(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0102860
(43) 공개일자 2021년08월20일

(51) 국제특허분류(Int. Cl.)
G06N 7/02 (2019.01)
(52) CPC특허분류
G06N 7/023 (2013.01)
(21) 출원번호 10-2021-0020128
(22) 출원일자 2021년02월15일
심사청구일자 2021년02월15일
(30) 우선권주장
1020200016477 2020년02월11일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
최영진
경상북도 구미시 백산로 153, 102동 601호
최희련
서울특별시 강북구 삼양로123길 30(수유동)
이홍철
서울특별시 서초구 동광로19길 51, 701호(방배동, 래미안방배에버뉴)
(74) 대리인
김홍석

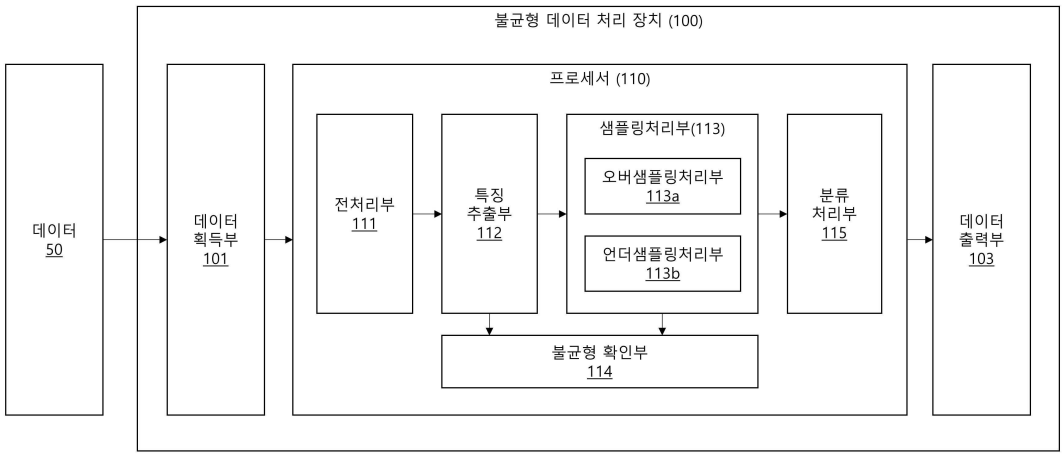
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **불균형 데이터 처리 장치 및 방법**

(57) 요 약

불균형 데이터 처리 장치 및 방법에 관한 것으로, 불균형 데이터 처리 장치는 적어도 하나의 데이터를 획득하되, 상기 적어도 하나의 데이터는 가상의 데이터를 더 생성하여 소수 표본에 추가하는 오버 샘플링부 및 다수 표본의 데이터 중에서 상기 소수 표본의 데이터와 근접한 데이터를 검출하고 상기 근접한 데이터를 다수 표본으로부터 제거하는 언더 샘플링부를 포함할 수 있다.

대 표 도 - 도1



명 세 서

청구범위

청구항 1

소수 표본에 속하는 적어도 하나의 데이터를 수신하고, 가상의 데이터를 더 생성하여 상기 소수 표본에 추가하는 오버 샘플링부; 및

다수 표본에 속하는 적어도 하나의 데이터를 수신하고, 상기 다수 표본의 데이터 중에서 상기 소수 표본의 데이터와 근접한 데이터를 검출하고 상기 근접한 데이터를 상기 다수 표본으로부터 제거하는 언더 샘플링부;를 포함하는 불균형 데이터 처리 장치.

청구항 2

제1항에 있어서,

상기 적어도 하나의 데이터는 음향 데이터를 포함하는 불균형 데이터 처리 장치.

청구항 3

제1항에 있어서,

음향 데이터에 대해 푸리에 변환을 수행하는 전처리부; 및

상기 전처리부에 의해 전처리된 음향 데이터로부터 특징을 추출하는 특징 추출부;를 더 포함하고,

상기 적어도 하나의 데이터는 추출된 상기 특징을 포함하는 불균형 데이터 처리 장치.

청구항 4

제1항에 있어서,

상기 적어도 하나의 데이터에 대해 또는 상기 오버 샘플링부 및 상기 언더 샘플링부에 의해 처리된 상기 소수 표본 및 상기 다수 표본의 데이터에 대해 t분포 확률적 임베딩을 기반으로 데이터 불균형을 확인하는 불균형 확인부;를 더 포함하는 불균형 데이터 처리 장치.

청구항 5

제1항에 있어서,

상기 오버 샘플링부 및 상기 언더 샘플링부에 의해 처리된 상기 소수 표본 및 상기 다수 표본의 데이터에 대해 적어도 하나의 학습 모델을 기반으로 클래스 분류를 수행하되, 상기 적어도 하나의 학습 모델은 랜덤 포레스트(Random Forest), 다층 퍼셉트론(MLP: Multi-Layer Perceptron), 심층 신경망(DNN: Deep Neural Network), 콘볼루션 신경망(CNN: Convolutional Neural Network) 및 순환 신경망(RNN: Recurrent Neural Network) 중 적어도 하나를 포함하는 분류 처리부;를 더 포함하는 불균형 데이터 처리 장치.

청구항 6

적어도 하나의 데이터를 획득하되, 상기 적어도 하나의 데이터는 소수 표본 또는 다수 표본 중 적어도 하나에 속하는 단계;

가상의 데이터를 더 생성하여 상기 소수 표본에 추가하여 오버 샘플링을 수행하는 단계; 및

상기 다수 표본의 데이터 중에서 상기 소수 표본의 데이터와 근접한 데이터를 검출하고 상기 근접한 데이터를 상기 다수 표본에서 제거하여 언더 샘플링을 수행하는 단계;를 포함하는 불균형 데이터 처리 방법.

