(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0088699
(43) 공개일자 2020년07월23일

(51) 국제특허분류(Int. Cl.)
*G06N 3/08* (2006.01) *G06F 16/35* (2019.01)
(52) CPC특허분류
*G06N 3/084* (2013.01)
*G06F 16/35* (2019.01)
(21) 출원번호 10-2019-0005334
(22) 출원일자 2019년01월15일
심사청구일자 2019년01월15일

(71) 출원인
**연세대학교 산학협력단**
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
**조성배**
서울특별시 강남구 선릉로76길 12, 101동 201호(대치동, 대치한신휴플러스)
**김진영**
경기도 하남시 미사강변대로 165, 110동 1303호(망월동, 미사강변 푸르지오)
*(뒷면에 계속)*
(74) 대리인
**민영준**

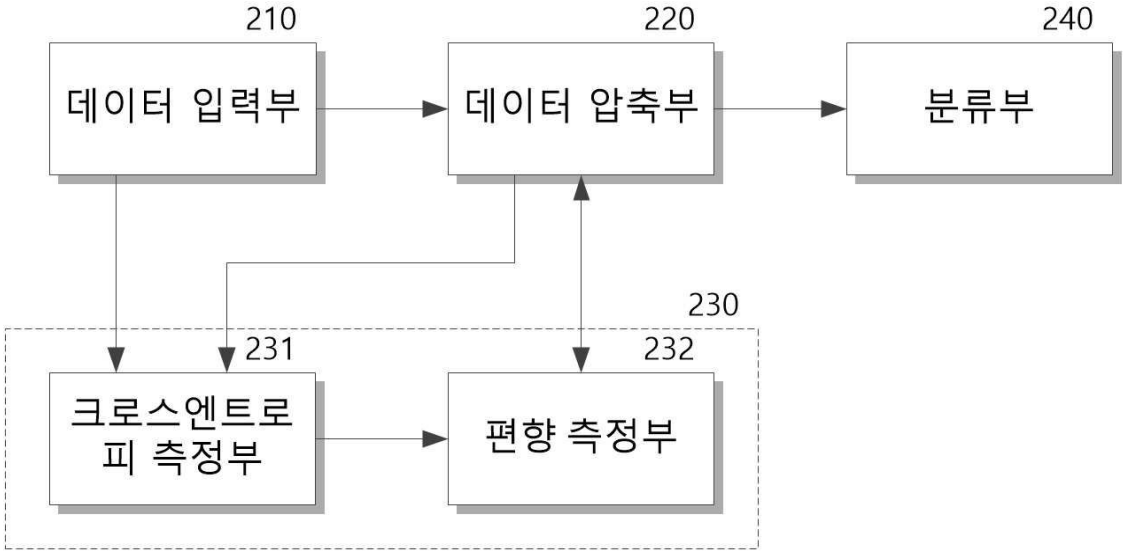
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **편향성이 감소된 분류장치 및 방법**

## (57) 요 약

본 발명은 미리 학습된 패턴 인식 방식에 따라 다수의 속성에 대한 속성값이 포함된 입력 데이터에서 특징값을 추출하여 기지정된 차원의 잠재 공간에 전사하는 데이터 압축부 및 데이터 압축부와 별도로 미리 학습된 패턴 인식 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류된 데이터를 복원하는 분류부를 포함하고, 데이터 압축부는 학습 시에 잠재 공간에 전사된 특징값에서 복원되는 데이터와 입력 데이터 사이의 오차인 전사 오차 및 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 통계적 특성을 분석하여 측정된 편향성 오차가 역전파되어 학습된 분류 장치 및 분류 방법을 제공할 수 있다.

*대 표 도* - 도5



(72) 발명자

**주세인**

서울특별시 강서구 마곡서로 175, 406동 201호(마곡동, 마곡엠밸리4단지)

**조유성**

서울특별시 마포구 독막로 288, 205호(대흥동, 대흥동세양아파트)

이 발명을 지원한 국가연구개발사업

과제고유번호 2016-0-00562

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터(NIPA산하)

연구사업명 정보통신방송연구개발사업

연구과제명 [이지바로][주관/한국과학기술원] 상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성 지능 연구개발 (3/5)

기 여 율 1/1

주관기관 한국과학기술원

연구기간 2018.07.01 ~ 2019.04.30

# 명 세 서

## 청구범위

### 청구항 1

미리 학습된 패턴 인식 방식에 따라 다수의 속성에 대한 속성값이 포함된 입력 데이터에서 특징값을 추출하여 기지정된 차원의 잠재 공간에 전사하는 데이터 압축부; 및

상기 데이터 압축부와 별도로 미리 학습된 패턴 인식 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류된 데이터를 복원하는 분류부; 를 포함하고,

상기 데이터 압축부는

학습 시에 상기 잠재 공간에 전사된 특징값에서 복원되는 데이터와 상기 입력 데이터 사이의 오차인 전사 오차 및 상기 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 통계적 특성을 분석하여 측정된 편향성 오차가 역전파되어 학습된 분류 장치.

### 청구항 2

제1 항에 있어서, 상기 분류 장치는

상기 데이터 압축부의 학습 시에 상기 전사 오차 및 상기 편향성 오차를 획득하기 위해 부가되는 편향 학습부; 를 더 포함하고,

상기 편향 학습부는

상기 입력 데이터(x)에서 잠재 공간에 전사된 특징값(f(x))으로부터 복원 데이터(g(f(x)))를 복원하여, 상기 입력 데이터(x)와 상기 복원 데이터(g(f(x))) 사이의 오차를 크로스엔트로피로 계산하여 상기 전사 오차를 획득하고,

잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 평균 및 분산을 기지정된 정규 확률 분포의 평균 및 분산과 비교하여 상기 편향성 오차를 측정하는 분류 장치.

### 청구항 3

제2 항에 있어서, 상기 편향 학습부는

수학식

$$\mathcal{L}_1 = crossentropy\left(x, g\left(f(x)\right)\right) + \mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$$

(여기서 crossentropy(x, g(f(x)))는 전사 오차를 계산하는 크로스엔트로피 함수를 나타내고, $\mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$ 은 각각 평균($\mu_{f(x)}$, 0)과 분산($\Sigma_{f(x)}$, I)을 갖는 2개의 정규 분포($\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)$, $\mathcal{N}(0, I)$) 형태로 잠재 공간에 전사된 특징값의 지정된 속성에 대한 잠재 변수와 정규 확률 분포를 갖는 기지정된 기준 잠재 변수 사이의 정보량 차이인 편향성 오차를 계산하는 KL 다이버전스(Kullback-Leibler divergence)를 나타낸다.)

