



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0132804
(43) 공개일자 2022년10월04일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) G06N 5/02 (2006.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06N 5/02 (2019.01)
(21) 출원번호 10-2021-0037802
(22) 출원일자 2021년03월24일
심사청구일자 2021년03월24일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
권오병
서울특별시 송파구 올림픽로4길 42, 21동 1202호 (잠실동, 우성아파트)
김정훈
서울특별시 동대문구 천장산로7길 46-4, 203호 (이문동)
(74) 대리인
김연권

전체 청구항 수 : 총 20 항

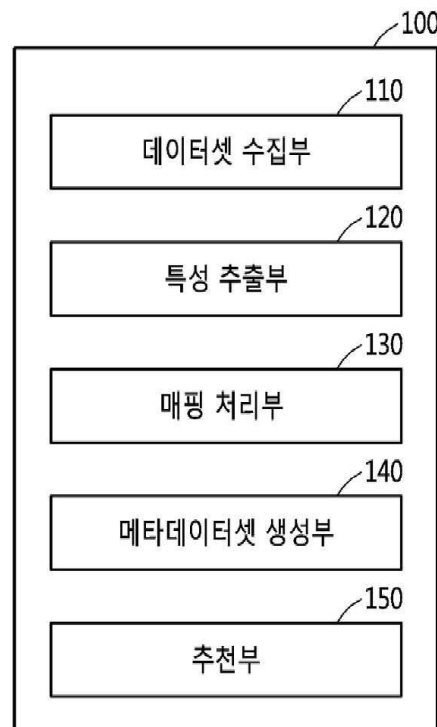
(54) 발명의 명칭 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치 및 방법

(57) 요약

본 발명은 오픈 데이터셋의 데이터셋 특성들을 이용하여 메타데이터셋을 생성하고, 사용자가 사용하려는 데이터셋이 입력될 경우, 메타데이터셋을 이용하여 사용자가 사용하려는 데이터셋에 대하여 최적의 알고리즘 종류 및 파라미터 값 그리고 데이터 샘플링 방법을 추천하는 기술에 관한 것으로, 본 발명의 일실시예에 따른 메타데이터

(뒷면에 계속)

대표도 - 도1



셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는 데이터셋 수집부, 상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 특성 추출부, 상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 매핑 처리부, 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 메타데이터셋 생성부 및 사용자로부터 입력된 사용자 데이터셋에 대하여 상기 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 추천부를 포함할 수 있다.

명세서

청구범위

청구항 1

오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는 데이터셋 수집부;

상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 특성 추출부;

상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 매핑 처리부;

상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 메타데이터셋 생성부; 및

사용자로부터 입력된 사용자 데이터셋에 대하여 상기 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 추천부를 포함하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 2

제1항에 있어서,

상기 특성 추출부는 상기 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 상기 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 3

제2항에 있어서,

상기 특성 추출부는 상기 수집된 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 복수의 데이터셋 특성을 추출하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 4

제3항에 있어서,

상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 5

제3항에 있어서,

상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 6

제3항에 있어서,

상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 7

제2항에 있어서,

상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 상기 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 상기 산출된 샘플링 방법 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 8

제2항에 있어서,

상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘 정확도를 산출하고, 상기 산출된 분류 알고리즘 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 9

제8항에 있어서,

상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성에 대한 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 상기 분류 알고리즘 정확도를 산출하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 10

제1항에 있어서,

상기 메타데이터셋 생성부는 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 적용된 상기 전처리된 복수의 데이터셋 특성을 필터링하고, 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘에 투입하여 기계학습하고, 상기 기계학습에 기반하여 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 11

제10항에 있어서,

상기 메타데이터셋 생성부는 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋과 상기 생성된 선정 룰 베이스를 포함하는 메타데이터셋을 생성하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 12

제11항에 있어서,

상기 생성된 메타데이터셋을 저장하는 메타데이터셋 저장부를 더 포함하는 것을 특징으로 하는
메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 13

제1항에 있어서,

상기 특성 추출부는 상기 입력된 사용자 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 14

제13항에 있어서,

상기 특성 추출부는 상기 입력된 사용자 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 복수의 데이터셋 특성을 추출하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 15

제14항에 있어서,

상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하고, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하고, 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 16

제15항에 있어서,

상기 추천부는 상기 사용자 데이터셋의 전처리된 복수의 데이터셋 특성을 인식하고, 상기 생성된 메타데이터셋에서 상기 인식된 복수의 데이터셋 특성과 관련된 복수의 데이터셋 특성을 확인하고, 상기 확인된 복수의 데이터셋 특성과 상기 생성된 선정 룰 베이스에 기반하여 상기 맞춤형 샘플링 방법 및 상기 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치.

청구항 17

데이터셋 수집부에서, 오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는 단계;

특성 추출부에서, 상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계;

매핑 처리부에서, 상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 단계;

메타데이터셋 생성부에서, 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 단계; 및

추천부에서, 사용자로부터 입력된 사용자 데이터셋에 대하여 상기 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 단계를 포함하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법.

청구항 18

제17항에 있어서,

상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계는,

상기 수집된 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 상기 복수의 데이터셋 특성을 추출하는 단계;

상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 상기 결측치를 처리하며, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계; 및

상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계를 포함하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법.

청구항 19

제17항에 있어서,

상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 단계는,

상기 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 상기 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 상기 산출된 샘플링 방법 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑하는 단계; 및

상기 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 분류 알고리즘 정확도를 산출하고, 상기 산출된 분류 알고리즘 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑하는 단계를 포함하는 것을 특징으로 하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법.

청구항 20

제17항에 있어서,

상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 단계는,

상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 적용된 상기 전처리된 복수의 데이터셋 특성을 필터링

하고, 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘에 투입하여 기계학습하고, 상기 기계학습에 기반하여 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하는 단계; 및

상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋과 상기 생성된 선정 룰 베이스를 포함하는 메타 데이터셋을 생성하는 단계를 포함하는 것을 특징으로하는

메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법.