청구항 7

제6항에 있어서,

상기 적어도 하나의 데이터는 음향 데이터를 포함하는 불균형 데이터 처리 방법.

청구항 8

제7항에 있어서,

상기 적어도 하나의 데이터는, 푸리에 변환이 수행된 음향 데이터로부터 추출된 특징을 포함하는 불균형 데이터 처리 방법.

청구항 9

제6항에 있어서,

상기 적어도 하나의 데이터에 대해 또는 상기 오버 샘플링부 및 상기 언더 샘플링부에 의해 처리된 상기 소수 표본 및 상기 다수 표본의 데이터에 대해 t분포 확률적 임베딩을 기반으로 데이터 불균형을 확인하는 단계;를 더 포함하는 불균형 데이터 처리 방법.

청구항 10

제5항에 있어서,

상기 오버 샘플링부 및 상기 언더 샘플링부에 의해 처리된 상기 소수 표본 및 상기 다수 표본의 데이터에 대해 적어도 하나의 학습 모델을 기반으로 클래스 분류를 수행하되, 상기 적어도 하나의 학습 모델은 랜덤 포레스트(Random Forest), 다층 퍼셉트론(MLP: Multi-Layer Perceptron), 심층 신경망(DNN: Deep Neural Network), 콘볼루션 신경망(CNN: Convolutional Neural Network) 및 순환 신경망(RNN: Recurrent Neural Network) 중 적어도 하나를 포함하는 단계;를 더 포함하는 불균형 데이터 처리 방법.

발명의 설명

기 술 분 야

[0001] 불균형 데이터 처리 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 정보 처리에 관한 하드웨어 기술의 발달에 따라 대량의 데이터의 신속한 처리가 가능해지고 있다. 이는 대량의 데이터를 기반으로 학습을 수행하고 학습 결과 획득된 규칙을 기반으로 다양한 정보를 획득하거나 연산 또는 제어 처리를 수행하는 기계학습 기술의 개발, 연구, 도입 및 이용을 더욱 촉진하고 있다. 이와 같은 기계학습 기술은 다양한 분야에서 이용되고 있다. 예를 들어, 데이터 검색, 날씨 등의 미래 데이터에 대한 예측, 음성 인식 및 텍스트 변환, 문자 인식, 언어 번역, 차량의 자율 주행, 소비자 성향 분석 및 이에 관한 서비스 제공, 게임, 스마트 그리드 및 사물 인터넷 등과 같이 다양한 분야에서 인공지능 기술이 개발 및 채용되어 이용되고 있다. 기계학습 중에서 지도 학습은 데이터 및 이에 대응하는 정답을 이용하여 학습을 수행하는 방법으로, 데이터를 어느 하나의 클래스로 분류시키는 분류 방법과 데이터의 연속으로 이루어진 함수를 추론하는 회귀 방법으로 나뉘어진다. 그런데, 지도 학습의 분류 방법은 데이터 불균형 시에는 그 성능이 저하되는 문제점이 있다. 구체적으로, 특정한 하나 또는 둘 이상의 클래스(다수 클래스)에 해당하는 데이터는 상대적으로 다수가 입력된 반면에 다른 하나 또는 둘 이상의 클래스(소수 클래스)에 해당하는 데이터는 상대적으로 소수가 입력되어 학습된 모델은, 추후 분류 대상 데이터가 입력되면 이들의 분류에 오류가 발생할 가능성이 높다. 즉, 소수 클래스에 해당하는 데이터가 분류 대상 데이터로 입력된 경우, 학습 모델은 이들 데이터를 다수 클래스로 오분류하기 쉽다. 그러므로, 이와 같은 클래스 간의 데이터 불균형은, 학습 모델의 예측 정확도는 향상시키는 반면에 그 재현율은 감소시켜 학습 모델의 전반적의 성능을 저하시키는 문제점이 있었다.

발명의 내용

해결하려는 과제

[0003] 클래스 간 데이터 불균형에 기인한 학습 모델의 성능 저하를 해결할 수 있는 불균형 데이터 처리 장치 및 방법을 제공하는 것을 해결하고자 하는 과제로 한다.

과제의 해결 수단

[0004] 상술한 과제를 해결하기 위하여 하기와 같이 불균형 데이터 처리 장치 및 방법이 제공된다.

[0005] 불균형 데이터 처리 장치는, 소수 표본에 속하는 적어도 하나의 데이터를 수신하고, 가상의 데이터를 더 생성하여 상기 소수 표본에 추가하는 오버 샘플링부 및 다수 표본에 속하는 적어도 하나의 데이터를 수신하고, 상기 다수 표본의 데이터 중에서 상기 소수 표본의 데이터와 근접한 데이터를 검출하고 상기 근접한 데이터를 상기 다수 표본으로부터 제거하는 언더 샘플링부를 포함할 수 있다.

[0006] 불균형 데이터 처리 방법은, 적어도 하나의 데이터를 획득하되, 상기 적어도 하나의 데이터는 소수 표본 또는 다수 표본 중 적어도 하나에 속하는 단계, 가상의 데이터를 더 생성하여 상기 소수 표본에 추가하여 오버 샘플링을 수행하는 단계 및 상기 다수 표본의 데이터 중에서 상기 소수 표본의 데이터와 근접한 데이터를 검출하고 상기 근접한 데이터를 상기 다수 표본에서 제거하여 언더 샘플링을 수행하는 단계를 포함할 수 있다.

발명의 효과

[0007] 상술한 불균형 데이터 처리 장치 및 방법에 의하면, 클래스 간 데이터 불균형에 따라 발생되는 여러 문제점을 해결할 수 있게 된다.

[0008] 상술한 불균형 데이터 처리 장치 및 방법에 의하면, 상대적으로 데이터가 많은 하나 이상의 다수 클래스 및 상대적으로 데이터가 적은 하나 이상의 소수 클래스가 존재하는 상황에서도 지도 학습 등을 위한 학습 모델의 성능 저하를 방지할 수 있게 된다.