에 따른 손실 함수($L_1$)의 계산 결과가 기지정된 제1 기준 손실값 이하가 되도록 상기 데이터 압축부를 학습시키는 분류 장치.

### 청구항 4

제2 항에 있어서, 상기 분류부는

상기 데이터 압축부에 의해 잠재 공간에 전사된 특징값(f(x))에 대해 분류 복원한 결과(h(f(x))와 기지정된 분

류값(y) 사이의 오차에 대한 손실 함수($L_2$)가 수학식

$$\mathcal{L}_2 = crossentropy\left(y, h(f(x))\right)$$

의 크로스엔트로피 함수로 계산되고, 손실 함수($L_2$)의 계산 결과가 기지정된 제2 기준 손실값 이하가 되도록 반복 학습된 분류 장치.

**청구항 5**

미리 학습된 패턴 인식 방식에 따라 다수의 속성에 대한 속성값이 포함된 입력 데이터에서 특징값을 추출하여 기지정된 차원의 잠재 공간에 전사하는 단계; 및

미리 학습된 패턴 인식 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류된 데이터를 복원하는 단계; 를 포함하고,

상기 전사하는 단계는 이전, 학습 단계에서 상기 잠재 공간에 전사된 특징값에서 복원되는 데이터와 상기 입력 데이터 사이의 오차인 전사 오차 및 상기 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 통계적 특성을 분석하여 측정된 편향성 오차가 역전파되어 미리 학습된 분류 방법.

**청구항 6**

제5 항에 있어서, 상기 학습 단계는

상기 입력 데이터(x)에서 잠재 공간에 전사된 특징값(f(x))으로부터 복원 데이터(g(f(x)))를 복원하여, 상기 입력 데이터(x)와 상기 복원 데이터(g(f(x))) 사이의 오차를 크로스엔트로피로 계산하여 상기 전사 오차를 획득하는 단계; 및

잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 평균 및 분산을 기지정된 정규 확률 분포의 평균 및 분산과 비교하여 상기 편향성 오차를 측정하는 단계; 를 포함하는 분류 방법.

**청구항 7**

제6 항에 있어서, 상기 학습 단계는

수학식

$$\mathcal{L}_1 = crossentropy\left(x, g(f(x))\right) + \mathcal{D}_{KL}[\mathcal{N}(\mu_{f(x)}, \Sigma_{f(x)})||\mathcal{N}(0, I)]$$

(여기서 crossentropy(x, g(f(x)))는 전사 오차를 계산하는 크로스엔트로피 함수를 나타내고, $\mathcal{D}_{KL}[\mathcal{N}(\mu_{f(x)}, \Sigma_{f(x)})||\mathcal{N}(0, I)]$ 은 각각 평균($\mu_{f(x)}$, 0)와 분산($\Sigma_{f(x)}$, I)를 갖는 2개의 정규 분포($\mathcal{N}(\mu_{f(x)}, \Sigma_{f(x)})$, $\mathcal{N}(0, I)$) 형태로 잠재 공간에 전사된 특징값의 지정된 속성에 대한 잠재 변수와 정규 확률 분포를 갖는 기지정된 기준 잠재 변수 사이의 정보량 차이인 편향성 오차를 계산하는 KL 다이버전스(Kullback-Leibler divergence)를 나타낸다.)

에 따른 손실 함수($L_1$)의 계산 결과가 기지정된 제1 기준 손실값 이하가 되도록 학습하는 분류 방법.

**청구항 8**

제7 항에 있어서, 상기 학습 단계는

잠재 공간에 전사된 특징값(f(x))에 대해 분류 복원한 결과(h(f(x))와 기지정된 분류값(y) 사이의 오차에 대한 손실 함수($L_2$)가 수학식

$$\mathcal{L}_2 = crossentropy\left(y, h(f(x))\right)$$

의 크로스엔트로피 함수로 계산되고, 손실 함수($L_2$)의 계산 결과가 기지정된 제2 기준 손실값 이하가 되도록 학

습되는 단계; 를 더 포함하는 분류 방법.

## 발명의 설명

### 기 술 분 야

[0001] 본 발명은 분류장치 및 방법에 관한 것으로, 편향성이 감소된 분류장치 및 방법에 관한 것이다.

### 배 경 기 술

[0002] 딥러닝(Deep learning) 기반 분류장치는 인공신경망(Artificial Neural Network)을 사용하여 입력된 데이터가 가진 속성을 판단 및 출력하는 장치를 뜻하며, 광범위한 분야에서 범용적으로 응용되고 있다.

[0003] 딥러닝 기반 분류장치는 다양한 방면에서 연구되고 있으나, 주로 분류의 정확도를 높이는 데에 편중되어 있어, 분류 과정을 규명하고 제어하는 연구가 미흡한 실정이다.

[0004] 이로 인해 딥러닝 기반 분류장치는 인공신경망의 구성과 학습 방법에 따라 높은 분류 정확도를 나타낼 수 있는 반면, 분류 과정에서 입력 데이터에서 사용자가 의도하지 않은 특정 정보에 과도하게 의존하는 편향성(bias)이 발생할 가능성이 있다.

[0005] 분류장치의 편향성 문제는 사용자의 요구에 적합한 분류를 수행하지 못하는 실용적인 문제를 발생시킬 수 있을 뿐만 아니라, 사회적, 경제적으로 민감한 속성을 판단할 때 인종, 성별 정보 등에 의존하여 윤리적인 문제로 이어질 수 있다는 문제가 있다.

[0006] 만일 딥러닝 기반 분류장치가 직원 채용 추천, 채무 이행 예측 또는 범죄 재발 예측과 같은 사회적, 경제적으로 민감한 속성을 판별하는 경우에, 기존에 주어진 학습 데이터를 기반으로 학습된 분류장치는 입력 데이터 중 가능한 공정성을 유지해야 하는 연령, 인종, 성별, 수입 등과 같이 윤리적으로 민감한 개인의 신상정보에 편향되어 판단하게 될 수 있다.

[0007] 일예로 딥러닝 기반 분류장치가 직원 채용 추천에 이용되어 입력 데이터를 추천 또는 비추천으로 분류하는 경우에, 추천으로 분류된 입력 데이터 중 남성에 대한 입력 데이터가 90% 이상이 되도록 과도하게 편향되어 분류할 수도 있다.

[0008] 이는 분류 장치의 분류 결과를 기반으로 업무를 추진하고자 하는 사용자에게 의도하지 않은 인종 차별 및 성차별과 같은 사회적, 윤리적 이슈를 발생시킬 수 있다는 문제가 있다.

