이터로 유리 속성과 건물 유리 그 외의 유리들로 두 가지의 클래스를 갖는다. (6) 차량 형태 데이터 : 차량의 형태를 분별하기 위한 데이터로 차량 형태의 속성과 밴 차량의 형태와 그 외의 차량 형태로 두 가지 클래스를 갖는다. (7) 갑상선 질환 데이터 : 갑상선 질환 발병에 대한 데이터로 갑상선 질환 관련 속성과 발병 유무를 클래스로 갖는다. (8) 숫자 분할 데이터 : 숫자의 분할 값을 분별하기 위한 데이터로 숫자 0과 나머지 1~8 숫자에 대한 두 가지 클래스를 갖는다. (9) 와인 종류 데이터 : 와인 종류에 대한 데이터로 와인 종류에 대한 속성과 세 가지 클래스를 갖는다.

- [0021] 본 발명에서는 성능 평가 척도인 예측 정확도(Accuracy), AUC, F1-Score를 사용하여 도출한 정확도 결과를 설명한다. 전술한 아홉 가지의 데이터 셋을 사용한 실험 결과로, 본 발명에 따른 방법과 기존의 오버샘플링 기법을 비교 실험 대상으로 사용하였다. 모든 실험은 5-fold cross validation을 사용하여 각각의 학습/테스트 데이터 셋에 대한 정확도의 평균을 사용하였다. 각 도표에서 제일 높은 성능은 굵은 글씨와 밑줄로 표시되어 있다.
- [0022] 도 12는 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 예측 정확도 결과를 도시한 도표이다. 도 12를 참조하면, 아홉 가지의 데이터 셋 중 여섯 가지의 데이터 셋이 기존 기법보다 제안 기법을 사용하였을 때 더 분류 정확도가 높았다. 하지만 예측 정확도는 단순히 참, 거짓만을 판별하기 때문에 대체적으로 불균형 문제를 해결하지 않고 기존 데이터(Original)만을 사용하였을 때, 세 가지 데이터 셋이 가장 높은 분류 정확도를 갖고 나머지 데이터 셋에 대해서도 대체적으로 분류 정확도가 높은 편이다.
- [0023] 도 13은 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 AUC 결과를 도시한 도표이다. 도 13을 참조하면, 아홉 가지의 데이터 셋 중 여덟 가지의 데이터 셋이 기존 기법보다 더 나은 결과를 도출하였다. ROC 커브의 민감도(Sensitivity)와 특이도(Specificity)를 반영하기 때문에 데이터 불균형 여부와 관계없이 정확도 도출이 가능하여 기존 데이터에 대해서는 분류 정확도가 가장 높은 것이 없었다.
- [0024] 도 14는 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 F1-Score 결과를 도시한 도표이다. 도 14를 참조하면, 아홉 가지 데이터 셋 중 여덟 가지의 데이터 셋이 기존 기법보다 정확도가 더 나은 결과를 도출하였다. F1-Score 또한 재현율(Recall)와 정밀도(Precision)의 조화평균으로 값을 도출하기 때문에 기존 데이터만을 사용하였을 때에는 전반적으로 낮은 정확도를 도출하였다.

도면의 간단한 설명

- [0025] 도 1은 SMOTE를 활용하여 합성 샘플을 생성하는 방법을 설명하는 그래프들이다.
- 도 2는 오토인코더를 도시한 구조도이다.
- 도 3은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법을 순차적으로 도시한 순서도이다.
- 도 4는 변분 오토인코더(VAE)를 도시한 구조도이다.
- 도 5는 적대적 생성 신경망(GAN) 모델을 도시한 구조도이다.
- 도 6은 적대적 오토인코더(AAE)를 도시한 구조도이다.
- 도 7은 본 발명에 따른 오버샘플링 방법에 있어서, 잠재 변수 공간에서의 특징 추출을 위한 적대적 오토인코더 모델에 대한 학습 과정을 예시적으로 도시한 알고리즘이다.
- 도 8은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 오버샘플링 단계 및 디코딩 단계에 의해 잠재 변수 공간에서 SMOTE를 활용하여 오버샘플링한 후 디코딩하는 과정을 도시한 도면이다.
- 도 9는 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 디코딩 단계의 일 실시 형태를 도시한 블록도이다.
- 도 10은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 디코딩 모듈의 다른 실시 형태를 도시한 블록도이다.
- 도 11은 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 사용된 데이터 셋(Data Set)을 도시한 도표이다.
- 도 12는 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 예측 정확도 결과를 도시한 도표이다.
- 도 13은 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 AUC 결과를 도시한 도표이다.
- 도 14는 본 발명과 기존 방법들의 성능을 비교 확인하기 위한 실험에 있어서, 모든 데이터 셋에 대한 F1-Score 결과를 도시한 도표이다.