발명의 설명

기술 분야

[0001] 본 발명은 메타데이터셋을 이용하여 최적화된 인공지능 알고리즘과 관련하여 샘플링 방법 및 분류 알고리즘을 추천하는 기술적 사상에 관한 것으로, 구체적으로, 오픈 데이터셋의 데이터셋 특성들을 이용하여 메타데이터셋을 생성하고, 사용자가 사용하려는 데이터셋이 입력될 경우, 메타데이터셋을 이용하여 사용자가 사용하려는 데이터셋에 대하여 최적의 알고리즘 종류 및 파라미터 값 그리고 데이터 샘플링 방법을 추천하는 기술에 관한 것이다.

배경 기술

[0002] 모든 산업분야에서 기업의 경쟁 우위를 선점하기 위해 기계학습 알고리즘을 통한 고객 서비스 및 기업 내 의사결정의 정확도를 향상시키기 위한 노력들이 지속되고 있다.

[0003] 기존 인공지능 개발 업체들은 성능이 뛰어난 알고리즘을 개발하기 위해 반복 실험을 하면서 시간과 컴퓨터 하드웨어 자원을 무분별하게 소비하고 있다.

[0004] 즉, 최적화된 인공지능 알고리즘을 선정하기 위해 많은 반복 실험과 시간 및 자원을 소모하고 있다.

[0005] 또한, 많은 인공지능 알고리즘 연구자들이 뛰어난 성능의 알고리즘을 발표하고 있지만, 데이터셋의 특성에 따라 성능 차이가 존재할 수 있다.

[0006] 최근, 제조 및 생산 분야에서 인공지능이 탑재된 스마트팩토리 구축을 위한 노력을 하고 있으며, 인공지능 구축 초기에 사전 지식이 부족한 인공지능 컨설팅업체 및 스마트 팩토리 구축에 활용될 수 있어 인공지능 알고리즘과 관련된 샘플링 방법 및 분류 알고리즘 선정은 중요한 이슈일 수 밖에 없다.

[0007] 또한, 기계학습을 근거로 한 솔루션, 시스템, 서비스 개발을 위한 플랫폼의 핵심 컴포넌트로서 인공지능 알고리즘의 자동적 선택과 추천은 필요하며, 기계학습 기반 개발 시장은 지속적으로 크게 성장하는 추세이다.

[0008] 데이터셋에 따라 알고리즘의 성능이 차이나는 것은 개발된 알고리즘이 특수한 상황에서 발생한 데이터셋에 대해 뛰어난 성능을 낼 수 있도록 개발되어 있기 때문이다.

[0009] 분류 알고리즘의 정확성은 알고리즘의 특성과 하이퍼파라미터뿐 만 아니라 데이터셋의 특성에 의해 결정될 수 있다.

[0010] 인공지능에 사용되는 분류 알고리즘들은 분류 전략이 다르기 때문에 특정한 데이터셋의 특성에 따라 효율적인 알고리즘이 존재하기 마련이므로, 데이터셋의 특성 파악이 중요할 수 있다.

[0011] 메타특징(meta-feature)이라고 하는 데이터셋의 특성과 분류 알고리즘 성능과의 연관성에 대한 연구가 아직 미흡하고, 다중 클래스(multi-class)의 불균형 특성을 반영하는 메타특징에 대한 연구가 이루어지지 않고있다.

선행기술문헌

특허문헌

[0012] (특허문헌 0001) 한국등록특허 제10-2103902호, "컴포넌트 기반 머신러닝 자동화 예측 장치 및 방법"

(특허문헌 0002) 한국등록특허 제10-2098897호, "기계학습 지식 및 자동화된 기계 학습 절차 기반의 자가 학습 시스템"

(특허문헌 0003) 한국등록특허 제10-1864286호, "머신 러닝 알고리즘을 이용하는 방법 및 장치"

(특허문헌 0004) 미국공개특허 제2020/0210775호, "DATA STITCHING AND HARMONIZATION FOR MACHINE LEARNING"

발명의 내용

해결하려는 과제

- [0013] 본 발명은 오픈 데이터셋의 데이터셋 특성들을 이용하여 메타데이터셋을 생성하고, 사용자가 사용하려는 데이터셋이 입력될 경우, 메타데이터셋을 이용하여 사용자가 사용하려는 데이터셋에 대하여 최적의 알고리즘 종류 및 파라미터 값 그리고 데이터 샘플링 방법을 추천하는 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치를 제공하는 것을 목적으로 한다.
- [0014] 본 발명은 오픈 데이터셋의 데이터셋 특성을 사전에 학습하여 메타데이터화함에 따라 생성된 메타데이터셋에 근거하여 최고의 성능을 나타내는 최적의 알고리즘 종류와 파라미터값, 그리고 데이터 샘플링 방법을 추천하는 것을 목적으로 한다.
- [0015] 본 발명은 데이터 사이언스, 인공지능 개발 업체에서 반복적인 실험을 줄여 시간과 비용을 줄일 수 있고, 하드웨어의 사용을 줄여 친환경적인 인공지능 개발 및 데이터 분석이 가능하도록 지원하는 것을 목적으로 한다.
- [0016] 본 발명은 사용자가 데이터 사이언스가 적용되지 않은 분야에 진입할 경우 사전 지식이 부족하여 어려움을 겪을 수 있는데, 오픈 데이터베이스를 통해 수집된 오픈 데이터셋들을 통해 사전에 학습된 데이터셋들을 통해 유사한 기계학습 알고리즘을 자동적으로 찾아서 추천하는 것을 목적으로 한다.
- [0017] 본 발명은 초보 데이터 과학자, 인공지능 개발자들이 데이터에 대한 노하우가 부족한 경우, 사전에 학습된 데이터셋들을 통해 유사한 기계학습 알고리즘을 자동적으로 찾아서 추천하는 방식을 참고하여 인공지능 알고리즘과 관련된 샘플링 방법 및 분류 알고리즘의 선택 방법에 대한 노하우 획득 지원을 목적으로 한다.