### 선행기술문헌

#### 특허문헌

[0009] (특허문헌 0001) 한국 등록 특허 제10-1855168호 (2018.04.30 등록)

## 발명의 내용

### 해결하려는 과제

[0010] 본 발명의 목적은 분류 과정에서 입력 데이터의 특정 속성에 관한 편향성을 저감하여 분류를 수행할 수 있는 분류 장치 및 방법을 제공하는데 있다.

[0011] 본 발명의 다른 목적은 인종, 성별 정보와 같이 윤리적으로 민감한 속성에 대한 편향성을 저감할 수 있는 분류 장치 및 방법을 제공하는데 있다.

### 과제의 해결 수단

[0012] 상기 목적을 달성하기 위한 본 발명의 일 실시예에 따른 분류장치는 미리 학습된 패턴 인식 방식에 따라 다수의 속성에 대한 속성값이 포함된 입력 데이터에서 특징값을 추출하여 기지정된 차원의 잠재 공간에 전사하는 데이터 압축부; 및 상기 데이터 압축부와 별도로 미리 학습된 패턴 인식 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류된 데이터를 복원하는 분류부; 를 포함하고, 상기 데이터 압축부는 학습

시에 상기 잠재 공간에 전사된 특징값에서 복원되는 데이터와 상기 입력 데이터 사이의 오차인 전사 오차 및 상기 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 통계적 특성을 분석하여 측정된 편향성 오차가 역전파되어 학습된다.

[0013] 상기 분류 장치는 상기 데이터 압축부의 학습 시에 상기 전사 오차 및 상기 편향성 오차를 획득하기 위해 부가되는 편향 학습부; 를 더 포함하고, 상기 편향 학습부는 상기 입력 데이터(x)에서 잠재 공간에 전사된 특징값(f(x))으로부터 복원 데이터(g(f(x)))를 복원하여, 상기 입력 데이터(x)와 상기 복원 데이터(g(f(x))) 사이의 오차를 크로스엔트로피로 계산하여 상기 전사 오차를 획득하고, 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 평균 및 분산을 기지정된 정규 확률 분포의 평균 및 분산과 비교하여 상기 편향성 오차를 측정할 수 있다.

[0014] 상기 편향 학습부는 수학식

[0015] $$\mathcal{L}_1 = crossentropy\left(x, g\left(f(x)\right)\right) + \mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$$

[0016] (여기서 crossentropy(x, g(f(x)))는 전사 오차를 계산하는 크로스엔트로피 함수를 나타내고, $\mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$ 은 각각 평균($\mu_{f(x)}$, 0)과 분산($\Sigma_{f(x)}$, I)을 갖는 2개의 정규 분포($\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)$, $\mathcal{N}(0, I)$) 형태로 잠재 공간에 전사된 특징값의 지정된 속성에 대한 잠재 변수와 정규 확률 분포를 갖는 기지정된 기준 잠재 변수 사이의 정보량 차이인 편향성 오차를 계산하는 KL 다이버전스(Kullback-Leibler divergence)를 나타낸다.) 에 따른 손실 함수($L_1$)의 계산 결과가 기지정된 제1 기준 손실값 이하가 되도록 상기 데이터 압축부를 학습시킬 수 있다.

[0017] 상기 분류부는 상기 데이터 압축부에 의해 잠재 공간에 전사된 특징값(f(x))에 대해 분류 복원한 결과(h(f(x))와 기지정된 분류값(y) 사이의 오차에 대한 손실 함수($L_2$)가 수학식

[0018] $$\mathcal{L}_2 = crossentropy\left(y, h\left(f(x)\right)\right)$$

[0019] 의 크로스엔트로피 함수로 계산되고, 손실 함수($L_2$)의 계산 결과가 기지정된 제2 기준 손실값 이하가 되도록 반복 학습될 수 있다.

[0020] 상기 목적을 달성하기 위한 본 발명의 다른 실시예에 따른 분류방법은 미리 학습된 패턴 인식 방식에 따라 다수의 속성에 대한 속성값이 포함된 입력 데이터에서 특징값을 추출하여 기지정된 차원의 잠재 공간에 전사하는 단계; 및 미리 학습된 패턴 인식 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류된 데이터를 복원하는 단계; 를 포함하고, 상기 전사하는 단계는 이전, 학습 단계에서 상기 잠재 공간에 전사된 특징값에서 복원되는 데이터와 상기 입력 데이터 사이의 오차인 전사 오차 및 상기 잠재 공간에 전사된 특징값에서 기지정된 속성에 대응하는 잠재 변수의 통계적 특성을 분석하여 측정된 편향성 오차가 역전파되어 미리 학습된다.

## *발명의 효과*

[0021] 따라서, 본 발명의 실시예에 따른 분류장치 및 방법은 잠재 공간에 전사된 압축 데이터의 지정된 속성에 관한 편향을 측정하고, 편향이 저감되도록 조절하여 분류를 수행함으로써, 분류 장치의 편향 문제를 최소화할 수 있다.

## *도면의 간단한 설명*

[0022] 도1 은 본 발명의 일 실시예에 따른 분류 장치의 개략적 구조를 나타낸다.

도2 는 본 실시예의 분류 장치가 편향성을 저감하는 개념을 나타낸다.

도3 은 도1 의 분류 장치를 학습시키는 개념을 나타낸다.

도4 는 본 발명의 일 실시예에 따른 분류 방법을 나타낸다.

도5 는 본 발명의 다른 실시예에 따른 분류 장치의 개략적 구조를 나타낸다.

도6 은 본 발명의 다른 실시예에 따른 분류 방법을 나타낸다.

## *발명을 실시하기 위한 구체적인 내용*

[0023] 본 발명과 본 발명의 동작상의 이점 및 본 발명의 실시에 의하여 달성되는 목적을 충분히 이해하기 위해서는 본 발명의 바람직한 실시예를 예시하는 첨부 도면 및 첨부 도면에 기재된 내용을 참조하여야만 한다.

[0024] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시예를 설명함으로써, 본 발명을 상세히 설명한다. 그러나, 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며, 설명하는 실시예에 한정되는 것이 아니다. 그리고, 본 발명을 명확하게 설명하기 위하여 설명과 관계없는 부분은 생략되며, 도면의 동일한 참조부호는 동일한 부재임을 나타낸다.

[0025] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라, 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈", "블록" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0026] 도1 은 본 발명의 일 실시예에 따른 분류 장치의 개략적 구조를 나타내고, 도2 는 본 실시예의 분류 장치가 편향성을 저감하는 개념을 나타낸다.

[0027] 도1 을 참조하면, 본 실시예에 따른 분류 장치는 데이터 입력부(110), 데이터 압축부(120), 편향 조절부(130) 및 분류부(140)를 포함한다.