[0009] 상술한 불균형 데이터 처리 장치 및 방법에 의하면, 시계열 데이터, 음향 데이터 또는 패턴 데이터 등의 분류 정확도를 높이면서도 재현율도 함께 높일 수 있는 장점이 있다.

도면의 간단한 설명

[0010] 도 1은 불균형 데이터 처리 장치의 일 실시예에 대한 블록도이다.

도 2는 데이터의 일 실시예에 대한 블록도이다.

도 3은 샘플링 처리 전의 데이터에 대한 t분포 확률적 임베딩 결과를 도시한 그래프 도면이다.

도 4는 샘플링 처리 후의 데이터에 대한 t분포 확률적 임베딩 결과를 도시한 그래프 도면이다.

도 5는 불균형 데이터 처리 방법의 일 실시예에 대한 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0011] 이하 명세서 전체에서 동일 참조 부호는 특별한 사정이 없는 한 동일 구성요소를 지칭한다. 이하에서 사용되는 '부'가 부가된 용어는, 소프트웨어 또는 하드웨어로 구현될 수 있으며, 설계자나 사용자의 고려나 임의적 선택에 따라 하나의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 복수의 '부'가 하나의 물리적 또는 논리적 부품으로 구현되거나, 하나의 '부'가 복수의 물리적 또는 논리적 부품들로 구현되는 것도 가능하다. 명세서 전체에서 어떤 부분이 다른 부분과 연결되어 있다고 할 때, 이는 어떤 부분과 다른 부분에 따라서 물리적 연결을 의미할 수도 있고, 또는 전기적으로 연결된 것을 의미할 수도 있다. 또한, 어떤 부분이 다른 부분을 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 부분 이외의 또 다른 부분을 제외하는 것이 아니며, 설계자의 선택에 따라서 또 다른 부분을 더 포함할 수 있음을 의미한다. 제1이나 제2 등의 용어는 하나의 부분을 다른 부분으로부터 구별하기 위해 사용되는 것으로, 특별한 기재가 없는 이상 이들이 순차적인 표현을 의미하는 것은 아니다. 또한 단수의 표현은 문맥상 명백하게 예외가 있지 않는 한, 복수의 표현을 포함할 수 있다.

[0012] 이하 도 1 내지 도 4를 참조하여 불균형 데이터 처리 장치의 일 실시예를 설명하도록 한다.

[0013] 도 1은 불균형 데이터 처리 장치의 일 실시예에 대한 블록도이고, 도 2는 데이터의 일 실시예에 대한 블록도이다.

[0014] 도 1 및 도 2에 도시된 바에 의하면, 불균형 데이터 처리 장치(100)는, 일 실시예에 있어서, 외부로부터 적어도 하나의 데이터(50)를 획득하는 데이터 획득부(101)와, 데이터 획득부(101)로부터 수신한 데이터(50)를 기반으로 불균형 데이터 처리, 학습 모델의 훈련 및/또는 훈련된 학습 모델 기반 결과 획득을 수행할 수 있는 프로세서(110)와, 프로세서(110)가 획득한 정보를 외부로 시각적 또는 청각적으로 출력 가능한 데이터 출력부(103)를 포

함할 수 있다. 여기서, 데이터 획득부(101), 데이터 출력부(103) 및 프로세서(110)는 일방으로 또는 쌍방으로 명령/지시나 데이터 등을 전송할 수 있되, 예를 들어, 케이블, 무선통신네트워크 및 회로라인 중 적어도 하나를 통해 명령/지시나 데이터 등을 순차적으로 또는 비순차적으로 전송할 수 있도록 마련될 수 있다.

[0015] 데이터 획득부(101)는 사용자의 조작에 따라, 메모리 장치나 다른 정보 처리 장치로부터의 수신에 따라 적어도 하나의 데이터(50)를 획득할 수 있다. 데이터 획득부(101)는, 실시예에 따라서, 키보드, 마우스, 태블릿, 터치스크린, 터치 패드, 트랙볼, 트랙패드, 스캐너 장치, 영상 촬영 모듈, 동작 감지 센서, 진동 센서, 수광 센서, 감압 센서, 근접 센서, 초음파 스캐너, 마이크로 폰, 외부의 다른 장치(일례로 휴대용 메모리 장치 등)로부터 데이터 등의 수신이 가능한 데이터 입출력 단자 및/또는 통신 모듈(일례로 랜카드, 근거리 통신 모듈 또는 이동통신 모듈 등) 등을 포함할 수 있다.

