



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2022년10월25일

(11) 등록번호 10-2458360

(24) 등록일자 2022년10월19일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) G06N 3/04 (2006.01)

(52) CPC특허분류
G06N 3/08 (2013.01)
G06N 3/0472 (2013.01)

(21) 출원번호 10-2021-0180528

(22) 출원일자 2021년12월16일

심사청구일자 2021년12월16일

(56) 선행기술조사문헌

Ryota Shimizu et al., "Balanced Mini-batch Training for Imbalanced Image Data Classification with Neural Network," 2018 First International Conference on Artificial Intelligence for Industries (2018.09.26.)*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

창원대학교 산학협력단

경상남도 창원시 의창구 창원대학로 20(퇴촌동)

(72) 발명자

차정원

경상남도 창원시 성산구 동산로 115, 105동 1407호 (상남동, 대동아파트)

성수진

경상남도 창원시 마산회원구 내서읍 광려천동로 22, 512호 (서광맨션)

(74) 대리인

김종선, 이형석

전체 청구항 수 : 총 5 항

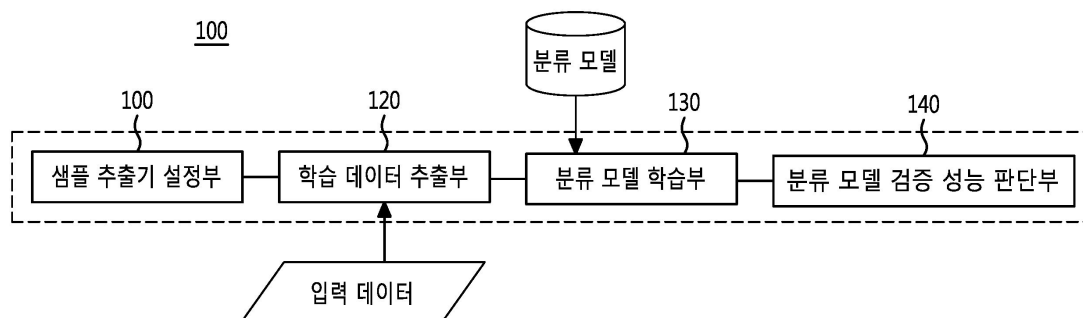
심사관 : 송근배

(54) 발명의 명칭 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법

(57) 요약

본 발명은 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것으로, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부 및 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부를 포함한다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호 1711134561

과제번호 2021-0-00354-001

부처명 과학기술정보통신부

과제관리(전문)기관명 정보통신기획평가원

연구사업명 차세대인공지능핵심원천기술개발(R&D)

기술 연구과제명 비정형 텍스트를 학습하여 쟁점별 사실과 논리적 근거 추론이 가능한 인공지능 원천

기 여 율 1/1

과제수행기관명 창원대학교 산학협력단

연구기간 2021.04.01 ~ 2021.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플별로 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부;

설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부; 및

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부;를 포함하며,

상기 샘플 추출기 설정부는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

상기 분류 모델 학습부는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

청구항 5

제1항에 있어서,

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 분류 모델 검증 성능 판단부를 포함하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치.

청구항 6

샘플 추출기 설정부에 의해, 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플별로 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 하나를 설정하는 단계;

학습 데이터 추출부에 의해, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 단계;

분류 모델 학습부에 의해, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계; 및

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계를 포함하며,

상기 분류 모델의 검증 성능 갱신 여부를 판단하는 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 단계를 포함하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법.

청구항 7

제6항에 있어서,

구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계는,

구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법.

청구항 8

삭제

청구항 9

삭제

발명의 설명

기술 분야

[0001] 본 발명은 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것으로, 보다 자세하게는 레이블 간 불균형이 존재하는 데이터 셋(dataset)에 대하여 전체 분류 성능을 향상시키기 위해 저빈도 레이블의 성능이 개선되도록 샘플 학습 데이터를 구성하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법에 관한 것이다.

배경 기술

[0002] 인공 신경망을 기반으로 하는 딥러닝이 다양한 분야에서 기존의 기계 학습을 능가하는 우수한 성능을 보여주고 있으나, 신경망 모델이 특정 작업을 수행하기 위해서는 모델 가중치의 학습이 선행되어야 한다.

[0003] 모델의 학습을 위해서는 학습 데이터가 필요하고 대부분의 분류를 위한 학습 데이터는 레이블 간의 불균형이 존재한다. 불균형이 존재하는 데이터를 사용하여 모델을 학습할 경우, 고빈도 레이블 샘플에 노이즈가 많거나 다른 레이블 샘플의 특징과 오버랩되는 등의 특수한 경우를 제외하고 저빈도 레이블의 성능이 떨어지기 때문에 레이블 샘플 수의 균형이 맞는 데이터를 사용한 모델보다 낮은 분류 성능이 나타난다.