[0028] 데이터 입력부(110)는 분류가 수행되어야 하는 다수의 입력 데이터를 획득한다. 여기서 다수의 입력 데이터 각각은 다수의 속성에 따른 속성값을 포함한다. 일예로 분류 장치가 소득 수준에 대한 분류 장치인 경우, 각 입력 데이터에는 다수의 속성으로 나이, 학력, 성별, 인종, 결혼 유무 등이 지정될 수 있으며, 각 속성에 따른 속성값이 포함될 수 있다.

[0029] 그리고 본 실시예에서는 입력 데이터에 포함된 다수의 속성 중 편향성을 갖지 않도록 보호된 속성으로 성별이 지정된 것으로 가정하여 설명한다. 그러나 다른 속성이 보호된 속성으로 지정될 수도 있으며, 사용자에 의해 미리 지정될 수 있다.

[0030] 데이터 압축부(120)는 미리 지정된 패턴 인식 방식으로 학습되어 데이터 입력부(110)에서 획득된 다수의 입력 데이터의 특징값을 잠재 공간(latent space)에 전사한다. 여기서 데이터 압축부(120)는 일예로 VAE(Variational Autoecoder)의 인코더로 구현될 수 있으며, 속성의 수에 따라 다차원으로 표현될 수 있는 입력 데이터에서 속성의 수보다 적은 기지정된 저차원의 잠재 공간에 표현할 수 있도록 특징값으로 추출하여 재구축(Reconstruction)한다.

[0031] 이때, 데이터 압축부(120)는 보호된 속성에 무관하게 특징값을 전사하며, VAE로 구현된 데이터 압축부(120)는 특징값이 잠재 공간 상에서 지정된 확률 분포를 갖는 잠재 변수(latent variable)로 표현되도록 전사할 수 있다.

[0032] 편향 조절부(130)는 데이터 압축부(120)에 의해 잠재 공간에 전사된 특징값에서 보호된 속성에 대응하는 잠재 변수(z)의 편향성을 분석하고, 분석된 편향성이 제거된 보정 데이터를 생성하고, 보정 데이터를 다시 잠재 공간에 전사한다.

[0033] 편향 조절부(130)는 편향 측정부(131), 데이터 재구성부(132) 및 데이터 재압축부(133)를 포함할 수 있다. 우선 편향 측정부(131)는 잠재 공간에 전사된 특징값에서 보호된 속성에 대한 잠재 변수(z)의 통계적 특성을 추출하고, 추출된 통계적 특성에 따라 잠재 변수(z)에 대한 편향성을 제거하기 위한 보상값을 획득한다. 잠재 변수(z)의 대한 편향성은 잠재 변수(z)의 평균값과 기지정된 기준 평균값(여기서는 일예로 0) 사이의 오차를 의미한다. 따라서 편향성을 제거하기 위한 보상값은 잠재 변수(z)의 평균값이 기준 평균값으로 보정되도록 오차를 제거하기 위한 값으로 획득될 수 있다.

[0034] 일예로, 보호된 속성이 상기한 바와 같이 성별일 때, 본 실시예의 분류 장치는 도2 에 도시된 바와 같이, 여성 및 남성 각각 대해 잠재 공간에 전사된 잠재 변수(z)의 평균값이 0이 되도록 보상값을 획득한다.

[0035] 이때 편향 측정부(131)는 단순히 편향성에 대한 오차뿐만 아니라 잠재 변수(z)의 분산 오차 또한 함께 획득하여 보상값을 획득할 수 있다. 일예로 편향 측정부(131)는 잠재 변수(z)의 분산과 정규 분포의 분산 사이의 오차를

보상하기 위한 보상값을 함께 획득할 수 있다.

[0036] 데이터 재구성부(132)는 편향 측정부(131)에서 획득한 보상값 따라 잠재 공간에 전사된 특징값에서 편향성이 제거된 보정 데이터를 획득한다. 데이터 재구성부(132)는 미리 학습된 패턴 인식 방식에 따라 학습되어 잠재 공간에 전사된 특징값을 다시 입력 데이터와 동일한 속성을 갖는 보정 데이터로 획득하며, 이때, 편향 측정부(131)에서 획득한 보상값이 반영된 데이터를 획득한다.

[0037] 데이터 압축부(120)가 VAE의 인코더로 구현된 경우, 데이터 재구성부(132)는 VAE의 디코더로 구현될 수 있다.

[0038] 데이터 재압축부(133)는 데이터 압축부(120)를 모의한 구성으로, 데이터 압축부(120)와 동일하게 학습된다. 이에 데이터 재압축부(133)는 데이터 압축부(120)와 유사하게 데이터 재구성부(132)에서 획득한 보정 데이터의 특징값을 잠재 공간에 전사한다. 이때 데이터 재구성부(132)에서 획득한 보정 데이터는 상기한 바와 같이 편향 측정부(131)에서 획득된 편향에 의한 보상값이 반영되어 편향성이 저감된 데이터이다. 따라서 데이터 재압축부(133)에서 잠재 공간에 전사한 특징값은 편향성이 저감된 상태이다.

[0039] 즉 데이터 재압축부(133)는 데이터 압축부(120)와 유사하게 특징값을 잠재 공간에 전사하기 위한 구성이지만, 데이터 압축부(120)와 달리 보호된 속성에 대한 편향성이 최소화 되도록 조절된 데이터를 잠재 공간에 전사한다.

[0040] 분류부(140)는 데이터 재압축부(133)에 의해 잠재 공간에 전사된 특징값을 미리 학습된 패턴 인식 방식에 따라 분류한다. 분류부(140)는 데이터 재구성부(132)와 유사하게 특징값으로부터 다수의 속성을 갖는 데이터를 획득하되, 데이터를 분류하여 획득한다.

[0041] 상기한 직원 채용 추천 예에서 분류부(140)는 특징값에서 입력 데이터에 포함된 나이, 학력, 성별, 인종, 결혼 유무 등의 속성값을 복원하되, 특징값에 따라 추천 및 비추천으로 분류하여 획득할 수 있다.

[0042] 그리고 편향 조절부(130)는 분류된 추천 및 비추천으로 분류된 데이터에서 보호된 속성이 균등하도록 조절하는 기능을 수행한다. 즉 추천으로 분류된 데이터에서 보호된 속성인 남성 데이터의 수와 여성 데이터의 수 및 분포가 균등하도록 조절하는 기능을 수행한다.