발명을 실시하기 위한 구체적인 내용

- [0026] 본 발명에 따른 오버샘플링 방법은 학습 데이터에 대해 적대적 오토인코더를 적용하여 잠재 변수 공간에서의 특징을 추출하고, 특징이 추출된 잠재 변수 공간에서 SMOTE를 적용하여 소수 범주 데이터를 오버샘플링한 후 복원하여 학습 데이터를 재구성함으로써, 데이터 불균형을 해결하는 것을 특징으로 한다. 본 발명은 데이터의 특징이 추출이 잘된 잠재변수 공간이라면 소수 범주의 데이터 분포와 다수 범주의 데이터 분포는 명확한 구분을 띠기 때문에 더 정확한 오버샘플링을 할 수 있을 것이라는 가정을 바탕으로 한다.
- [0027] 이하, 첨부된 도면을 참조하여 본 발명의 바람직한 실시예에 따른 오버샘플링 방법의 구성 및 동작에 대하여 구체적으로 설명한다.
- [0028] 도 3은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법을 순차적으로 도시한 순서도이다. 본 발명에 따른 오버샘플링 방법은 컴퓨터 등과 같은 데이터 처리 장치의 프로세서(processor) 등에 의해 실행될 수 있는 프로그램으로 구현될 수 있으며, 인코딩 단계(S100), 오버샘플링 단계(S110) 및 디코딩 단계(S120)를 구비한다. 이하, 전술한 오버샘플링 방법을 구성하는 각 단계의 동작들에 대하여 구체적으로 설명하도록 한다.
- [0029] 상기 인코딩 단계(S100)는 입력된 학습 데이터들에 대하여 인코더를 적용하여 차원이 축소된 잠재 변수 공간에 대한 특징들을 추출한다. 잠재 변수 공간에 대한 특징을 추출하기 위한 인코더로는 적대적 오토인코더(Adversarial Auto-Encoder, AAE) 사용하는데, 적대적 오토인코더는 생성 모델(Generative Model)로서 데이터를 만들어 내기 위한 목적으로 잠재변수를 활용한다.
- [0030] 상기 적대적 오토인코더(Adversarial Autoencoder, AAE)는 오토인코더의 구조를 갖는 생성 모델(Generative Model)로서, 변분 오토인코더(Variational Autoencoder, VAE)와 적대적 생성 네트워크(Generative Adversarial Network, GAN)을 결합한 구조이다.
- [0031] 상기 변분 오토인코더(Variational Autoencoder, VAE)는 생성 모델의 일종이다. 생성 모델은 잠재 변수(z)가 주어졌을 때의 조건부 확률 $p(x|z)$ 를 모델링하는 것이다. VAE는 (x)의 특징을 잘 찾아내는 잠재변수(z)를 찾아내는데 목적이 있다. 도 4는 변분 오토인코더를 도시한 구조도이다. 도 4를 참조하면, 인코더를 통해 입력 데이터(x)를 받아서 잠재 변수(z)를 얻은 확률 $q(z|x)$ 을 만들어 내고, 디코더를 통해 인코더가 만든 잠재 변수(z)를 활용해 입력 데이터(x)를 복원해낼 $p(x|z)$ 를 만들어낸다.
- [0032] 디코더는 잠재 변수(z)로부터 입력 데이터(x)를 만들어내는 신경망으로 입력데이터(x)의 확률 분포 $p(x)$ 를 알기 위해 디코더는 $p(x|z)$ 를 학습한다. 인코더는 입력데이터(x)로부터 잠재 변수(z)를 얻는 신경망으로 무작위적인 잠재 변수(z)를 샘플링 하는 것이 아니라 입력 데이터(x)를 잘 생성할 수 있는 잠재 변수(z)를 생성하는 것이 목적이다. VAE에서는 이러한 잠재 변수(z)가 정규분포를 따르도록 학습한다. 하지만, 입력 데이터(x)가 주어졌을 때 잠재 변수(z)의 확률 분포 $p(z|x)$ 는 사후 분포로써 이 값을 직접 계산하기는 어렵다. 그래서 이 문제를 해결하기 위해 변분 추론(Variational Inference) 방법으로, 계산하기 힘든 $p(z|x)$ 대신 $q(z|x)$ 를 도입해 $p(z|x)$ 로 근사시키는 방법을 사용한다.