과제의 해결 수단

- [0018] 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는 데이터셋 수집부, 상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 특성 추출부, 상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 매핑 처리부, 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 메타데이터셋 생성부 및 사용자로부터 입력된 사용자 데이터셋에 대하여 상기 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 추천부를 포함할 수 있다.
- [0019] 상기 특성 추출부는 상기 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 상기 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0020] 상기 특성 추출부는 상기 수집된 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 복수의 데이터셋 특성을 추출할 수 있다.
- [0021] 상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0022] 상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0023] 상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및

소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리할 수 있다.

- [0024] 상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 상기 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 상기 산출된 샘플링 방법 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑할 수 있다.
- [0025] 상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘 정확도를 산출하고, 상기 산출된 분류 알고리즘 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑할 수 있다.
- [0026] 상기 매핑 처리부는 상기 전처리된 복수의 데이터셋 특성에 대한 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 상기 분류 알고리즘 정확도를 산출할 수 있다.
- [0027] 상기 메타데이터셋 생성부는 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 적용된 상기 전처리된 복수의 데이터셋 특성을 필터링하고, 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘에 투입하여 기계학습하고, 상기 기계학습에 기반하여 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성할 수 있다.
- [0028] 상기 메타데이터셋 생성부는 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋과 상기 생성된 선정 룰 베이스를 포함하는 메타데이터셋을 생성할 수 있다.
- [0029] 본 발명의 일실시예에 따르면 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 상기 생성된 메타데이터셋을 저장하는 메타데이터셋 저장부를 더 포함할 수 있다.
- [0030] 상기 특성 추출부는 상기 입력된 사용자 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출할 수 있다.
- [0031] 상기 특성 추출부는 상기 입력된 사용자 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 복수의 데이터셋 특성을 추출할 수 있다.
- [0032] 상기 특성 추출부는 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하고, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하고, 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0033] 상기 추천부는 상기 사용자 데이터셋의 전처리된 복수의 데이터셋 특성을 인식하고, 상기 생성된 메타데이터셋에서 상기 인식된 복수의 데이터셋 특성과 관련된 복수의 데이터셋 특성을 확인하고, 상기 확인된 복수의 데이터셋 특성과 상기 생성된 선정 룰 베이스에 기반하여 상기 맞춤형 샘플링 방법 및 상기 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0034] 본 발명의 일실시예에 따르면 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 데이터셋 수집부에서, 오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는 단계, 특성 추출부에서, 상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계, 매핑 처리부에서, 상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 단계, 메타데이터셋 생성부에서, 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 단계 및 추천부에서, 사용자로부터 입력된 사용자 데이터셋에 대하여 상기 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 단계를 포함할 수 있다.

[0035] 상기 수집된 오픈 데이터셋의 복수의 데이터셋 특성을 추출하고, 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계는, 상기 수집된 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 상기 결정된 복수의 훈련 데이터셋으로부터 상기 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 상기 복수의 데이터셋 특성을 추출하는 단계, 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 상기 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 상기 결측치를 처리하며, 상기 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 상기 결측치를 처리함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계 및 상기 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 상기 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 상기 존재하는 클래스 불균형을 해소함에 따라 상기 추출된 복수의 데이터셋 특성을 전처리하는 단계를 포함할 수 있다.

[0036] 상기 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑하는 단계는, 상기 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 상기 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 상기 산출된 샘플링 방법 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑하는 단계 및 상기 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 상기 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 분류 알고리즘 정확도를 산출하고, 상기 산출된 분류 알고리즘 정확도에 따라 상기 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑하는 단계를 포함할 수 있다.

[0037] 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하고, 상기 생성된 선정 룰 베이스 및 상기 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성하는 단계는, 상기 매핑된 샘플링 방법과 상기 매핑된 분류 알고리즘에 적용된 상기 전처리된 복수의 데이터셋 특성을 필터링하고, 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘에 투입하여 기계학습하고, 상기 기계학습에 기반하여 상기 추천 샘플링 방법 및 상기 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성하는 단계 및 상기 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋과 상기 생성된 선정 룰 베이스를 포함하는 메타데이터셋을 생성하는 단계를 포함할 수 있다.

발명의 효과

[0038] 본 발명은 오픈 데이터셋의 데이터셋 특성들을 이용하여 메타데이터셋을 생성하고, 사용자가 사용하려는 데이터셋이 입력될 경우, 메타데이터셋을 이용하여 사용자가 사용하려는 데이터셋에 대하여 최적의 알고리즘 종류 및 파라미터 값 그리고 데이터 샘플링 방법을 추천하는 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치를 제공할 수 있다.

[0039] 본 발명은 오픈 데이터셋의 데이터셋 특성을 사전에 학습하여 메타데이터화함에 따라 생성된 메타데이터셋에 근거하여 최고의 성능을 나타내는 최적의 알고리즘 종류와 파라미터값, 그리고 데이터 샘플링 방법을 추천할 수 있다.

[0040] 본 발명은 데이터 사이언스, 인공지능 개발 업체에서 반복적인 실험을 줄여 시간과 비용을 줄일 수 있고, 하드웨어의 사용을 줄여 친환경적인 인공지능 개발 및 데이터 분석이 가능하도록 지원할 수 있다.

[0041] 본 발명은 사용자가 데이터 사이언스가 적용되지 않은 분야에 진입할 경우 사전 지식이 부족하여 어려움을 겪을 수 있는데, 오픈 데이터베이스를 통해 수집된 오픈 데이터셋들을 통해 사전에 학습된 데이터셋들을 통해 유사한 기계학습 알고리즘을 자동적으로 찾아서 추천할 수 있다.