[0004] 이러한 레이블 불균형으로 인해 발생하는 학습 편향을 해결하기 위해 학습 데이터를 표본 추출하여 정보량을 조정하거나, 손실함수를 개선하여 샘플에 가중치를 부여하는 방법 등이 제안되어 왔다.

[0005] 이와 관련하여, 한국공개특허 제2021-0120489호는 "레이블 데이터 생성 방법 및 이를 이용한 학습방법 및 장치"에 관하여 개시하고 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 상기와 같은 문제점을 해결하기 위해 발명된 것으로서, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법을 제공하는데 그 목적이 있다.

[0007] 또한, 본 발명은 레이블 샘플 수를 기반으로 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법을 제공하는데 그 목적이 있다.

과제의 해결 수단

[0008] 상기의 목적을 달성하기 위한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한

레이블 기반 샘플 추출 장치는 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 샘플 추출기 설정부; 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 학습 데이터 추출부; 및 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 분류 모델 학습부;를 포함한다.

[0009] 또한, 상기 샘플 추출기 설정부는 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 제2 샘플 추출기 중 적어도 어느 하나로 설정되는 것을 특징으로 한다.

[0010] 또한, 상기 샘플 추출기 설정부는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 것을 특징으로 한다.

[0011] 또한, 상기 분류 모델 학습부는 구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 한다.

[0012] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 분류 모델 검증 성능 판단부를 포함하는 것을 특징으로 한다.

[0013] 상기의 목적을 달성하기 위한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법은 샘플 추출기 설정부에 의해, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정하는 단계; 학습 데이터 추출부에 의해, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하는 단계; 및 분류 모델 학습부에 의해, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계;를 포함한다.

[0014] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계는, 구성된 샘플 학습 데이터를 사용하여 딥러닝 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장하는 것을 특징으로 한다.

[0015] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트하는 단계 이후에, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계를 포함하는 것을 특징으로 한다.

[0016] 또한, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단하는 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키는 단계를 포함하는 것을 특징으로 한다.

발명의 효과

[0017] 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치 및 그 방법은 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 이용하여 전체 학습 데이터에서 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성하고, 레이블 샘플 수를 기반으로 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트함으로써, 저빈도 레이블 샘플의 추출 확률을 높여 저빈도 레이블에 대한 학습량을 늘리고 이로 인해 과생되는 학습 편향을 억제할 수 있는 효과가 있다.

도면의 간단한 설명

[0018] 도 1은 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치의 구성을 설명하기 위한 도면이다.

도 2는 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법의 순서를 설명하기 위한 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0019] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다.
- [0020] 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 각 도면을 설명하면서 유사한 참조부호를 유사한 구성요소에 대해 사용하였다.
- [0021] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.
- [0022] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 출원에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.
- [0023] 이하, 첨부한 도면들을 참조하여, 본 발명의 바람직한 실시예를 보다 상세하게 설명하고자 한다. 이하, 도면상의 동일한 구성요소에 대해서는 동일한 참조부호를 사용하고 동일한 구성요소에 대해서 중복된 설명은 생략한다.
- [0025] 도 1은 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치의 구성을 설명하기 위한 도면이다.
- [0026] 도 1을 참조하여 설명하면, 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치(100)는 크게 샘플 추출기 설정부(110), 학습 데이터 추출부(120), 분류 모델 학습부(130) 및 분류 모델 검증 성능 판단부(140)를 포함한다.
- [0027] 샘플 추출기 설정부(110)는 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정한다.
- [0028] 샘플 추출기 설정부(110)는 이후 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습한 결과, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 설정된 샘플 추출기를 다른 샘플 추출기로 전환시킬 수 있다.
- [0029] 즉, 샘플 추출기 설정부(110)는 2가지 방식의 샘플 추출기를 가지며 현재 학습 경향에 따라 두 가지 샘플 추출기를 전환하여 이후 분류 모델의 학습 데이터를 추출에 사용되도록 한다.
- [0030] 여기서, 제1 샘플 추출기는 전체 학습 데이터에 동일한 추출 확률값을 할당한 후 비복원 추출을 수행하는 추출기이고, 제2 샘플 추출기는 전체 학습 데이터에 나타나는 정답 레이블 샘플을 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하는 추출기이다.
- [0031] 학습 데이터 추출부(120)는 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 이후 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성한다. 여기서, 입력되는 전체 학습 데이터는 분류 레이블이 부착된 모든 분야의 데이터로 딥러닝 기반 모델이 처리할 수 있는 형태로 정제된 상태이다.
- [0032] 학습 데이터 추출부(120)는 제1 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 모두 동일한 추출 확률값을 할당하고 비복원 추출을 수행하여 샘플 학습 데이터를 구성한다.
- [0033] 학습 데이터 추출부(120)는 제2 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하여 샘플 학습 데이터를 구성한다. 보다 자세하게, 학습 데이터 추출부(120)는 제2 샘플 추출기를 이용하는 경우 정답 레이블 샘플수를 토대로 저빈도 샘플의 추출 확률이 높아지는 것을 목적으로 하는 하기의 수식을 통해 샘플 각각에 대한 추출 확률을 계산할 수 있다.