[0016] 데이터 획득부(101)가 획득한 데이터(50)는, 예를 들어, 도 2에 도시된 바와 같이 어느 하나의 표본(51, 이하 소수 표본)에 속하는 적어도 하나의 데이터(1 내지 4) 및/또는 다른 하나의 표본(52, 이하 다수 표본)에 속하는 적어도 하나의 데이터(5 내지 13)을 포함할 수 있다. 이 경우, 소수 표본(51)은 다수 표본(52)에 비해 상대적으로 적은 수의 데이터(1 내지 4)를 포함하고, 반대로 다수 표본(52)은 소수 표본(51)에 비해 상대적으로 많은 수의 데이터(5 내지 13)를 포함할 수 있다. 상황에 따라서 소수 표본(51)은 다수 표본(52)에 비해 포함하는 데이터가 극히 적을 수도 있다. 소수 표본(51)의 데이터(1 내지 4)는, 지도 학습 등을 통해 어느 하나의 클래스(이하 소수 클래스)로 분류될 수 있는 데이터이고, 다수 표본(52)의 데이터(5 내지 13)는 다른 하나의 클래스(이하 다수 클래스)로 분류될 수 있는 데이터이다. 즉, 소수 표본(51)은 소수 클래스에 대응하고, 다수 표본(52)은 다수 클래스에 대응한다. 이 경우, 소수 표본(51)의 데이터(1 내지 4)의 개수가 다수 표본(52)의 데이터(5 내지 13)의 개수에 비해 상당히 작으면, 양 표본(51, 52) 데이터(1 내지 4, 5 내지 13)의 불균형으로 인해 정확도는 증가해도 재현율이 감소하게 되는 문제점이 발생한다. 도 2에는 오직 두 개의 표본(51, 52)만이 도시되어 있으나, 이는 예시적인 것으로, 표본(51, 52)의 개수나 데이터(1 내지 13)의 개수는 이에 한정되는 것은 아니다. 실시예에 따라서 데이터 획득부(101)는 셋 이상의 표본에 속하는 데이터를 획득할 수도 있다. 여기서, 셋 이상의 표본 중 적어도 하나의 표본은 다른 둘 이상의 표본에 비해 상대적으로 소수의 데이터를 포함하고, 적어도 하나의 다른 표본은 다른 둘 이상의 표본에 대해 상대적으로 다수의 데이터를 포함할 수 있다. 또한, 실시예에 따라서 데이터(1 내지 13)는 후술하는 전처리부(111)에 의해 처리되거나 특징 추출부(112)에 의해 추출된 특징을 포함할 수도 있다.

[0017] 데이터(1 내지 13)는 적어도 하나의 패턴 데이터를 포함할 수 있으며, 패턴 데이터는 시간의 흐름에 따라 변화하는 시계열 데이터를 포함할 수 있다. 패턴 데이터는, 예를 들어, 정현파나 고조파를 포함할 수 있다. 실시예에 따라서, 데이터(1 내지 13)는 청진음 데이터(일례로 심음(心音)), 음성, 주변 음향, 잡음 또는 음악 등과 같은 음향 데이터를 포함할 수 있다. 여기서, 심음은 심장의 박동 소리로, 심음의 정상 여부를 통해 판막 질환 등과 같은 심장의 상태 이상 여부를 용이하게 측정할 수 있다. 이러한 심음은 시계열적 패턴 데이터이다. 데이터 획득부(101)가 획득한 데이터(1 내지 13)는 즉시 또는 순차적으로 프로세서(110)로 전달될 수 있다. 프로세서(110)로의 전달 전에 데이터(1 내지 13)는 일시적 또는 비일시적으로 저장부(미도시)에 저장될 수도 있다. 여기서 저장부는 주기억장치 및 보조기억장치 중 적어도 하나를 포함할 수 있다.

[0018] 프로세서(110)는 적어도 하나의 데이터(1 내지 13)를 획득하고 획득한 적어도 하나의 데이터(1 내지 13)에 대해 오버 샘플링 및 언더 샘플링을 수행하여 불균형된 데이터를 적절하게 처리할 수 있다.

[0019] 도 1에 도시된 일 실시예에 의하면, 프로세서(110)는 전처리부(111), 특징 추출부(112), 샘플링 처리부(113), 불균형 확인부(114) 및 분류 처리부(115)를 포함할 수 있다. 실시예에 따라서, 전처리부(111), 특징 추출부(112), 불균형 확인부(114) 및 분류처리부(115) 중 적어도 하나는 생략 가능하다. 여기서, 전처리부(111), 특징 추출부(112), 샘플링처리부(113), 불균형 확인부(114) 및 분류처리부(115)는 논리적으로 구분되는 것일 수도 있고 물리적으로 구분되는 것일 수도 있다.

[0020] 전처리부(111)는 데이터(1 내지 13)를 처리가 가능하거나 용이한 상태로 변환할 수 있다. 예를 들어, 전처리부(111)는 데이터(1 내지 13)가 주파수 기반의 데이터인 경우, 데이터(1 내지 13)에 대해 푸리에 변환을 적용하여 데이터(1 내지 13)의 전처리를 수행할 수도 있다.

[0021] 특징 추출부(112)는 데이터(1 내지 13), 일례로 음향 데이터로부터 적어도 하나의 특징(feature)를 추출할 수 있다. 특징은 소수 표본(51)에 대응하는 특징 및 다수 표본(52)에 대응하는 특징을 포함할 수 있다. 이 경우, 특징 추출부(112)는 시간, 주파수 및 통계 중 어느 하나를 기반으로 특징을 추출할 수도 있고, 또는 이들 중 복수를 기반으로 특징을 추출한 후 융합할 수도 있다. 특징 추출부(112)는 예를 들어, 멜 주파수 켑스트럴 계수

(MFCC: Mel-frequency cepstral coefficients), 멜 스펙트로그램(Mel spectrogram), 제로 크로싱율(zero crossing rate), 크로마(chroma) 및 콘트라스트(contrast) 방법 중 하나를 이용하거나 또는 이들 중 둘 이상을 조합 이용하여 특징을 추출할 수도 있다. 여기서, 멜 스펙트로그램은 입력되는 신호를 프레임으로 분할하고 이에 푸리에 변환을 적용하여 스펙트럼을 획득한 후, 멜 필터를 통해 여과시켜 데이터를 주파수 영역으로 표현하는 것이고, 멜 주파수 켑스트럴 계수는 멜 스펙트로그램에 대해 켑스트럴 분석을 통해 획득되는 벡터이다. 제로 크로싱율은 음성 신호나 시계열 데이터의 값이 0으로 변화하는 횟수를 측정한 값을 특징에 대한 값으로 이용하는 방법이다. 크로마는 음높이를 12음계를 나타내는 피치 클래스로 나타내는 벡터이다. 콘트라스트는 옥타브 기반의 스펙트럼 대역대의 특성을 나타내면서 특징을 통해 음악의 장르를 분류할 수 있도록 하는 방법이다.