[0043] 결과적으로 도1 에 도시된 분류 장치는 편향 조절부(130)가 편향 측정부(131), 데이터 재구성부(132) 및 데이터 재압축부(133)를 포함하여, 잠재 공간에 전사된 특징값에서 보호된 속성의 편향성을 측정하고, 측정된 편향성을 보상한 특징값에 대응하는 보정데이터를 획득하여 다시 잠재 공간에 전사하도록 함으로써, 편향성을 저감시킬 수 있다.

[0044] 그러나 도1 의 분류 장치가 요구되는 성능을 나타내기 위해서는 데이터 압축부(120)와 데이터 재압축부(133), 데이터 재구성부(132) 및 분류부(140)가 미리 학습되어야 한다. 여기서 데이터 재압축부(133)는 데이터 압축부(120)와 동일하게 학습되어야 하므로, 학습된 데이터 압축부(120)가 동일하게 이용될 수 있다.

[0045] 도3 은 도1 의 분류 장치를 학습시키는 개념을 나타낸다.

[0046] 도3 에서 (a)는 VAE 또는 오토인코더(Autoencoder)의 일반적 구성으로, VAE 및 오토인코더는 입력값(x)의 특징값(f(x))을 추출하여 잠재 공간에 전사하는 인코더(Encoder)와 잠재 공간에 전사된 특징값(f(x))으로부터 입력값(x)을 복원(g(f(x)))하기 위한 디코더(Decoder)를 포함할 수 있다.

[0047] 인코더(Encoder)는 도1 의 데이터 압축부(120) 및 데이터 재압축부(133)에 대응하는 구성이며, 디코더(Decoder)는 도1 의 데이터 재구성부(132)에 대응하는 구성으로 볼 수 있다.

[0048] VAE 및 오토인코더가 우수한 성능을 나타내기 위해서는 인코더(Encoder)의 입력값(x)과 디코더(Decoder)의 출력값(g(f(x)))의 오차, 즉 전사 오차의 수준을 손실 함수(loss function)($L_1$)가 최소가 되도록 수학식 1에 따라 비지도 방식으로 미리 학습될 수 있다.

*수학식 1*

[0049] $$\mathcal{L}_1 = crossentropy\left(x, g\left(f(x)\right)\right)$$

[0050] 여기서 crossentropy(x, g(f(x)))는 VAE의 인코더와 디코더에 의한 재구축 오차를 계산하는 크로스엔트로피 함

수를 나타낸다.

[0051] 한편, (b)는 분류부(140)를 학습시키기 위한 구성을 나타낸다.

[0052] (b)에서 인코더(Encoder)는 (a)와 동일한 인코더로서 데이터 재압축부(133)이며, 미리 학습된 상태이다. 그리고 분류기(classifier)는 분류부(140)에 대응하는 구성으로 잠재 공간에 전사된 잠재 변수(z)의 특징값에 따라 0또는 1의 이진 분류값(y)으로 분류되어 복원된 데이터를 출력하는 것으로 가정한다. 상기한 직원 채용 추천의 예에서 1이 추천이고, 0이 비추천일 수 있다.

[0053] 인코더(Encoder)가 미리 학습된 상태이므로, 분류부(140)는 수학식 2와 같이 구성된 손실 함수($L_2$)가 최소가 되도록 학습될 수 있다.

*수학식 2*

[0054] $$\mathcal{L}_2 = crossentropy\left(y, h(f(x))\right)$$

[0055] 이때, 분류부(140)는 학습 데이터를 이용하여 지도 학습 방식으로 학습될 수 있다.

[0056] 도4 는 본 발명의 일 실시예에 따른 분류 방법을 나타낸다.

[0057] 도1 을 참조하여 도4 의 분류 방법을 설명하면, 우선 분류를 수행할 입력 데이터를 획득한다(S11). 이때, 입력 데이터의 다수의 속성 중 보호될 속성이 함께 지정될 수 있다. 그리고 입력 데이터의 다수의 속성값에 따른 특징값을 추출하여 잠재 공간에 기지정된 확률 분포를 갖는 잠재 변수 형태로 전사한다(S12).

[0058] 입력 데이터에 대한 특징값이 잠재 공간에 전사되면, 잠재 공간에 전사된 특징값에서 보호된 속성에 대응하는 잠재 변수(z)의 통계적 특성을 추출하고, 추출된 통계적 특성에 따라 보호된 속성에 대한 편향성을 측정한다(S13). 그리고 잠재 공간에 전사된 특징값을 다시 입력 데이터에 대응하는 속성을 갖는 데이터로 획득하며, 이때, 측정된 편향성을 보상하기 위한 오차가 반영된 보정 데이터를 획득한다(S14).

[0059] 보정 데이터가 획득되면, 획득된 보정 데이터에서 특징값을 추출하여 다시 잠재 공간에 전사한다(S15). 즉 편향성이 보상된 보정 데이터에서 특징값을 추출하여 잠재 공간에 전사한다. 이후, 잠재 공간에 전사된 특징값을 분류하여 분류된 데이터를 획득한다(S16).

[0060] 도5 는 본 발명의 다른 실시예에 따른 분류 장치의 개략적 구조를 나타낸다.

[0061] 도1 의 분류 장치를 다시 참조하면, 도1 의 분류 장치에서는 편향 조절부(130)의 편향 측정부(131)가 잠재 공간에 전사된 특징값에서 보호된 속성에 대응하는 잠재 변수(z)의 통계적 특성을 추출하고, 추출된 통계적 특성에 따라 보호된 속성에 대한 편향성을 통계적 특성에 따라 측정하며, 데이터 재구성부(132)가 측정된 편향성에 따라 잠재 공간에 전사된 특징값으로부터 보정 데이터를 획득한 후, 데이터 재압축부(133) 다시 잠재 공간에 보상 데이터에 대한 특징값을 전사한다.

[0062] 그러나 데이터 재압축부(133)는 상기한 바와 같이 데이터 압축부(120)를 모의한 구성으로 데이터 압축부(120)와 동일하다. 그리고 데이터 압축부(120)는 인공 신경망으로 구현되어 미리 학습된 패턴 인식 방식에 따라 입력 데이터로부터 특징값을 추출하여 잠재 공간에 전사한다. 따라서 데이터 압축부(120)가 입력 데이터로부터 곧바로 편향성이 저감된 특징값을 전사하도록 미리 학습될 수 있다면, 편향 조절부(130)를 제거할 수 있다.

[0063] 이에 도5 의 분류 장치는 데이터 입력부(210), 데이터 압축부(220) 및 분류부(240)를 포함한다.