수학식 1

[0033]
$$\log p(x) \geq L(x)$$

수학식 2

[0034]
$$L(x) = -D_{KL}(q_{\phi}(z|x) \| p(z)) + E_{q_{\phi}(z|x)}[\log p(x|z)]$$

- [0035] 수학식 1은 확률 분포인 $p(x)$ 의 계산 편의를 위해 $\log p(x)$ 를 대신 풀어보면 $\log p(x)$ 의 하계(lower bound) 값인 $L(x)$ 를 구할 수 있다. 수학식 2의 우변의 식은 복원 오차 값을 구하는 것이고, 좌변의 식은 정규화 하는 부분으

로 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)를 사용한다. KLD는 잠재 변수(z)를 샘플링 하는 $q_\phi(z|x)$ 와 정규분포인 $p(z)$ 의 거리를 최소화하는 역할을 한다. 여기서 θ , ϕ 는 $p(z|x)$, $q(z|x)$ 의 변수로 사용된다. 그래서 θ , ϕ 를 조절하여 $L(x)$ 를 최대화하는 점을 찾으면서 $\log p(x)$ 와 $L(x)$ 가 같아지도록 한다.

[0036] 한편, 적대적 생성 신경망(Generative Adversarial Network, GAN)은 생성자(Generator)와 판별자(Discriminator)의 역할을 하는 두 개의 모델로 구성되어 있다. 도 5는 적대적 생성 신경망(GAN) 모델을 도시한 구조도이다. 도 5를 참조하면, 판별자는 실제 데이터만을 참으로 판단하기 위해 힘쓰고, 생성자는 판별자가 거짓으로 판별하지 못하도록 가짜 데이터를 생성해가며 두 모델의 성능을 같이 올라가게 되고 최종적으로는 판별자가 실제 데이터와 가짜 데이터를 구분하지 못하게 된다.

[0037] 수학적 3은 GAN의 목적함수이다.

수학적 3

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))]$$

[0038]

[0039] 수학적 3에서, 먼저 판별자(D)가 $V(D, G)$ 를 최대화 하는 관점에서 보면 $x \sim p_{data}(x)$ 의 확률 분포이고 x 는 그 중 샘플링 데이터이다. 판별자(D)는 실제 데이터가 들어오면 1에 가깝게 확률을 추정하여 출력하고, 생성자(G)가 만들어낸 가짜 데이터가 들어오면 0에 가깝게 확률을 추정하여 출력한다. 따라서 $\log$ 를 사용했기 때문에 실제 데이터라면 최대값인 0에 가까운 값이 나오고, 가짜 데이터라면 무한대로 발산하게 하기 때문에 $V(D, G)$ 를 최대화하는 방향으로 학습하게 된다. 생성자(G) 부분인 오른쪽 부분이 $V(D, G)$ 를 최소화하는 관점에서 보면, $z \sim p_z(z)$ 는 보통 정규분포로 사용하는 임의의 노이즈 분포이고 z 는 노이즈 분포에서 샘플링한 것이다. 이 입력을 생성자(G)에 넣어 만든 데이터를 판별자 (D)가 진짜로 판별하면 $\log(1 - D(G(z)))$ 이기 때문에 $D(G(z))$ 는 1이며, D 가 가짜로 판별하면 $D(G(z))$ 는 0이기 때문에 $\log(1 - D(G(z)))$ 는 0에 가까운 최대값이 나오게 될 것이다. 따라서 생성자(G)는 $V(D, G)$ 를 최소화하는 방향으로 학습하게 된다. 즉 $V(D, G)$ 에 있어서 G 는 이를 최소화하는 방향으로, D 는 최대화하는 방향으로 간다

[0040] 도 6은 적대적 오토인코더(AAE)를 도시한 구조도이다. 전술한 바와 같이, 적대적 오토인코더(AAE)는 VAE와 GAN을 결합한 구조로서, VAE 구조가 GAN에서의 생성자 역할을 한다. 생성자의 인코더는 데이터(x)를 받아서 잠재 변수(z)를 샘플링하고, 생성자의 디코더는 이로부터 다시 잠재 변수(x)를 복원한다. AAE가 기존 VAE와 다른 점은 GAN의 판별자 역할을 하는 네트워크가 추가되어, 생성자의 인코더가 샘플링한 가짜 잠재 변수(z)와 $p(z)$ 로부터 직접 샘플링한 진짜 잠재 변수(z)를 구분하는 역할을 한다.