[0022] 샘플링 처리부(113)는 주어진 데이터(1 내지 13)에 대해 새로운 데이터를 더 부가하거나 또는 주어진 데이터(1 내지 13) 중에서 일부를 제거하는 등 데이터(1 내지 13)에 대한 가공을 수행할 수 있다. 여기서, 주어진 데이터(1 내지 13)는, 데이터 획득부(101)가 획득한 데이터(1 내지 13)일 수도 있고, 전처리부(111)에 의해 전처리된 데이터일 수도 있으며, 및/ 또는 특징추출부(112)에서 추출된 데이터(1 내지 13)에 상응하는 특징일 수도 있다.

[0023] 일 실시예에 의하면, 샘플링 처리부(113)는 오버 샘플링 처리를 수행하는 오버 샘플링 처리부(113a) 및 언더 샘플링 처리를 수행하는 언더 샘플링 처리부(113b)를 포함할 수 있다. 오버 샘플링 처리부(113a)는 가상의 데이터를 더 생성하여 소수 표본(51)에 추가할 수 있다. 즉, 소수 표본(51)은 기존의 데이터(1 내지 4, 상술한 바와 같이 특징을 포함할 수도 있음) 외에 다른 가상의 데이터(또는 특징)를 더 포함하게 된다. 이 경우, 오버 샘플링 처리부(113a)는 소수 표본(51)의 하나 이상의 데이터(1 내지 4) 중에서 적어도 하나(일례로 중심이 되는 데이터)를 선택하고, 선택된 데이터와 다른 데이터를 가상의 선분으로 연결한 후, 가상의 선분 위에 적어도 하나의 가상의 데이터를 임의적으로 추가함으로써 적어도 하나의 가상의 데이터를 생성할 수 있다. 이에 따라 소수 표본(51)에 해당하는 데이터(1 내지 4)의 개수는 증가하게 된다. 언더 샘플링 처리부(113b)는 다수 표본(52)의 데이터(5 내지 13) 중에서 적어도 하나의 데이터를 삭제하여 언더 샘플링을 수행할 수 있다. 이 경우, 언더 샘플링 처리부(113b)는 표본(51, 52) 내의 데이터(1 내지 13, 상술한 바와 같이 다수 표본(52)에 속하는 특징을 포함할 수 있음) 간의 거리를 이용하여 언더 샘플링을 수행할 수 있다. 예를 들어, 언더 샘플링 처리부(113b)는 다수 표본(52)의 데이터(5 내지 13, 또는 추출된 특징) 중에서 소수 표본(51)의 데이터(1 내지 4, 또는 추출된 특징) 중 적어도 하나와 근접한 적어도 하나의 데이터를 획득하고, 소수 표본(51)의 데이터(1 내지 4) 중 적어도 하나와 근접한 적어도 하나의 데이터를 다수 표본(52)에서 제거하여 다수 표본(51)의 데이터(5 내지 13)의 개수를 감소시킬 수 있다. 실시예에 따라서 샘플링 처리부(113)는 불균형 확인부(114)에 의한 데이터 불균형 발생 여부의 확인 결과에 응하여 상술한 오버 샘플링 및 언더 샘플링 중 적어도 하나를 수행할 수도 있다.

[0024] 도 3은 샘플링 처리 전의 데이터에 대한 t분포 확률적 임베딩 결과를 도시한 그래프 도면이고, 도 4는 샘플링 처리 후의 데이터에 대한 t분포 확률적 임베딩 결과를 도시한 그래프 도면이다.

[0025] 불균형 확인부(114)는 데이터 획득부(101)가 획득한 데이터(1 내지 13), 전처리부(111)에 의해 전처리된 데이터 및/또는 특징추출부(112)에 의해 추출된 특징을 기반으로 클래스 불균형 여부를 판단할 수 있다. 일 실시예에 의하면, 불균형 확인부(114)는, 오버 샘플링부(113a)에 의해 처리된 소수 표본(51)의 데이터(1 내지 4) 및 언더 샘플링부(113b)에 의해 처리된 다수 표본(52)의 데이터에(5 내지 13)에 대해 t분포 확률적 임베딩(T-SNE: t-Stochastic Nearest Neighbor)를 적용함으로써 클래스 불균형 여부를 판단할 수도 있다. t분포 확률적 임베딩은 고차원의 데이터를 저차원으로 축소하는 기계 학습 방법으로, 3 이상의 차원의 데이터를 2차원 또는 3차원의 데이터로 축소하여 도 3 및 도 4에 도시된 바와 같이 그래프의 형태로 시각화할 수 있다. 도 3에 도시된 바와 같이 오버 샘플링 또는 언더 샘플링이 수행되지 않은 데이터의 경우, 클래스 불균형으로 인해 데이터가 혼재한 양상을 나타내고 있다. 반면에 오버 샘플링부(113a) 및 언더 샘플링부(113b) 각각에 의해 오버 샘플링 및 언더 샘플링이 수행된 이후에는 도 4에 도시된 바와 같이 좀더 개선된 형태로 분류되고 있다.