[0064] 데이터 입력부(210)는 분류가 수행되어야 하는 다수의 입력 데이터를 획득한다. 그리고 데이터 압축부(220)는 미리 학습된 패턴 인식 방식에 따라 데이터 입력부(210)에서 획득된 다수의 입력 데이터의 특징값을 잠재 공간에 전사한다. 여기서 데이터 압축부(220)는 기지정된 보호되어야 하는 속성값의 편향성이 저감되도록 특징값을 잠재 공간에 전사할 수 있도록 미리 학습된다.

[0065] 그리고 분류부(240)는 잠재 공간에 전사된 특징값을 미리 학습된 패턴 인식 방식에 따라 분류하고, 분류된 특징값으로부터 다수의 속성을 갖는 데이터를 복원한다.

[0066] 다만 도5 의 분류 장치는 상기한 바와 같이, 데이터 압축부(220)가 입력 데이터로부터 보호된 속성값의 편향성

을 저감하여 특징값을 잠재 공간에 전사할 수 있도록 미리 학습되어야 한다. 이에 데이터 압축부(220)를 학습시키기 위한 편향 학습부(230)가 학습 시에 부가되어 이용될 수 있다.

[0067] 편향 학습부(230)는 데이터 압축부(220)가 입력 데이터의 특징값을 잠재 공간에 복원 가능하도록 전사할 수 있도록 학습 시킴과 동시에, 잠재 공간에 전사되는 특징값의 보호된 속성에 대한 편향성이 저감된 형태로 전사되도록 학습시키기 위한 구성이다.

[0068] 편향 학습부(230)는 크로스엔트로피 측정부(231) 및 편향 측정부(232)를 포함할 수 있다.

[0069] 크로스엔트로피 측정부(231)는 도2 의 (a)에 도시된 디코더(Decoder)로 구현될 수 있다. 크로스엔트로피 측정부(231)는 데이터 압축부(220)에 의해 입력 데이터(x)가 잠재 공간에 전사된 특징값(f(x))을 데이터(g(f(x))로 복원하고, 입력 데이터(x)와 복원된 데이터(g(f(x)) 사이의 오차를 전사 오차로서 계산한다.

[0070] 한편 편향 측정부(232)는 데이터 압축부(220)에 의해 잠재 공간에 전사된 특징값에서 보호된 속성에 대한 통계적 특성을 추출하고, 추출된 통계적 특성에 따라 보호된 속성에 대한 편향성을 나타내는 편향성 오차를 획득한다. 그리고 획득된 편향성 오차를 크로스엔트로피 측정부(231)에서 측정된 전사 오차와 함께 데이터 압축부(220)로 역전파한다.

[0071] 여기서 편향 측정부(232)는 도1 의 편향 측정부(131)와 유사하게 잠재 변수(z)의 평균값과 기준 평균값 사이의 오차를 측정하여 편향성 오차를 획득할 수 있다. 또한 편향 학습부(230)는 잠재 공간에 전사된 잠재 변수(z)의 분산과 정규 분포의 분산 사이의 분산 오차를 함께 측정하여 편향성 오차에 포함할 수 있다. 이는 데이터 압축부(220)가 입력 데이터로부터 특징값을 추출하여 잠재 공간에 전사할 때, 보호된 속성에 대해 가능한 균등한 분포로 특징값을 전사할 수 있도록 하기 위함이다.

[0072] 일예로 편향성 측정부(232)는 잠재 변수(z)를 기지정된 정규 확률 분포와 비교하여 편향성 오차를 획득할 수 있다.

[0073] 그리고 편향 학습부(230)는 획득된 전사 오차와 편향성 오차를 데이터 압축부(220)로 역전파으로써, 데이터 압축부(220)가 입력 데이터로부터 특징값을 추출하여 잠재 공간에 전사할 때, 보호된 속성의 잠재 변수(z)의 평균값과 분산이 기지정된 기준 평균값 및 분산을 갖도록 학습시킬 수 있다.

[0074] 데이터 압축부(220)는 편향 학습부(230)에 의해 수학식 3에 따른 손실 함수($L_1$)가 기지정된 제1 기준 손실값 이하가 될 때까지 반복 학습 될 수 있다.

*수학식 3*

[0075] $$\mathcal{L}_1 = crossentropy\left(x, g\left(f(x)\right)\right) + \mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$$

[0076] 여기서 crossentropy(x, g(f(x)))는 전사 오차를 계산하는 크로스엔트로피 함수를 나타내고, $\mathcal{D}_{KL}[\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)||\mathcal{N}(0, I)]$ 은 각각 평균($\mu_{f(x)}$, 0)과 분산($\Sigma_{f(x)}$, I)을 갖는 2개의 정규 분포($\mathcal{N}\left(\mu_{f(x)}, \Sigma_{f(x)}\right)$, $\mathcal{N}(0, I)$) 형태로 잠재 공간에 전사된 특징값의 지정된 속성에 대한 잠재 변수와 정규 확률 분포를 갖는 기지정된 기준 잠재 변수 사이의 정보량 차이를 계산하는 KL 다이버전스(Kullback-Leibler divergence)를 나타낸다.

[0077] 편향 학습부(230)는 데이터 압축부(200)의 학습이 완료되면, 즉 분류 장치가 실제 이용될 때는 제거되어 이용될 수 있다.

[0078] 도6 은 본 발명의 다른 실시예에 따른 분류 방법을 나타낸다.

[0079] 도5 의 분류 장치를 이용한 분류 방법을 설명하면, 우선 분류 장치가 이용되기 위해서는 데이터 압축부(220)와 분류부(240)가 학습되어야 한다.

[0080] 이에 도5 의 분류 방법에서는 우선 데이터 압축부(220)를 학습시키는 단계(S20)와 분류부(240)를 학습시키는 단계(S30) 및 데이터 분류 단계(S40)를 포함한다.

[0081] 데이터 압축부(220)를 학습시키는 단계에서는 우선 데이터 압축부(220)는 학습을 위한 데이터를 인가받아 학습

특징값을 잠재 공간에 전사한다(S21). 이때 전사된 학습 특징값은 데이터 압축부(220)가 학습되지 않은 상태이므로, 전사된 특징값에 대한 데이터 복원 시에 학습 데이터와 복원된 데이터 사이의 전사 오차가 크게 발생될 수 있다. 또한 보호된 속성에 대한 고려가 되지 않은 상태로 잠재 공간에 전사되므로 보호된 속성에 대한 편향성이 존재하는 특징값일 수 있다.

[0082] 이에 편향 학습부(230)는 우선 학습 특징값을 복원하여 복원된 데이터와 학습 데이터 사이의 전사 오차를 크로스엔트로피를 이용하여 계산한다(S22).