[0034] 즉, 제2 샘플 추출기는 레이블 샘플 수의 역수를 추출 확률로 정의한다. 입력 문장 x_i 와 정답 레이블 y_i 로 구성된 학습 데이터 $D = \{(x_i, y_i)_{i=1}^N\}$ 이 주어질 때, i 번째 샘플의 추출 확률 $P(x_i | D, y_i)$ 을 구하는 수식은 다음과 같다.

수학식 1

$$N(a, b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases}$$

$$P(x_i | D, y_i) = \frac{1}{\sum_j N(j, y_i)}$$

[0035]

[0036] 분류 모델 학습부(130)는 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한다.

[0037] 분류 모델 학습부(130)는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장한다.

[0038] 분류 모델 검증 성능 판단부(140)는 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단한다.

[0040] 도 2는 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법의 순서를 설명하기 위한 순서도이다.

[0041] 도 2를 참조하여 설명하면, 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 방법은 앞서 설명한 본 발명에 따른 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치를 이용하는 것으로, 이하 중복되는 설명은 생략하기로 한다.

[0042] 먼저, 전체 학습 데이터의 레이블 샘플수를 기반으로 하는 제1 샘플 추출기 및 전체 학습 데이터에 나타나는 정답 레이블 샘플수를 기반으로 하는 제2 샘플 추출기 중 하나를 설정한다(S100).

[0043] S100 단계는 2가지 방식의 샘플 추출기를 가지며 현재 학습 경향에 따라 두 가지 샘플 추출기를 전환하여 이후 분류 모델의 학습 데이터를 추출에 사용되도록 한다.

[0044] 다음, 설정된 제1 샘플 추출기 또는 제2 샘플 추출기를 이용하여 전체 학습 데이터에서 이후 분류 모델의 학습에 사용할 샘플을 추출하여 샘플 학습 데이터를 구성한다(S200). 여기서, 입력되는 전체 학습 데이터는 분류 레이블이 부착된 모든 분야의 데이터로 딥러닝 기반 모델이 처리할 수 있는 형태로 정제된 상태이다.

[0045] S200 단계에서, 제1 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 모두 동일한 추출 확률값을 할당하고 비복원 추출을 수행하여 샘플 학습 데이터를 구성한다.

[0046] S200 단계에서, 제2 샘플 추출기를 이용하는 경우, 전체 학습 데이터에 나타나는 정답 레이블 샘플 수를 산출하고, 산출된 정답 레이블 샘플에 동일한 추출 확률값을 할당한 후 복원 추출을 수행하여 샘플 학습 데이터를 구성한다.

[0047] 다음, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한다(S300).

[0048] S300 단계는 구성된 샘플 학습 데이터를 사용하여 딥러닝(Deep Learning) 기반의 분류 모델을 학습하고 업데이트되는 분류 모델의 가중치를 저장한다.

[0049] 다음, 구성된 샘플 학습 데이터를 사용하여 분류 모델을 학습하고 업데이트한 결과에 대하여 분류 모델의 검증 성능 갱신 여부를 판단한다(S400).

[0050] S400 단계에서, 분류 모델의 검증 성능이 갱신된 것으로 판단되면, S300 단계로 돌아간다.

[0051] 한편, S400 단계에서, 분류 모델의 검증 성능이 갱신되지 않은 것으로 판단되면, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수를 비교한다(S500).

[0052] S500 단계에서, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수만큼 갱신에 실패하지 않은 경우 설정된 샘플 추출기를 다른 샘플 추출기로 전환시키기 위해 S100 단계로 돌아간다.