[0026] 분류 처리부(115)는, 일 실시예에 따르면, 소수 표본(51)의 데이터(1 내지 4)와, 오버 샘플링부(113a)에 의해 소수 표본(51)에 부가된 가상의 데이터와, 언더 샘플링부(113b)에 의해 일부가 제거된 다수 표본(52)의 데이터(5 내지 13 중 적어도 하나)를 기반으로 클래스 분류를 수행할 수 있다. 이를 위해 분류 처리부(115)는 적어도 하나의 학습 모델을 이용할 수 있다. 여기서, 적어도 하나의 학습 모델은, 예를 들어, 심층 신경망(DNN, Deep Neural Network), 콘볼루션 신경망(CNN, Convolutional Neural Network), 순환 신경망(RNN, Recurrent Neural Network), 심층 신뢰 신경망(DBN, Deep Belief Network), 심층 Q-네트워크(Deep Q-Networks), 장단기 메모리(LSTM: Long short term memory), 랜덤 포레스트(Random Forest), 다층 퍼셉트론(Multi-layer Perceptron), 서포트 벡터 머신(SVM: support vector machine), 생성적 적대 신경망(GAN: Generative Adversarial Network) 및/또는 조건적 생성적 적대 신경망(cGAN: Conditional GAN) 등과 같은 학습 모델을 기반으로 구축된 것이되, 하

나의 학습 모델을 이용하여 구현된 것일 수도 있고 또는 둘 이상의 학습 모델을 조합하여 구현된 것일 수도 있다. 한편, 상술한 학습 모델들은 예시적인 것으로 분류 처리부(115)는 이들 외에도 설계자나 사용자가 고려 가능한 적어도 하나의 다른 학습 모델을 기반으로 클래스 분류를 수행할 수도 있다.

[0027] 상술한 프로세서(100)는, 예를 들어, 중앙 처리 장치(CPU, Central Processing Unit), 마이크로 컨트롤러 유닛(MCU, Micro Controller Unit), 마이컴(Micom, Micro Processor), 애플리케이션 프로세서(AP, Application Processor), 전자 제어 유닛(ECU, Electronic Controlling Unit) 및/또는 각종 연산 처리 및 제어 신호의 생성이 가능한 다른 전자 장치 등을 포함할 수 있다. 이들 장치는, 예를 들어, 하나 또는 둘 이상의 반도체 칩 등을 이용하여 구현된 것일 수 있다. 또한, 프로세서(100)는 저장부에 저장된 프로그램을 실행시켜, 미리 정의된 연산, 판단, 처리 및/또는 제어 동작 등을 수행할 수 있으며, 저장부에 저장된 프로그램은 사용자나 설계자 등에 의해 직접 작성 및 수정된 것일 수도 있고, 다른 메모리 장치 등으로부터 전달, 설치 및 갱신된 것일 수도 있으며, 및/또는 유선 또는 무선 통신 네트워크를 통해 접속 가능한 전자 소프트웨어 유통망을 통하여 획득 또는 갱신된 것일 수도 있다.

[0028] 데이터 출력부(103)는 샘플링 처리부(113)의 처리 결과, 불균형 확인부(114)가 획득한 불균형에 관한 그래프나, 분류 처리부(115)에 의함 분류 처리 결과 등을 시각적 또는 청각적으로 외부로 출력하거나, 및/또는 외부의 메모리 장치나 다른 정보 처리 장치로 전달할 수 있다. 데이터 출력부(103)는 예를 들어, 디스플레이, 프린터 장치, 스피커 장치, 영상 출력 단자, 데이터 입출력 단자 및/또는 통신 모듈 등을 포함할 수 있으나, 이에 한정되는 것은 아니다.

[0029] 상술한 불균형 데이터 처리 장치(100)는, 하나의 정보처리장치를 이용하여 또는 둘 이상의 정보처리장치를 조합 이용하여 구현될 수 있다. 여기서, 정보 처리 장치는, 실시예에 따라서, 서버용 컴퓨터 장치, 데스크톱 컴퓨터, 랩톱 컴퓨터, 스마트 폰, 태블릿 피씨, 스마트 시계, 두부 장착형 디스플레이(HMD: Head Mounted Display) 장치, 내비게이션 장치, 휴대용 게임기, 개인용 디지털 보조기(PDA: Personal Digital Assistant), 디지털 텔레비전, 전자 광고판, 셋 톱 박스, 가전 기기(냉장고나 세탁기 등), 인공 지능 음향 재생 장치, 유인 이동체(승용차, 버스나 이륜차 등의 차량 등), 무인 이동체(로봇 청소기 등), 건설기계, 유인 비행체, 무인 비행체, 가정용 또는 산업용 로봇 및/또는 산업용 기계 등을 포함할 수 있으나, 이에 한정되는 것은 아니다. 하나 또는 둘 이상의 정보처리장치는 예시된 바 이외에도 데이터의 기반의 학습 모델의 훈련 및/또는 훈련된 학습 모델 기반의 정보 획득을 수행할 수 있는 다양한 장치를 포함할 수 있다.

[0030] 이하 도 5를 참조하여 불균형 데이터 처리 방법의 일 실시예를 설명하도록 한다.

[0031] 도 5는 불균형 데이터 처리 방법의 일 실시예에 대한 블록도이다.

[0032] 도 5에 도시된 바에 의하면, 먼저 적어도 하나의 데이터가 획득되되, 획득된 적어도 하나의 데이터는 복수의 표본 중 어느 하나에 속할 수 있다(200). 이 경우, 복수의 표본은 포함하는 데이터의 개수에 따라서 적어도 하나의 소수 표본 또는 적어도 하나의 다수 표본으로 구분될 수 있다. 적어도 하나의 데이터는 적어도 하나의 소수 표본 및 적어도 하나의 다수 표본 중 적어도 하나에 속할 수 있다. 일 실시예에 의하면, 데이터는 심음과 같은 음향 데이터 또는 음향 데이터로부터 추출된 특징을 포함할 수 있다. 음향 데이터로부터 추출된 특징은, 예를 들어, 음향 데이터에 대해 푸리에 변환 등을 이용하여 전처리를 수행한 후, 시간, 주파수 및 통계 중 어느 하나를 기반으로 전처리된 데이터로부터 추출된 것일 수 있다. 특징의 추출은, 멜 주파수 켑스트럴 계수, 멜 스펙트로그램, 제로 크로싱, 크로마 및/또는 콘트라스트 등을 기반으로 수행 가능하다.