[0041] 이렇게 복잡한 네트워크를 만든 이유는 VAE 단점 때문이다. VAE는 사전확률 분포 $p(z)$ 를 표준정규분포로 가정하고 $q(z|x)$ 를 이와 비슷하게 맞추는 과정에서 학습이 이루어진다. VAE 구조가 이처럼 구성되어 있는 이유는 표준정규분포와 같이 단순한 확률 함수여야 샘플링에 용이하고, KLD 계산을 쉽게 할 수 있기 때문이다. 그런데 실제 데이터 분포가 정규분포를 따르지 않거나 이보다 복잡할 경우 VAE 성능에 문제가 발생할 수 있다.

[0042] 하지만 GAN의 경우 모델에 특정 확률분포를 전제할 필요가 없이 데이터가 어떤 분포를 따르든, 데이터의 실제 분포와 생성자가 만들어내는 분포 사이의 차이를 줄이도록 학습되기 때문이다. VAE의 수학적 2의 좌변의 정규화 부분을 GAN의 손실로 대체할 경우 사전확률과 사후확률 분포를 정규분포 이외에 다른 분포를 쓸 수 있게 되어 모델 선택의 폭이 넓어지는 효과를 누릴 수 있다.

[0043] AAE의 학습 과정은 오토인코더, 판별자, 생성자 순으로 진행된다. 먼저, 수학적 4와 같이 오토인코더 학습을 진행하는데, 입력 데이터(x)에 대해 복원 오차를 구해 인코더 파라미터(ϕ)와 디코더 파라미터(θ)를 업데이트 한다. 그리고 수학적 5의 목적 분포인 z 와 인코더를 통해 샘플링된 $q_\phi(x;)$ 판별자를 통해 학습을 진행하여 판별

자 파라미터(λ)를 업데이트한다. 마지막으로 수식식 6과 같이 생성자 학습을 통해 인코더 파라미터(ϕ)를 업데이트한다.

수식식 4

$$L_i(\phi, \theta, x_i) = -E_{q_\theta(z|x_i)}[\log(p_\theta(x_i|z))]$$

수식식 5

$$-V_i(\phi, \lambda, x_i, z_i) = -\log d_\lambda(z_i) - \log(1 - d_\lambda(q_\phi(x_i)))$$

수식식 6

$$-V_i(\phi, \lambda, x_i, z_i) = -\log(d_\lambda(q_\phi(x_i)))$$

전술한 바와 같이, 본 발명에 따른 오버샘플링 방법은 각 데이터의 특성에 맞게 AAE 모델을 학습시켜 잠재 변수 공간에서 데이터의 특징을 잘 추출해야 된다. AAE 학습 모델을 구축하기 위하여 다양한 파라미터의 결정이 필요하며, 결정해야 될 파라미터로는 은닉층의 수, 은닉층의 뉴런의 수, 활성화함수 등이 있다.

도 7은 본 발명에 따른 오버샘플링 방법에 있어서, 잠재 변수 공간에서의 특징 추출을 위한 적대적 오토인코더 모델에 대한 학습 과정을 예시적으로 도시한 알고리즘이다. 도 7을 참조하면, 적대적 오토인코더 모델의 학습을 위하여, 파라미터 설정 과정, AAE 학습 및 특징 추출 과정을 알 수 있다.

적대적 오토인코더의 학습을 위해, 아홉 가지의 데이터 셋을 사용하여 파라미터 변경을 해가며 실험을 진행해본 결과, 가장 높은 성능을 내는 최적의 파라미터들을 표 2를 통해 나타냈다. 표 2는 도 11에 설명된 각 데이터 셋에 대한 최적의 파라미터들을 도시한 도표이다.

표 2

데이터 셋	레이어의 수		뉴런의 수		잠재변수의 차원
	인코더	디코더	인코더	디코더	
Breast	2	2	9	9	2
Ionosphere	3	3	20	20	2
Pima	2	2	5	5	2
Heart	2	2	10	10	2
Glass	2	2	9	9	2
Vehicle	3	3	15	15	2
Thyroid	2	2	5	5	2
Segment	2	3	15	15	2
Wine	2	2	10	10	2

[0051] 표 1을 참조하면, 오토인코더의 인코더와 디코더는 레이어의 수, 뉴런의 수는 동일하게 사용하였다. 실험에 사용한 데이터의 수, 속성의 수가 크지 않기 때문에 레이어의 수는 2개, 3개, 4개로 진행하였고, 각 레이어의 뉴런의 수는 데이터 속성의 수보다 작게 설정하였다. 또한 상대적으로 단순한 데이터들을 사용하여서 잠재 변수 차원도 이차원으로 고정하여 실험을 진행하였고, 목적 분포는 정규분포를 사용하였다. 활성화함수로는 ReLU를 사용하였는데, 이 함수는 Sigmoid 함수의 기울기 값이 사라지는 문제를 보완하여 최근 가장 많이 사용되고 있는 함수이다. 그리고 학습 데이터의 과적합 방지를 위한 dropout을 사용하였다.