[0042] 본 발명은 초보 데이터 과학자, 인공지능 개발자들이 데이터에 대한 노하우가 부족한 경우, 사전에 학습된 데이터셋들을 통해 유사한 기계학습 알고리즘을 자동적으로 찾아서 추천하는 방식을 참고하여 인공지능 알고리즘과 관련된 샘플링 방법 및 분류 알고리즘의 선택 방법에 대한 노하우 획득 지원할 수 있다.

도면의 간단한 설명

- [0043] 도 1은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치를 설명하는 도면이다.
- 도 2는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 메타데이터셋을 생성하는 절차를 설명하는 도면이다.
- 도 3 내지 도 5는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법을 설명하는 도면이다.
- 도 6은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 수집하는 데이터셋 특성의 구조를 설명하는 도면이다.
- 도 7은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 생성하는 메타데이터셋을 예시하는 도면이다.
- 도 8은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 분류 알고리즘과 관련하여 성능 평가한 결과를 예시하는 도면이다.
- 도 9은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 샘플링 방법과 관련하여 성능 평가한 결과를 예시하는 도면이다.

발명을 실시하기 위한 구체적인 내용

- [0044] 이하, 본 문서의 다양한 실시 예들이 첨부된 도면을 참조하여 기재된다.
- [0045] 실시 예 및 이에 사용된 용어들은 본 문서에 기재된 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 해당 실시 예의 다양한 변경, 균등물, 및/또는 대체물을 포함하는 것으로 이해되어야 한다.
- [0046] 하기에 다양한 실시 예들을 설명에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.
- [0047] 그리고 후술되는 용어들은 다양한 실시 예들에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다.
- [0048] 도면의 설명과 관련하여, 유사한 구성요소에 대해서는 유사한 참조 부호가 사용될 수 있다.
- [0049] 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [0050] 본 문서에서, "A 또는 B" 또는 "A 및/또는 B 중 적어도 하나" 등의 표현은 함께 나열된 항목들의 모든 가능한 조합을 포함할 수 있다.
- [0051] "제1," "제2," "첫째," 또는 "둘째," 등의 표현들은 해당 구성요소들을, 순서 또는 중요도에 상관없이 수식할 수 있고, 한 구성요소를 다른 구성요소와 구분하기 위해 사용될 뿐 해당 구성요소들을 한정하지 않는다.
- [0052] 어떤(예: 제1) 구성요소가 다른(예: 제2) 구성요소에 "(기능적으로 또는 통신적으로) 연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 상기 어떤 구성요소가 상기 다른 구성요소에 직접적으로 연결되거나, 다른 구성요소(예: 제3 구성요소)를 통하여 연결될 수 있다.
- [0053] 본 명세서에서, "~하도록 구성된(또는 설정된)(configured to)"은 상황에 따라, 예를 들면, 하드웨어적 또는 소프트웨어적으로 "~에 적합한," "~하는 능력을 가지는," "~하도록 변경된," "~하도록 만들어진," "~를 할 수 있는," 또는 "~하도록 설계된"과 상호 호환적으로(interchangeably) 사용될 수 있다.
- [0054] 어떤 상황에서는, "~하도록 구성된 장치"라는 표현은, 그 장치가 다른 장치 또는 부품들과 함께 "~할 수 있는" 것을 의미할 수 있다.
- [0055] 예를 들면, 문구 "A, B, 및 C를 수행하도록 구성된(또는 설정된) 프로세서"는 해당 동작을 수행하기 위한 전용 프로세서(예: 임베디드 프로세서), 또는 메모리 장치에 저장된 하나 이상의 소프트웨어 프로그램들을 실행함으로써, 해당 동작들을 수행할 수 있는 범용 프로세서(예: CPU 또는 application processor)를 의미할 수 있다.
- [0056] 또한, '또는' 이라는 용어는 배타적 논리합 'exclusive or' 이기보다는 포함적인 논리합 'inclusive or' 를 의미한다.

- [0057] 즉, 달리 언급되지 않는 한 또는 문맥으로부터 명확하지 않는 한, 'x가 a 또는 b를 이용한다' 라는 표현은 포함적인 자연 순열들(natural inclusive permutations) 중 어느 하나를 의미한다.
- [0058] 이하 사용되는 '..부', '..기' 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어, 또는, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.
- [0060] 도 1은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치를 설명하는 도면이다.
- [0061] 도 1은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치의 구성요소를 예시한다.
- [0062] 도 1을 참고하면, 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치(100)는 데이터셋 수집부(110), 특성 추출부(120), 매핑처리부(130), 메타데이터셋 생성부(140) 및 추천부(150)를 포함한다.
- [0063] 본 발명의 일실시예에 따르면 데이터셋 수집부(110)는 오픈 데이터 베이스로부터 오픈 데이터셋을 수집한다.
- [0064] 여기서, 오픈 데이터 베이스는 공개된 데이터를 저장하고 있는 데이터 베이스를 지칭할 수 있다.
- [0065] 즉, 데이터셋 수집부(110)는 사전 데이터학습을 위하여 오픈 데이터 베이스로부터 오픈 데이터셋을 수집하는데, 오픈 데이터셋은 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치(100)를 이용하는 사용자에게 의해 오픈 데이터베이스 기반으로 입력되는 데이터일 수 있고, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치(100)가 오픈 데이터 베이스에 접근하여 수집하는 데이터일 수 있다.
- [0066] 한편, 데이터셋 수집부(110)는 사용자에게 의해서 테스트를 위한 사용자 데이터셋이 입력되는 경우에도 사용자 데이터셋을 수집할 수 있다.
- [0067] 본 발명의 일실시예에 따르면 특성 추출부(120)는 오픈 데이터셋 또는 사용자 데이터셋의 복수의 데이터셋 특성을 추출하고, 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0068] 일례로, 특성 추출부(120)는 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출하고, 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0069] 또한, 특성 추출부(120)는 입력된 사용자 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출하고, 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0070] 즉, 특성 추출부(120)는 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출할 수 있다.
- [0071] 구체적으로, 특성 추출부(120)는 사용자 데이터셋 또는 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 결정된 복수의 훈련 데이터셋으로부터 복수의 데이터셋 특성을 추출할 수 있다.
- [0072] 여기서, 분류된 복수의 폴드(fold) 중 하나에 해당하는 데이터셋은 복수의 훈련 데이터셋을 이용한 훈련 과정이 완료되어 선정된 분류 알고리즘의 테스트 데이터셋으로 이용될 수 있다.
- [0073] 본 발명의 일실시예에 따르면 특성 추출부(120)는 복수의 데이터셋 특성에서 결측치 및 클래스 불균형이 존재하는지 판단하고, 결측치 및 클래스 불균형을 해소하기 위한 전처리 과정을 수행할 수 있다.
- [0074] 일례로, 특성 추출부(120)는 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 결측치를 처리함에 따라 추출된 복수의 데이터셋 특성을 전처리할 수 있다.