[0033] 필요에 따라서 데이터에 대해 불균형 여부가 확인될 수 있다(402). 불균형 여부는 데이터에 대해 t분포 확률적 임베딩을 적용함으로써 수행될 수 있다. 데이터에 대한 불균형 여부의 확인은 도 5에 도시된 바와 같이 오버 샘플링 과정(204) 및 언더 샘플링 과정(206) 전에 수행될 수도 있고, 오버 샘플링 과정(204) 및 언더 샘플링 과정(206) 중 적어도 하나와 동시에 또는 순차적으로 수행될 수도 있으며, 및/또는 오버 샘플링 과정(204) 및 언더 샘플링 과정(206)이 처리된 이후에 수행될 수도 있다. 실시예에 따라서, 데이터 불균형 여부의 확인은 오버 샘플링 및 언더 샘플링이 처리되기 전의 데이터에 대해서도 수행되고, 아울러 오버 샘플링 및 언더 샘플링이 처리된 이후의 데이터에 대해서도 수행될 수도 있다.

[0034] 소수 표본에 대해서 오버 샘플링이 수행될 수 있다(404). 오버 샘플링은 소수 표본에 가상의 데이터를 더 추가함으로써 수행될 수 있다. 일 실시예에 따르면, 오버 샘플링은 소수 표본의 하나 이상의 데이터 중에서 적어도 하나를 선택하고, 선택된 데이터와 다른 데이터를 가상의 선분으로 연결한 후, 가상의 선분 위에 적어도 하나의 가상의 데이터를 임의적으로 추가함으로써 수행될 수도 있다.

[0035] 오버 샘플링에 후행하거나, 선행하거나 또는 오버 샘플링과 동시에 다수 표본에 대해서 언더 샘플링이 수행될 수 있다(406). 언더 샘플링은 다수 표본의 데이터의 일부를 제거하는 방식으로 수행될 수도 있다. 구체적으로 예를 들어, 언더 샘플링은 다수 표본에 속하는 적어도 하나의 데이터 중에서 소수 표본의 적어도 하나의 데이터와 거리가 가까운 적어도 하나의 데이터를 선택하고, 선택된 적어도 하나의 데이터를 다수 표본에서 삭제함으로써 수행될 수도 있다.

[0036] 오버 샘플링(404) 및 언더 샘플링(406)이 수행되면, 실시예에 따라서, 오버 샘플링된 소수 표본의 데이터와 언더 샘플링된 다수 표본의 데이터를 기반으로 클래스 분류가 더 수행될 수도 있다(408). 클래스 분류는 적어도 하나의 학습 모델을 기반으로 수행될 수 있으며, 예를 들어, 심층 신경망, 콘볼루션 신경망, 순환 신경망, 심층 신뢰 신경망, 심층 Q-네트워크, 장단기 메모리, 랜덤 포레스트, 다층 퍼셉트론, 서포트 벡터 머신, 생성적 적대 신경망 및 조건적 생성적 적대 신경망 중 적어도 하나의 학습 모델을 기반으로 수행될 수 있다. 클래스 분류는 생략 가능하다.

[0037] 상술한 실시예에 따른 불균형 데이터 처리 방법은, 컴퓨터 장치에 의해 구동될 수 있는 프로그램의 형태로 구현될 수 있다. 여기서 프로그램은, 프로그램 명령, 데이터 파일 및 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 프로그램은 기계어 코드나 고급 언어 코드를 이용하여 설계 및 제작된 것일 수 있다. 프로그램은 상술한 방법을 구현하기 위하여 특별히 설계된 것일 수도 있고, 컴퓨터 소프트웨어 분야에서 통상의 기술자에게 기 공지되어 사용 가능한 각종 함수나 정의를 이용하여 구현된 것일 수도 있다. 또한, 여기서, 컴퓨터 장치는, 프로그램의 기능을 실현 가능하게 하는 프로세서나 메모리 등을 포함하여 구현된 것일 수 있으며, 필요에 따라 통신 장치를 더 포함할 수도 있다. 상술한 불균형 데이터 처리 방법을 구현하기 위한 프로그램은, 컴퓨터에 의해 판독 가능한 기록 매체에 기록될 수 있다. 컴퓨터에 의해 판독 가능한 기록 매체는, 예를 들어, 솔리드 스테이트 드라이브(SSD), 롬, 램 또는 플래시 메모리 등과 같은 반도체 저장 장치, 하드 디스크나 플로피 디스크 등과 같은 자기 디스크 저장 매체, 콤팩트 디스크나 디브이디 등과 같은 광 기록 매체, 플롭티컬 디스크 등과 같은 자기-광 기록 매체 및 자기 테이프 등 컴퓨터 등의 호출에 따라 실행되는 특정 프로그램을 저장 가능한 적어도 한 종류의 물리적 장치를 포함할 수 있다.

[0038] 이상 불균형 데이터 처리 장치 및 방법의 일 실시예에 대해 설명하였으나, 불균형 데이터 처리 장치 및 방법은 오직 상술한 실시예에 한정되는 것은 아니다. 해당 기술 분야에서 통상의 지식을 가진 자가 상술한 실시예를 기초로 수정 및 변형하여 구현 가능한 장치나 방법 역시 상술한 불균형 데이터 처리 장치 및 방법의 일례가 될 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로, 기능, 부품 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 또는 치환되더라도 상술한 불균형 데이터 처리 장치 및 방법의 일 실시예가 될 수 있다.