[0052] 도 8은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 오버샘플링 단계 및 디코딩 단계에 의해 잠재 변수 공간에서 SMOTE를 활용하여 오버샘플링한 후 디코딩하는 과정을 도시한 도면이다.

[0053] SMOTE(Synthetic Minority Over-Sampling Technique)는 소수 범주의 샘플들을 중심으로 최 근접 이웃 k개(k-Nearest Neighbor)를 합성하여 새로운 샘플을 생성한다. 합성 샘플(x_{new})은 수학적 식 7을 통해 얻을 수 있다. 소수 범주의 샘플(x_i)에 대하여 최 근접 이웃($\hat{x}_i$)을 선택하고 x_i 와 $\hat{x}_i$ 사이의 임의의 난수(λ)를 생성하여 수학적 식 7에 따라 합성 샘플을 생성하게 된다. 이 방법을 소수 범주의 샘플이 다수 범주가 원하는 균형 비율에 도달할 때까지 무작위로 선택된 소수 범주에 대해 수행한다.

수학적 식 7

$$x_{new} = x_i + \lambda (\hat{x}_i - x_i), \quad \lambda \in [0, 1]$$

[0054]

[0055] 도 8 및 도 2를 참조하면, 본 발명에 따른 오버샘플링 방법의 상기 오버샘플링 단계(S110)는 인코딩 단계에서 최종적으로 학습을 마친 학습 데이터의 잠재 변수 공간에서 SMOTE 알고리즘을 적용하여, 소수 범주의 데이터를 오버샘플링함으로써, 소수 범주의 데이터를 다수 범주의 데이터와 동일한 비율로 맞춘다.

[0056] 도 8 및 도 2를 참조하면, 본 발명에 따른 오버샘플링 방법의 상기 디코딩 단계(S120)는 오버샘플링 단계에 의해 소수 범주의 데이터와 다수 범주의 데이터를 균형을 맞춘 잠재 변수 공간에서의 특징들(z')을 디코딩하여 데이터들을 복원한다.

[0057] 이러한 과정을 통해, 학습 데이터들을 재구성하게 되며, 재구성된 학습 데이터를 기반으로 기계 학습을 함으로써 정확도를 향상시킬 수 있게 된다.

[0058] 한편, 상기 디코딩 단계(S120)는 두 가지의 실시 형태가 있다. 도 9는 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 디코딩 단계의 일 실시 형태를 도시한 블록도이다. 도 9를 참조하면, 일 실시형태에 따른 디코딩 단계는, 전술한 오버샘플링 모듈에 의해 소수 범주의 데이터가 생성된 경우, 오버샘플링된 소수 범주의 데이터만을 복원하고, 원래의 학습 데이터(x)와 상기 복원된 오버샘플들(Reconstructed Oversamples)을 학습 데이터로 재구성하게 된다.

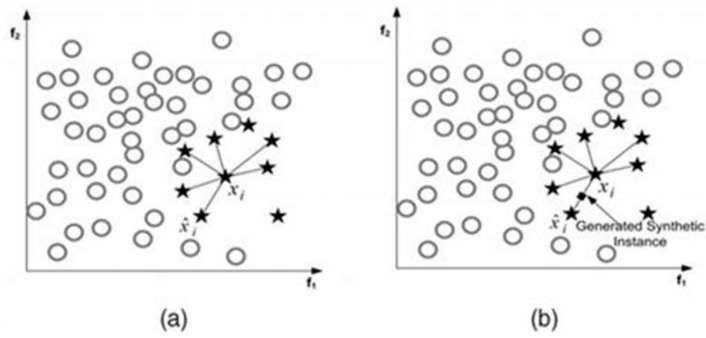
[0059] 도 10은 본 발명의 바람직한 실시예에 따른 오버샘플링 방법에 있어서, 디코딩 모듈의 일 실시 형태를 도시한 블록도이다. 도 10을 참조하면, 디코딩 모듈(120)의 다른 실시 형태는 비율을 맞춘 소수 범주의 데이터와 다수 범주의 데이터, 즉 학습 데이터(x)와 오버샘플링된 데이터를 모두 복원하여, 복원된 학습 데이터(Reconstructed x)와 복원된 오버샘플들(Reconstructed Oversamples)을 학습 데이터로 재구성하게 된다.