- [0075] 본 발명의 일실시예에 따르면 특성 추출부(120)는 복수의 데이터셋 특성이 추출된 데이터셋에서 결측치가 존재하고, 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 결측치를 처리함에 따라 추출된 복수의 데이터셋 특성을 전처리할 수 있다.
- [0076] 일례로, 특성 추출부(120)는 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 클래스 불균형을 해소함에 따라 복수의 데이터셋 특성을 전처리할 수 있다.
- [0077] 본 발명의 일실시예에 따르면 매핑 처리부(130)는 특성 추출부(120)에 의해 전처리된 복수의 데이터셋 특성에 따른 샘플링 방법 및 분류 알고리즘을 매핑할 수 있다.
- [0078] 일례로, 매핑 처리부(130)는 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 산출된 샘플링 방법 정확도에 따라 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑할 수 있다.
- [0079] 본 발명의 일실시예에 따르면 매핑 처리부(130)는 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘 정확도를 산출하고, 산출된 분류 알고리즘 정확도에 따라 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑할 수 있다.
- [0080] 여기서, 샘플링 방법 정확도 및 분류 알고리즘 정확도는 F1-Score 및 G-mean이 사용되고, F1-score는 긍정 참값 비율(True positive Rate)과 긍정 예측 값(Positive Predictive Vale)를 고려하여 결정되는 값이고, G-mean은 참 긍정 값과 참 부정 값을 고려한 산술평균으로 볼 수 있다.
- [0081] 예를 들어, F1-score는 정밀도(precision)와 재현율(recall)을 산출하여 조화 평균을 사용하는 것을 특징으로 한다.
- [0082] F1-score는 하기 수학적 식 1을 이용하여 산출될 수 있고, G-mean은 하기 수학적 식 2를 이용하여 산출될 수 있다.
- [0084] [수학적 식 1]
- [0085]
$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$
- [0087] 수학적 식 1은 정밀도(precision)와 재현율(recall)을 고려하여 산출될 수 있다.
- [0089] [수학적 식 2]
- [0090]
$$G - mean = \sqrt{TP_{rate} \cdot TN_{rate}}$$
- [0092] 수학적 식 2는 참 긍정(True Positive, TP) 비율과 참 부정(True Negative, TN) 비율을 이용하여 산출될 수 있다.
- [0093] 본 발명의 일실시예에 따르면 매핑 처리부(130)는 전처리된 복수의 데이터셋 특성에 대한 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 분류 알고리즘 정확도를 산출할 수 있다.
- [0094] 즉, 매핑 처리부(130)는 분류 알고리즘의 특성과 하이퍼파라미터에 대하여 정밀도(precision)와 재현율(recall) 또는 참 긍정(True Positive, TP) 비율과 참 부정(True Negative, TN) 비율을 고려하여 분류 알고리즘 정확도를 산출할 수 있다.
- [0095] 본 발명의 일실시예에 따르면 메타데이터셋 생성부(140)는 매핑 처리부(130)에 의해 매핑된 샘플링 방법과 매핑된 분류 알고리즘에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성

할 수 있다.