부호의 설명

[0039] 100: 불균형 데이터 처리 장치 101: 데이터 획득부

111: 전처리부 112: 특징 추출부

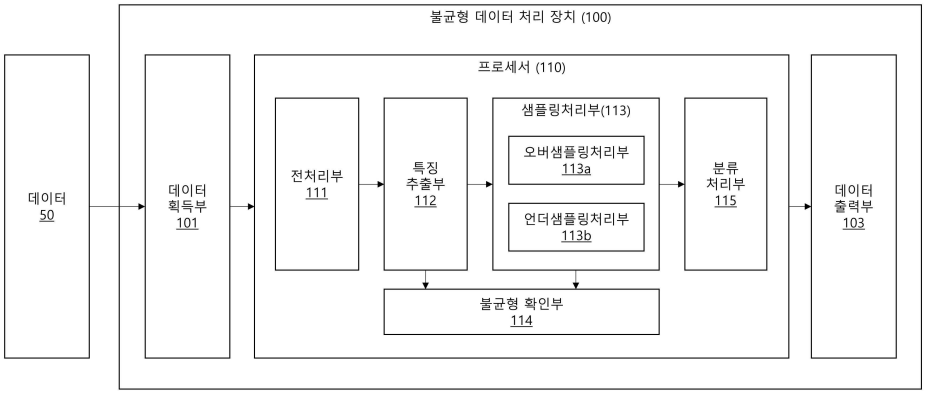
113: 샘플링 처리부 113a: 오버 샘플링 처리부

113b: 언더 샘플링 처리부 114: 불균형 확인부

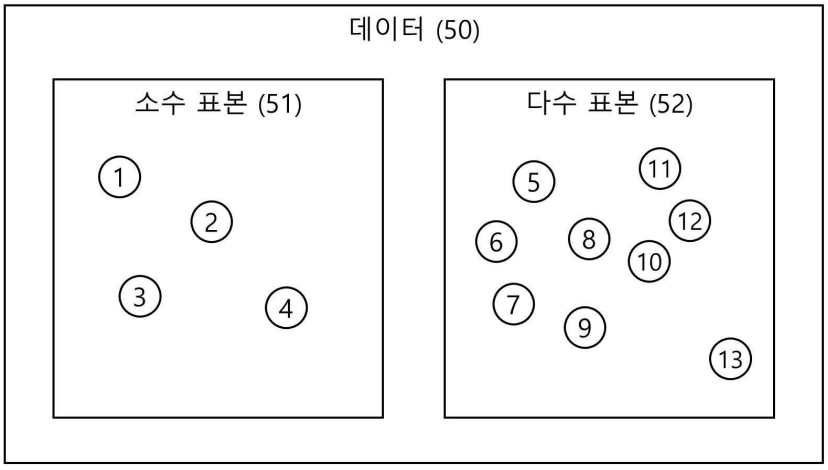
115: 분류 처리부

도면

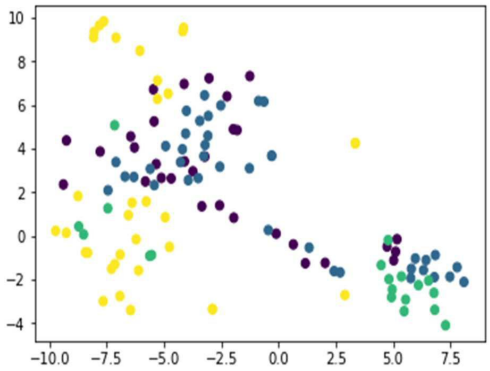
도면1


불균형 데이터 처리 장치 (100)
데이터
50
데이터
획득부
101
프로세서 (110)
전처리부
111
특징
추출부
112
샘플링처리부(113)
오버샘플링처리부
113a
언더샘플링처리부
113b
분류
처리부
115
데이터
출력부
103
불균형 확인부
114

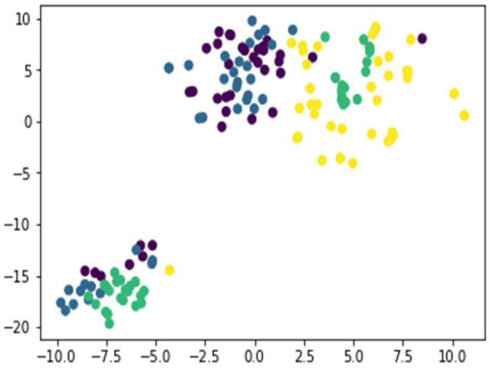
도면2


데이터 (50)
소수 표본 (51)
다수 표본 (52)
1
2
3
4
5
6
7
8
9
10
11
12
13

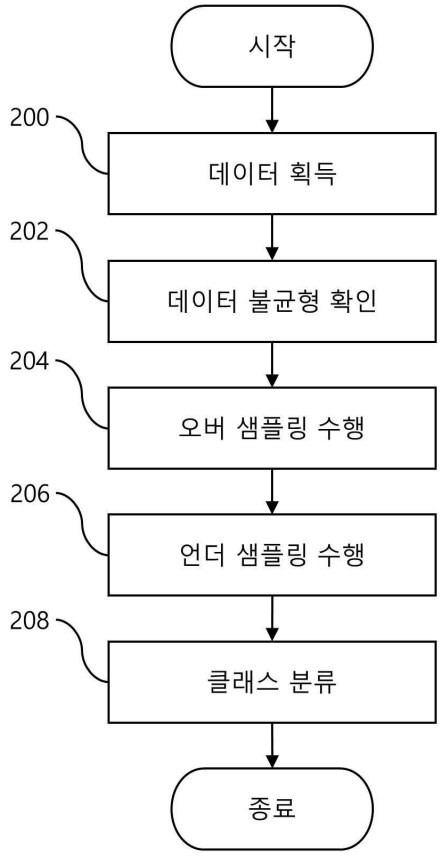
도면3


10
8
6
4
2
0
−2
−4
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

도면4


10
5
0
−5
−10
−15
−20
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

도면5


시작
200
데이터 획득
202
데이터 불균형 확인
204
오버 샘플링 수행
206
언더 샘플링 수행
208
클래스 분류
종료