[0060] 전술한 과정들에 의해 재구성된 학습 데이터는 소수 범주 데이터와 다수 범주 데이터의 균형을 맞추게 되며, 기계 학습에 적용됨으로써 학습의 성능 및 정확도를 향상시킬 수 있게 된다.

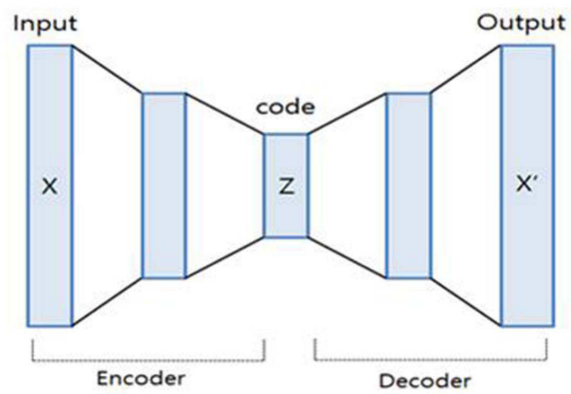
[0061] 이상에서 본 발명에 대하여 그 바람직한 실시예를 중심으로 설명하였으나, 이는 단지 예시일 뿐 본 발명을 한정하는 것이 아니며, 본 발명이 속하는 분야의 통상의 지식을 가진 자라면 본 발명의 본질적인 특성을 벗어나지 않는 범위에서 이상에 예시되지 않은 여러 가지의 변형과 응용이 가능함을 알 수 있을 것이다. 그리고, 이러한 변형과 응용에 관계된 차이점들은 첨부된 청구 범위에서 규정하는 본 발명의 범위에 포함되는 것으로 해석되어야 할 것이다.