[0083] 그리고 전사된 학습 특징값에서 보호된 속성에 대한 잠재 변수(z)의 통계적 특성을 추출하고, 추출된 통계적 특성에 따라 잠재 변수(z)의 편향성을 측정하여 편향성 오차를 측정한다(S23). 이때 편향 학습부(230)는 잠재 공간에 전사된 잠재 변수(z)의 평균과 분산을 기지정된 평균 및 분산을 갖는 정규 확률 분포와 비교 측정하여 편향성 오차를 측정할 수 있다.

[0084] 편향 학습부(230)는 획득된 크로스엔트로피와 편향성 오차를 이용하여 수학식 3의 손실 함수($L_1$)를 계산하고, 계산된 손실 함수($L_1$)의 값을 함수 데이터 압축부(220)로 역전파하여 데이터 압축부(220)가 학습되도록 한다(S24).

[0085] 상기한 데이터 압축부(220)의 학습은 데이터 압축부(220)가 요구되는 성능을 나타낼 수 있을 때까지 반복 수행될 수 있다. 즉 데이터 압축부(220)가 입력 데이터로부터 제1 기준 손실값 이하가 되도록 특징값을 잠재 공간에 전사할 수 있도록 학습된다.

[0086] 그리고 데이터 압축부(220)의 학습이 완료되면, 분류부(240)가 학습된다(S30). 분류부(240)는 도2 의 (b)와 같이 학습된 데이터 압축부(220)를 인코더(Encoder)로 이용하여 수학식 2와 같이 크로스엔트로피를 이용한 손실 함수($L_2$)가 기지정된 제2 기준 손실값 이하가 되도록 학습될 수 있다.

[0087] 데이터 압축부(220)와 분류부(240)가 학습되면, 분류 장치는 분류되어야 할 입력 데이터를 인가받아 분류를 수행한다(S40). 우선 데이터 입력부(210)는 분류를 수행해야 할 입력 데이터를 획득한다(S41). 그리고 1차 학습 및 편향성 저감 학습된 데이터 압축부(220)는 입력 데이터로부터 특징값을 추출하여 보호된 속성에 대한 잠재 변수(z)의 편향성이 저감되도록 잠재 공간에 전사한다(S42).

[0088] 이에 분류부(240)는 이미 학습된 방식에 따라 잠재 공간에 전사된 특징값을 분류하고, 분류된 특징값으로부터 분류 데이터를 획득한다(S43).

[0089] 도5 에서 편향 학습부(230)는 데이터 압축부(220)가 보호된 속성에 대한 편향성이 저감된 특징값을 잠재 공간에 전사할 수 있도록 하는 학습용 구성으로, 데이터 압축부(220)의 학습이 완료된 이후에는 분류 장치에서 생략될 수 있다.

[0090] 결과적으로 도5 및 도6 에 도시된 분류 장치 및 방법은 입력 데이터로부터 특징값을 추출하여, 특징값에서 보호된 속성에 대한 잠재 변수(z)가 잠재 공간 상에 편향성이 저감된 패턴으로 전사할 수 있다.

[0091] 이는 일예로 직원 채용 추천이라는 상기의 예에서 성별이라는 보호된 속성에 대해 남성과 여성이 균등하게 추천되도록 입력 데이터를 분류할 수 있다.

[0092] 본 발명에 따른 방법은 컴퓨터에서 실행 시키기 위한 매체에 저장된 컴퓨터 프로그램으로 구현될 수 있다. 여기서 컴퓨터 판독가능 매체는 컴퓨터에 의해 액세스 될 수 있는 임의의 가용 매체일 수 있고, 또한 컴퓨터 저장 매체를 모두 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함하며, ROM(판독 전용 메모리), RAM(랜덤 액세스 메모리), CD(컴팩트 디스크)-ROM, DVD(디지털 비디오 디스크)-ROM, 자기 테이프, 플로피 디스크, 광데이터 저장장치 등을 포함할 수 있다.

[0093] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다.

[0094] 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 청구범위의 기술적 사상에 의해 정해져야 할 것이다.