- [0096] 또한, 메타데이터셋 생성부(140)는 생성된 선정 룰 베이스 및 전처리된 복수의 데이터셋 특성을 포함하는 메타데이터셋을 생성할 수 있다.
- [0097] 여기서, 선정 룰 베이스는 추가적으로 테스트 데이터셋이 입력될 경우에 추천 샘플링 방법과 추천 분류 알고리즘을 선별하기 위한 기준이 될 수 있다.
- [0098] 본 발명의 일실시예에 따르면 메타데이터셋 생성부(140)는 매핑된 샘플링 방법과 매핑된 분류 알고리즘에 적용된 전처리된 복수의 데이터셋 특성을 필터링하고, 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 추천 샘플링 방법 및 추천 분류 알고리즘에 투입하여 기계학습하고, 기계학습에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성할 수 있다. 여기서, 기계학습에 기반하는 것은 기계학습 결과를 이용하는 것을 나타낼 수 있다.
- [0099] 또한, 메타데이터셋 생성부(140)는 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋과 상기 생성된 선정 룰 베이스를 포함하는 메타데이터셋을 생성할 수 있다.
- [0100] 본 발명의 일실시예에 따르면 추천부(150)는 사용자로부터 입력된 사용자 데이터셋에 대하여 생성된 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0101] 즉, 추천부(150)는 사용자가 사용하려는 사용자 데이터셋을 입력하면 입력된 데이터셋을 스캔하여 사용자 데이터셋 특성을 특성 추출부(120)를 통해 자동인식하고, 사전 학습되어 생성된 메타데이터셋에 기반하여 사용자 데이터셋 특성에 적합한 최적의 알고리즘 종류, 파라미터 값 그리고 데이터를 샘플링하기 위한 샘플링 방법을 자동적으로 추천 또는 선택하여 사용하도록 할 수 있다.
- [0102] 일례로, 추천부(150)는 사용자 데이터셋의 전처리된 복수의 데이터셋 특성을 인식하고, 생성된 메타데이터셋에서 인식된 복수의 데이터셋 특성과 관련된 복수의 데이터셋 특성을 확인하고, 확인된 복수의 데이터셋 특성과 생성된 선정 룰 베이스에 기반하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0103] 본 발명의 일실시예에 따르면 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치(100)는 메타데이터셋 생성부(140)에 의해 생성된 메타데이터셋을 저장하는 메타데이터셋 저장부(미도시)를 더 포함할 수 있다.
- [0104] 본 발명의 일실시예에 따르면 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치(100)는 사용자 데이터셋이 입력된 경우, 메타데이터셋 저장부(미도시)에 저장된 메타데이터셋을 바로 호출하여 사용자 데이터셋에 맞는 샘플링 방법 및 분류 알고리즘을 추천하는데 활용할 수 있다.
- [0105] 따라서, 본 발명은 오픈 데이터셋의 데이터셋 특성들을 이용하여 메타데이터셋을 생성하고, 사용자가 사용하려는 데이터셋이 입력될 경우, 메타데이터셋을 이용하여 사용자가 사용하려는 데이터셋에 대하여 최적의 알고리즘 종류 및 파라미터 값 그리고 데이터 샘플링 방법을 추천하는 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치를 제공할 수 있다.
- [0106] 또한, 본 발명은 오픈 데이터셋의 데이터셋 특성을 사전에 학습하여 메타데이터베이스에 따라 생성된 메타데이터셋에 근거하여 최고의 성능을 나타내는 최적의 알고리즘 종류와 파라미터값, 그리고 데이터 샘플링 방법을 추천할 수 있다.
- [0108] 도 2는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 메타데이터셋을 생성하는 절차를 설명하는 도면이다.
- [0109] 도 2는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 오픈 데이터셋을 이용하여 메타데이터셋을 생성하는 과정을 예시한다.
- [0110] 도 2를 참고하면, 단계(S201)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 오픈 데이터베이스로부터 수집한 오픈 데이터셋을 복수의 폴드로 분류한다.
- [0111] 예를 들어, 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 하나의 오픈 데이터셋을 10개의 폴드로 분류하고, 9개의 폴드의 데이터셋은 훈련 데이터셋으로 이용하고, 하나의 폴드에서의 테스트 데이터셋으로 이용한다.

- [0112] 단계(S202)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 9개의 훈련 데이터셋 및 하나의 테스트 데이터셋으로부터 복수의 데이터셋 특성을 추출한다.
- [0113] 여기서, 추출된 복수의 데이터셋 특성은 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함할 수 있다.
- [0114] 즉, 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 하나의 데이터셋에 대하여 복수의 데이터셋 특성을 추출할 수 있다. 또한, 추출된 복수의 데이터셋 특성은 메타데이터셋에 포함될 수 있다.
- [0115] 단계(S203)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 추출된 데이터셋 특성의 결측치를 보간할 수 있다.
- [0116] 즉, 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 결측치가 존재하는 변수가 수치형인 경우 해당 클래스의 평균값을 이용하여 결측치를 보간하고, 결측치가 존재하는 변수가 명목형인 경우 해당 클래스의 최빈값을 이용하여 결측치를 보간할 수 있다.
- [0117] 단계(S204)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 클래스 불균형 해소 방법을 적용하여 클래스 불균형을 해소할 수 있다.
- [0118] 클래스 불균형은 분류 성능에 많은 영향을 미치는 것으로 분류해야 할 목표 변수(target variable)의 속성이 불균형을 이루는 것을 나타내는데, 기존 알고리즘을 수정하여 다수 클래스 불균형을 완화하는 비용 민감 학습 방법(cost sensitive learning)이 있고, 비용 민감 학습은 클래스의 분포에 맞게 데이터 추출 방법과 다르게 오분류 데이터에 대해 비용 매트릭스(Cost Matrix)를 사용하여 분류 오류를 줄이는 방법일 수 있다.
- [0119] 또한, 과대 표집 방법 및 과소 표집 방법이 있는데, 과소 표집 방법(Under Sampling)은 다수 클래스(Majority Class)를 제거하여 클래스 분포의 균형을 맞추는 것이다. 과소 표집 방법의 문제점은 정보의 손실이다. 반면, 과대 표집 방법(Over Sampling)은 소수 클래스(Minority Class)를 다수 클래스에 맞게 복제하여 균형을 맞추는 방법일 수 있다.
- [0120] 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 복수의 데이터셋 특성이 추출된 데이터셋에서 클래스 불균형이 존재하는 경우, 존재하는 클래스 불균형에 따라 다수 클래스(majority class)를 제거하는 과소 표집 방법(under sampling) 및 소수 클래스(minority class)를 다수 클래스(majority class)에 맞게 복제하는 과대 표집 방법(over sampling) 중 어느 하나의 클래스 불균형 해소 방법을 이용하여 클래스 불균형을 해소하는 클래스 불균형 해소 방법을 적용할 수 있다.
- [0121] 단계(S205)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 전처리된 데이터셋 특성을 추출한다.
- [0122] 또한, 단계(S206)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 훈련 데이터셋에 해당하는 전처리된 데이터셋 특성을 분류 알고리즘에 적용하여 테스트를 진행한다.
- [0123] 단계(S207)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 단계(S206)에 기반하여 성능 평가를 진행한다.
- [0124] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 K-겹 교차검증법(K-fold cross validation)을 사용하여 각 폴드(fold) 별 샘플링을 수행하고, 각 폴드에 대한 데이터 특성 및 성능 테스트를 수행하고, 성능 테스트 결과를 도출하여 성능 평가를 진행할 수 있다.
- [0125] 단계(S208)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 메타데이터셋을 생성한다.
- [0126] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 단계(S202)에서 추출된 복수의 데이터셋 특성, 단계(S205)에서 추출된 전처리된 데이터셋 특성, 단계(S206)에서 적용된 분류 알고리즘 및 단계(S207)에서 진행된 성능 평가 결과에 기반하여 메타데이터셋을 생성한다.
- [0127] 다시 말해, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치는 매핑된 샘플링 방법과 매핑된 분류 알고리즘에 적용된 전처리된 복수의 데이터셋 특성을 필터링하고, 필터링된 복수의 데이터셋 특성과 관련된

복수의 데이터셋을 추천 샘플링 방법 및 추천 분류 알고리즘에 투입하여 기계학습하고, 기계학습에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성 결과 및 복수의 데이터셋 특성을 모두 포함하는 메타데이터셋을 생성한다.