도면

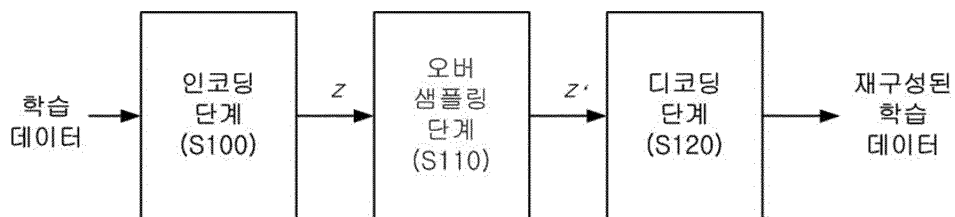
도면1



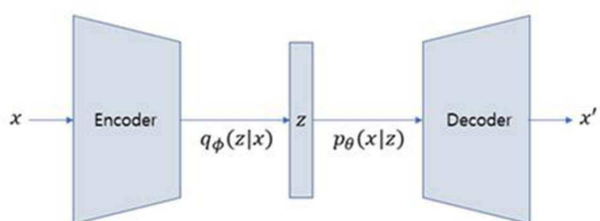
도면2



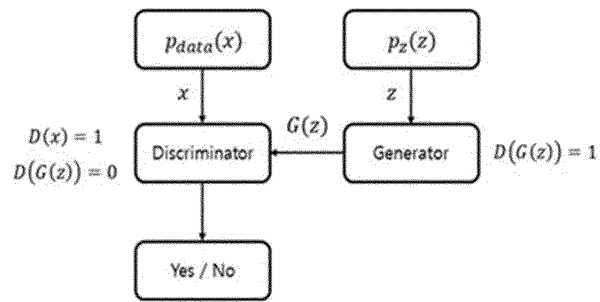
도면3



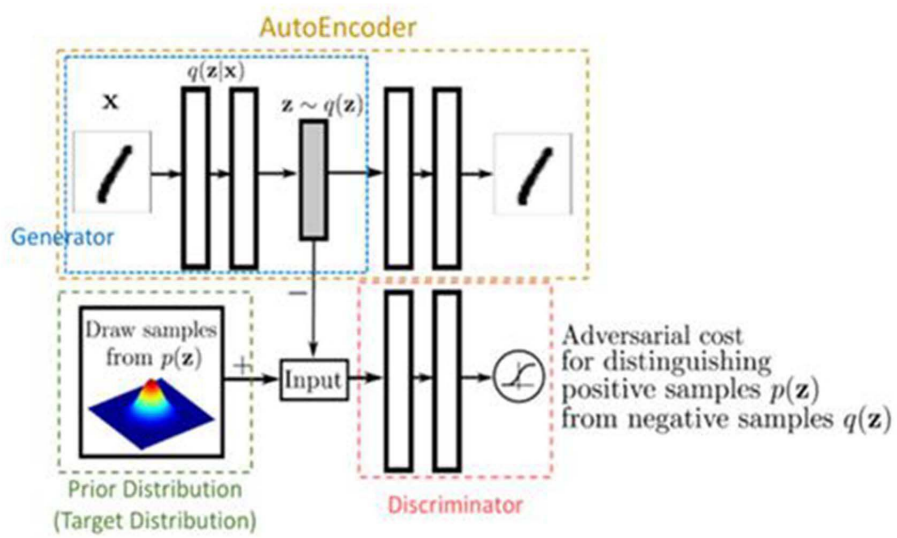
도면4



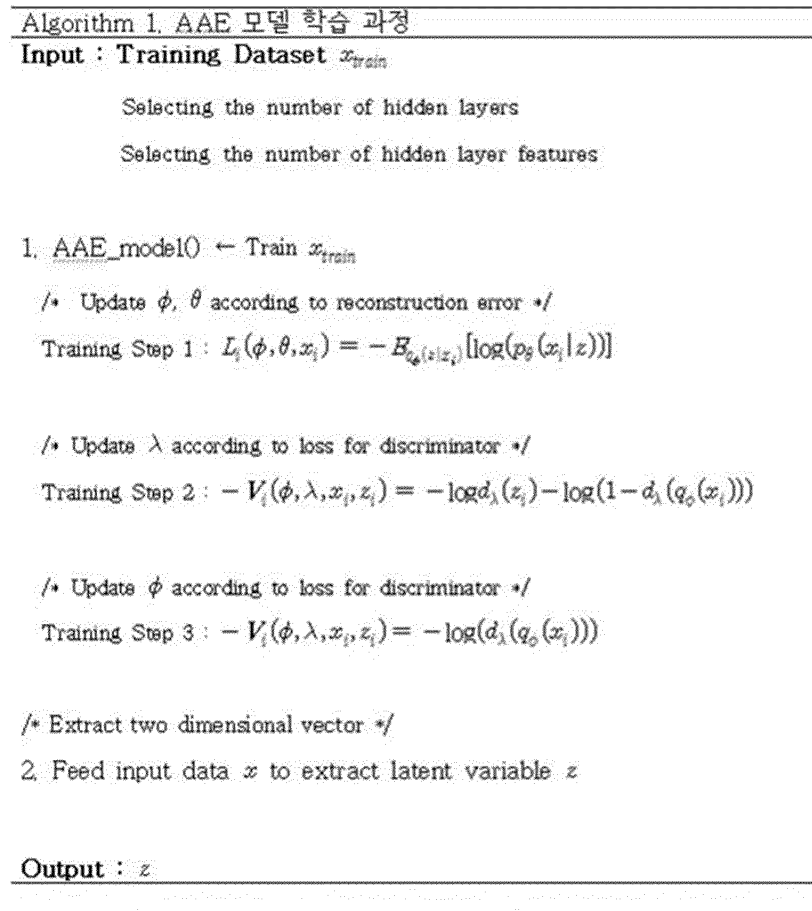
도면5



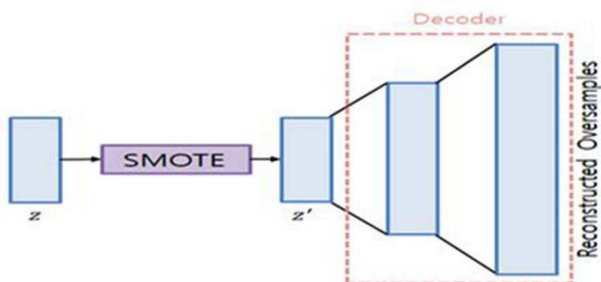
도면6



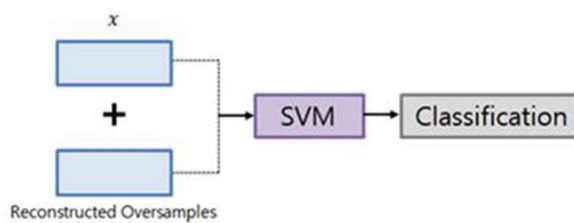
도면7



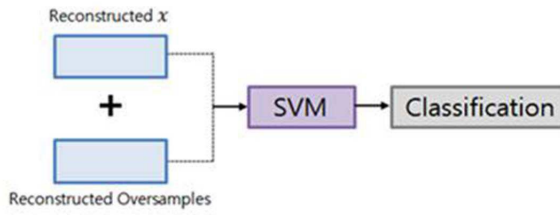
도면8



도면9



도면10



도면11

데이터 셋	전체 데이터의 개수	데이터 속성의 개수	다수 범주의 개수	소수 범주의 개수	불균형 비율
Breast	569	30	357	212	1.68
Ionosphere	351	34	225	126	1.79
Pima	768	8	500	268	1.87
Heart	303	13	164	139	1.18
Glass	214	9	138	76	1.82
Vehicle	846	18	629	217	2.9
Thyroid	215	5	180	35	5.14
Segment	2308	19	1979	329	6.02
Wine	178	13	71	59	1.20
			71	48	1.48

도면12

데이터 셋 모델	Breast	Ionosphere	Pima	Heart	Glass	Vehicle	Thyroid	Segment	Wine
Original	97.36%	87.04%	<u>77.09%</u>	80.29%	<u>64.44%</u>	76.25%	92.09%	99.61%	<u>97.75%</u>
Random	96.84%	87.04%	74.74%	79.28%	49.10%	75.17%	98.14%	99.65%	97.17%
SMOTE	96.84%	86.46%	74.53%	79.61%	49.10%	74.94%	98.14%	99.65%	96.63%
Borderline-SMOTE	95.43%	85.31%	73.69%	80.62%	50.96%	74.35%	97.21%	99.98%	97.19%
ADASYN	95.25%	85.60%	74.87%	79.17%	50.50%	74.26%	97.21%	99.85%	96.84%
AAE-SMOTE-Original	96.66%	<u>91.36%</u>	72.65%	<u>81.29%</u>	52.39%	72.69%	98.14%	<u>99.70%</u>	97.73%