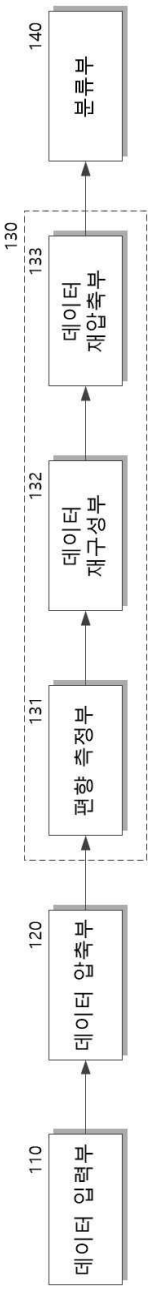
## *부호의 설명*

[0095] 210: 데이터 입력부 220: 데이터 압축부

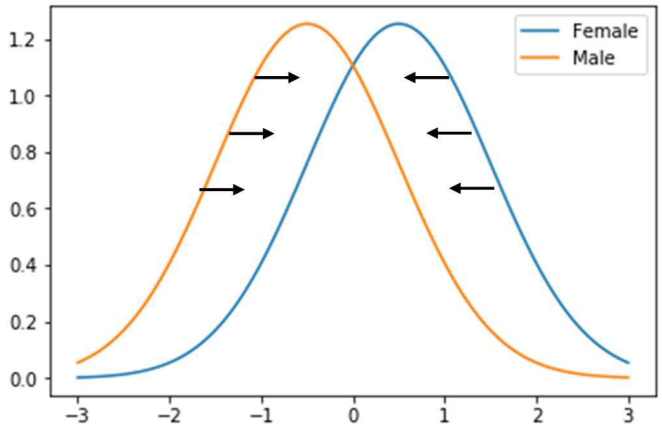
230: 편향 학습부 240: 분류부

## 도면

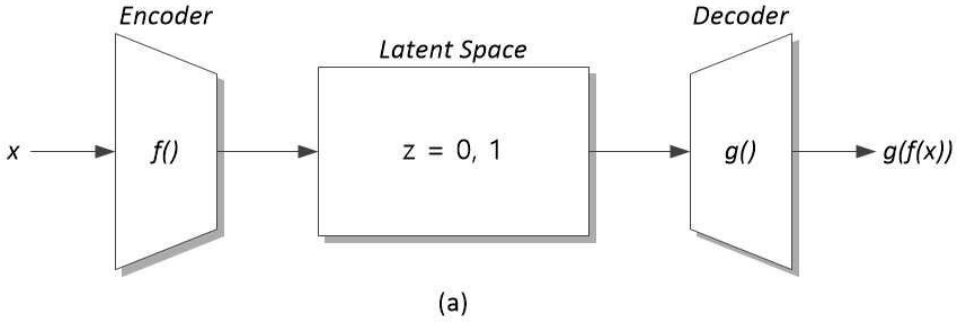
### 도면1



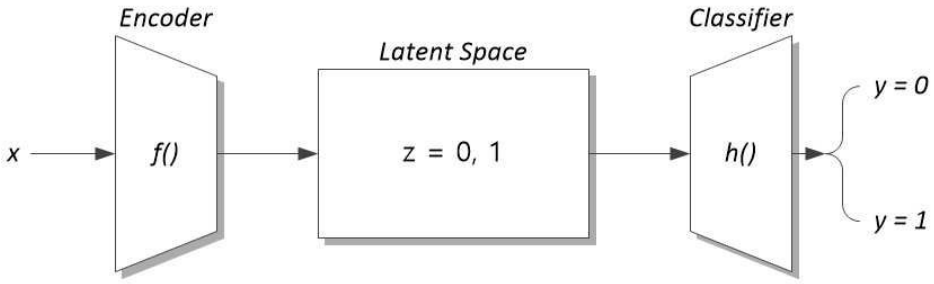
도면2


Female
Male
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3

도면3

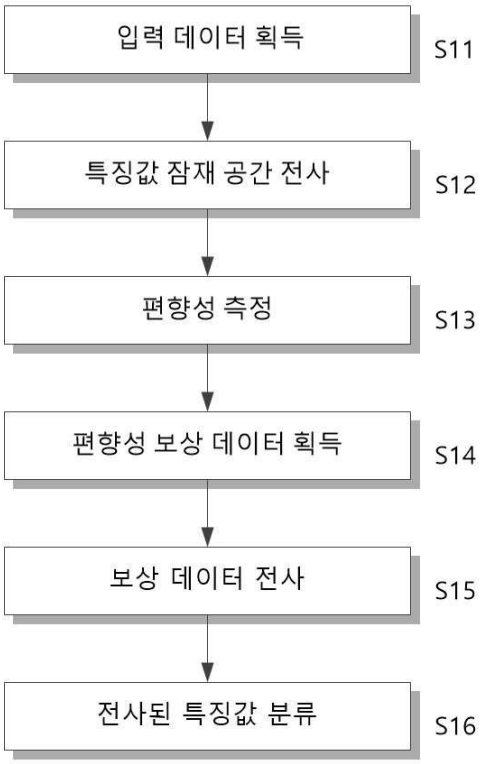

Encoder
Latent Space
Decoder
x
f()
z = 0, 1
g()
g(f(x))

(a)

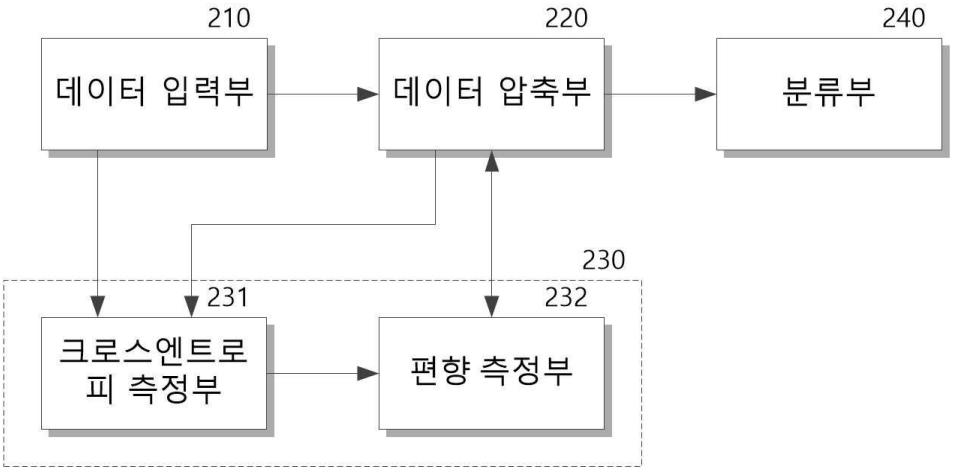

Encoder
Latent Space
Classifier
x
f()
z = 0, 1
h()
y = 0
y = 1

(b)

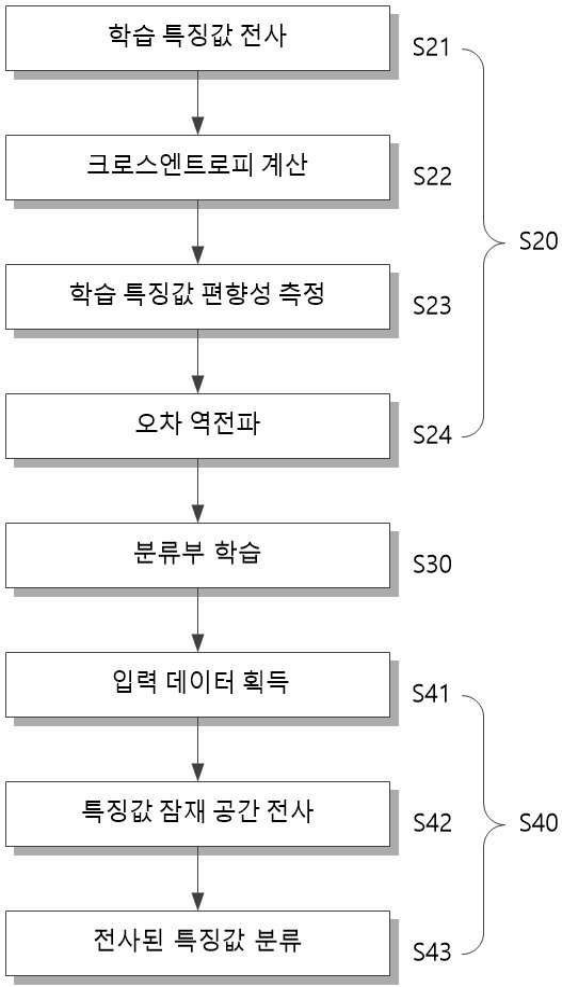
도면4


입력 데이터 획득
S11
특징값 잠재 공간 전사
S12
편향성 측정
S13
편향성 보상 데이터 획득
S14
보상 데이터 전사
S15
전사된 특징값 분류
S16

도면5


210
데이터 입력부
220
데이터 압축부
240
분류부
230
231
크로스엔트로피 측정부
232
편향 측정부

도면6


학습 특징값 전사
S21
크로스엔트로피 계산
S22
학습 특징값 편향성 측정
S23
오차 역전파
S24
S20
분류부 학습
S30
입력 데이터 획득
S41
특징값 잠재 공간 전사
S42
전사된 특징값 분류
S43
S40