- [0129] 도 3은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법을 설명하는 도면이다.
- [0130] 도 3은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법이 오픈 데이터셋으로부터 추출한 데이터셋의 특성을 파악하여 맞춤형 알고리즘을 추천하는 실시예를 설명한다.
- [0131] 도 3을 참고하면, 단계(301)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 오픈 데이터 베이스로부터 오픈 데이터셋을 수집한다.
- [0132] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 오픈 데이터 베이스로부터 오픈 데이터셋을 수집할 수 있다.
- [0133] 단계(302)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 오픈 데이터셋의 복수의 데이터셋 특성을 추출한다.
- [0134] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 수집된 오픈 데이터셋을 복수의 폴드(fold)로 분류하고, 상기 분류된 복수의 폴드(fold) 중 하나를 제외한 나머지 폴드(fold)에 포함된 데이터셋을 복수의 훈련 데이터셋으로 결정하고, 복수의 훈련 데이터셋으로부터 수집된 오픈 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출한다.
- [0135] 단계(303)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 복수의 데이터셋 특성을 전처리한다.
- [0136] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 복수의 데이터셋 특성의 결측치 보간 및 클래스 불균형 해소를 위한 전처리 과정을 수행한다.
- [0137] 단계(304)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 복수의 데이터셋 특성에 따른 샘플링 방법을 매핑한다.
- [0138] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 전처리된 복수의 데이터셋 특성을 복수의 샘플링 방법에 적용하고, 적용된 복수의 샘플링 방법 각각에서의 샘플링 방법 정확도를 산출하고, 산출된 샘플링 방법 정확도에 따라 전처리된 복수의 데이터셋 특성과 샘플링 방법을 매핑할 수 있다.
- [0139] 단계(305)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 복수의 데이터셋 특성에 따른 분류 알고리즘을 매핑한다.
- [0140] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 전처리된 복수의 데이터셋 특성을 복수의 분류 알고리즘에 적용하고, 적용된 복수의 분류 알고리즘 각각에서의 분류 알고리즘의 특성과 하이퍼파라미터에 기반하여 분류 알고리즘 정확도를 산출하고, 산출된 분류 알고리즘 정확도에 따라 전처리된 복수의 데이터셋 특성과 분류 알고리즘을 매핑할 수 있다.
- [0141] 단계(306)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 매핑된 샘플링 방법 및 분류 알고리즘에 기반하여 선정 룰 베이스를 생성한다.
- [0142] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 매핑된 샘플링 방법과 매핑된 분류 알고리즘에 적용된 전처리된 복수의 데이터셋 특성을 필터링하고, 필터링된 복수의 데이터셋 특성과 관련된 복수의 데이터셋을 추천 샘플링 방법 및 추천 분류 알고리즘에 투입하여 기계학습하고, 기계학습에 기반하여 추천 샘플링 방법 및 추천 분류 알고리즘을 선정하기 위한 선정 룰 베이스를 생성할 수 있다.
- [0143] 단계(307)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 선정 룰 베이스를 메타 데이터셋으로 저장한다.
- [0144] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 단계(306)에서 생성된 선정 룰 베이스와

함께 복수의 데이터셋 특성을 메타데이터셋으로 저장할 수 있다.

- [0145] 단계(308)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자로부터 입력된 사용자 데이터셋에 대하여 메타데이터셋을 이용하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0146] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자에게 의해 입력된 사용자 데이터셋에 따른 복수의 데이터셋 특성과 메타데이터셋에 포함된 복수의 데이터셋 특성을 비교하여 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0148] 도 4는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법을 설명하는 도면이다.
- [0149] 도 4는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법이 입력된 사용자 데이터셋에 따른 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천하는 실시예를 설명한다.
- [0150] 도 4를 참고하면, 단계(401)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자 데이터셋을 입력 받는다.
- [0151] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자로부터 사용자 데이터셋을 입력 받는데, 사용자 데이터셋은 사용자가 인공지능 알고리즘에 활용하기 위한 데이터셋으로 볼 수 있다.
- [0152] 단계(402)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자 데이터셋의 복수의 데이터셋 특성을 추출한다.
- [0153] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 입력된 사용자 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 추출할 수 있다.
- [0154] 단계(403)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 복수의 데이터셋 특성을 전처리한다.
- [0155] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자 데이터셋으로부터 추출된 복수의 데이터셋 특성에 대하여 결측치 보간 및 클래스 불균형 해소를 위한 전처리 과정을 수행한다.
- [0156] 단계(404)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 메타데이터셋 기반 복수의 데이터셋 특성에 따른 추천 샘플링 방법을 결정한다.
- [0157] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 메타데이터셋에 포함된 사전 학습된 추천 샘플링 방법들에 사용자 데이터셋의 복수의 데이터셋 특성을 적용하고, 적용 결과에 따라 성능이 가장 우수한 샘플링 방법을 추천 샘플링 방법으로 결정한다.
- [0158] 단계(405)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 메타데이터셋 기반 복수의 데이터셋 특성에 따른 추천 분류 알고리즘을 결정한다.
- [0159] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 메타데이터셋에 포함된 사전 학습된 추천 분류 알고리즘들에 사용자 데이터셋의 복수의 데이터셋 특성을 적용하고, 적용 결과에 따라 성능이 가장 우수한 분류 알고리즘을 추천 분류 알고리즘으로 결정한다. 여기서, 추천 분류 알고리즘의 결정에 따라 분류 알고리즘의 종류 및 파라미터 값도 결정될 수 있다.
- [0160] 단계(406)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 추천 샘플링 방법 및 추천 분류 알고리즘을 학습하고, 맞춤형 샘플링 방법 및 맞춤형 분류 알고리즘 중 적어도 하나를 추천할 수 있다.
- [0161] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 사용자 데이터셋의 복수의 데이터셋 특성과 메타데이터셋 내에 포함된 복수의 데이터셋 특성을 동시에 고려하여 결정된 추천 샘플링 방법 및 추천 분류 알고리즘을 학습하고, 이에 따라 사용자 데이터셋에 맞춤형으로 분류 알고리즘 및 데이터 샘플링 방법을 자