AAE-SMOTE-Reproduce	<u>97.37%</u>	87.32%	74.87%	79.95%	49.57%	<u>79.20%</u>	<u>98.60%</u>	99.61%	97.56%

도면13

데이터 셋 모델	Breast	Ionosphere	Pima	Heart	Glass	Vehicle	Thyroid	Segment	Wine
Original	96.70%	84.56%	71.49%	79.81%	50.00%	54.29%	77.11%	98.85%	97.23%
Random	96.42%	84.56%	73.56%	78.88%	58.33%	75.46%	98.01%	99.14%	97.02%
SMOTE	96.59%	83.52%	74.58%	79.15%	58.06%	75.05%	98.82%	99.14%	96.67%
Borderline -SMOTE	95.79%	82.25%	74.39%	80.04%	60.26%	76.36%	98.27%	96.36%	97.45%
ADASYN	95.51%	82.60%	74.74%	79.66%	60.34%	76.76%	98.27%	96.36%	96.56%
AAE -SMOTE -Original	96.81%	84.10%	74.65%	79.35%	58.95%	75.07%	99.12%	99.18%	96.67%
AAE -SMOTE -Reproduce	96.25%	88.93%	71.70%	81.09%	61.87%	73.14%	96.37%	99.54%	97.94%

도면14

데이터 셋 모델	Breast	Ionosphere	Pima	Heart	Glass	Vehicle	Thyroid	Segment	Wine
Original	97.16%	85.76%	72.57%	79.69%	39.11%	50.78%	81.99%	99.16%	97.56%
Random	96.61%	85.33%	72.72%	78.72%	47.73%	71.25%	97.29%	99.27%	96.93%
SMOTE	96.59%	84.51%	73.64%	79.03%	47.85%	70.95%	97.35%	99.27%	96.93%
Borderline -SMOTE	95.16%	83.19%	72.39%	80.02%	49.62%	71.09%	95.68%	89.48%	97.02%
ADASYN	94.97%	83.44%	73.32%	79.13%	48.97%	71.14%	95.68%	89.48%	96.13%
AAE -SMOTE -Original	97.16%	85.35%	73.83%	79.29%	48.13%	73.43%	98.02%	99.11%	96.78%
AAE -SMOTE -Reproduce	96.44%	90.10%	70.57%	80.81%	51.57%	68.73%	96.90%	99.36%	97.12%