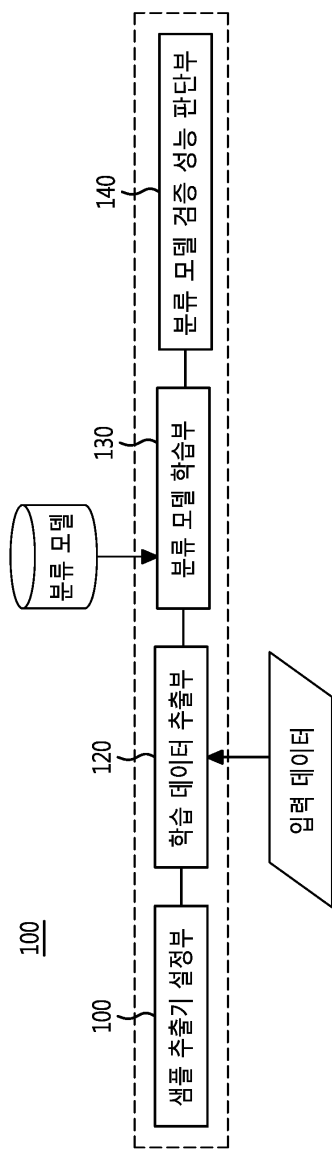
- [0053] 한편, S500 단계에서, 분류 모델의 검증 성능이 갱신 횟수와 사용자가 사전에 지정한 횟수만큼 갱신에 실패한 경우 전체 학습을 중단하고 최종 모델을 저장한다.
- [0055] 이상 본 명세서에서 설명한 기능적 동작과 본 주제에 관한 실시형태들은 본 명세서에서 개시한 구조들 및 그들의 구조적인 등가물을 포함하여 디지털 전자 회로나 컴퓨터 소프트웨어, 펌웨어 또는 하드웨어에서 또는 이들 중 하나 이상이 조합에서 구현 가능하다.
- [0056] 본 명세서에서 기술하는 주제의 실시형태는 하나 이상이 컴퓨터 프로그램 제품, 다시 말해 데이터 처리 장치에 의한 실행을 위하여 또는 그 동작을 제어하기 위하여 유형의 프로그램 매체 상에 인코딩되는 컴퓨터 프로그램 명령에 관한 하나 이상이 모듈로서 구현될 수 있다. 유형의 프로그램 매체는 전파형 신호이거나 컴퓨터로 판독 가능한 매체일 수 있다. 전파형 신호는 컴퓨터에 의한 실행을 위하여 적절한 수신기 장치로 전송하기 위한 정보를 인코딩하기 위하여 생성되는 예컨대 기계가 생성한 전기적, 광학적 또는 전자기 신호와 같은 인공적으로 생성된 신호이다. 컴퓨터로 판독 가능한 매체는 기계로 판독 가능한 저장장치, 기계로 판독 가능한 저장 기판, 메모리 장치, 기계로 판독 가능한 전파형 신호에 영향을 미치는 물질의 조합 또는 이들 중 하나 이상이 조합일 수 있다.
- [0057] 부가적으로, 본 특허문헌에서 기술하는 논리 흐름과 구조적인 블록도는 개시된 구조적인 수단의 지원을 받는 대응하는 기능과 단계의 지원을 받는 대응하는 행위 및/또는 특정한 방법을 기술하는 것으로, 대응하는 소프트웨어 구조와 알고리즘과 그 등가물을 구축하는 데에도 사용 가능하다.
- [0058] 그리고 본 기술한 설명은 본 발명의 최상의 모드를 제시하고 있으며, 본 발명을 설명하기 위하여, 그리고 당업자가 본 발명을 제작 및 이용할 수 있도록 하기 위한 예를 제공하고 있다. 이렇게 작성된 명세서는 그 제시된 구체적인 용어에 본 발명을 제한하는 것이 아니다.
- [0059] 따라서, 상술한 예를 참조하여 본 발명을 상세하게 설명하였지만, 당업자라면 본 발명의 범위를 벗어나지 않으면서도 본 예들에 대한 개조, 변경 및 변형을 가할 수 있다. 요컨대 본 발명이 의도하는 효과를 달성하기 위해 도면에 도시된 모든 기능 블록을 별도로 포함하거나 도면에 도시된 모든 순서를 도시된 순서 그대로 따라야만 하는 것은 아니며, 그렇지 않더라도 얼마든지 청구항에 기재된 본 발명의 기술적 범위에 속할 수 있음에 주의한다.

부호의 설명

- [0060] 100 : 불균형 데이터에 대한 딥러닝 분류 모델 성능을 향상시키기 위한 레이블 기반 샘플 추출 장치
- 110 : 샘플 추출기 설정부
- 120 : 학습 데이터 추출부
- 130 : 분류 모델 학습부
- 140 : 분류 모델 검증 성능 판단부

도면

도면1



도면2