동적으로 결정하여 추천할 수 있다.

- [0163] 도 5는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법을 설명하는 도면이다.
- [0164] 도 5는 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법이 복수의 데이터셋 특성을 수집함에 따라 메타데이터셋을 생성하는 실시예를 설명한다.
- [0165] 도 5를 참고하면, 단계(501)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 표집 방법을 훈련 데이터셋에 적용하여 복수의 훈련 데이터셋을 생성한다.
- [0166] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 데이터 샘플링 방법 중 하나인 분류 알고리즘에 적합하도록 훈련 데이터를 샘플링하는 방법인 표집 방법을 이용하여 복수의 훈련 데이터셋을 생성한다.
- [0167] 단계(502)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 데이터셋의 특성을 추출한다.
- [0168] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 데이터셋에서 변수의 개수, 인스턴스의 개수, 클래스의 개수, 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성 및 이웃성을 포함하는 복수의 데이터셋 특성을 각 훈련 데이터셋으로부터 추출할 수 있다.
- [0169] 단계(503)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 결측치 존재 여부를 판단한다.
- [0170] 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 결측치가 존재할 경우 단계(504)를 진행하고, 결측치가 존재하지 않을 경우, 단계(507)로 진행한다.
- [0171] 단계(504)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 결측치가 존재하는 변수가 수치형인지 판단한다.
- [0172] 일례로, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 결측치가 존재하는 변수가 수치형일 경우, 단계(505)로 진행하고, 결측치가 존재하는 변수가 수치형이 아닐 경우, 단계(506)로 진행한다.
- [0173] 단계(505)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 해당 클래스의 평균을 입력하여 결측치를 처리한다.
- [0174] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 수치형 결측치를 보간하기 위해 클래스의 평균을 결측치에 입력하여 결측치를 보간할 수 있다.
- [0175] 단계(506)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 해당 클래스의 최빈값을 입력하여 결측치를 처리한다.
- [0176] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 명목형 결측치를 보간하기 위해 클래스의 최빈값을 결측치에 입력하여 결측치를 보간할 수 있다.
- [0177] 단계(507)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 K-겹 교차 검증법을 이용하여 각 폴드 별 샘플링을 수행하고, 테스트 결과를 저장한다.
- [0178] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 K만큼 분류된 각 폴드 별 훈련 데이터셋 특성을 샘플링하고, 샘플링된 훈련 데이터셋 특성에 따른 테스트 결과를 도출하여 저장한다. 예를 들어, K는 10일 수 있다.
- [0179] 단계(508)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 클래스 불균형 방법 및 분류 알고리즘을 적용한다.
- [0180] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 클래스 불균형을 해소하기 위한 방법을 적용하고, 샘플링된 훈련 데이터셋 특성과 관련이 있는 테스트 데이터셋 특성을 분류 알고리즘에 적용할 수 있다.

- [0181] 단계(509)에서 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 분류 성능을 측정한다.
- [0182] 즉, 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 방법은 분류 알고리즘에 대한 적용 결과에 대한 분류 성능을 측정하는데, 분류 성능 측정에는 F1-Score 및 G-mean을 이용할 수 있다.
- [0184] 도 6은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 수집하는 데이터셋 특성의 구조를 설명하는 도면이다.
- [0185] 도 6을 참고하면, 데이터셋 특성의 구조 데이터셋 특성(600)이 데이터셋 복잡성(610)과 기본 데이터셋 특성(620)으로 구분될 수 있는 구조를 나타낼 수 있다.
- [0186] 예를 들어, 데이터셋 복잡성(610)은 클래스의 치우침 정도, 클래스의 엔트로피, 변수의 겹침정도, 실루엣 점수, 허브스코어, 변수의 엔트로피, 데이터셋의 선형성, 이웃성, HHI(Herfindahl-Hirschman Index) 및 Silhouette 등을 포함할 수 있다.
- [0187] 한편, 기본 데이터셋 특성(620)은 인스턴트 수, 결측 값의 수, 변수의 개수, 클래스의 개수를 포함할 수 있다.
- [0188] HHI(Herfindahl-Hirschman Index)는 산업의 경쟁 상황인 시장상황을 0 내지 1로 나타내는 것일 수 있고, Silhouette는 군집 분석에서 군집 평가를 위해 사용되는 평가 지표일 수 있다.
- [0189] Silhouette는 높을수록 클러스터(Cluster)가 잘 묶일수록 클러스터내의 동질성을 갖고 있다고 볼 수 있고, 동시에 클러스터간의 이질성을 잘 나타낸다고 할 수 있다.
- [0190] 인스턴스 수는 데이터 셋의 크기를 나타내는 것으로, 이는 정보가 많을수록 정보가 더 많다는 것을 의미한다.
- [0191] 그러나 더 큰 데이터일수록 잡음이 발생하고 머신러닝 분야에서 학습 시간이 더 많이 소요되고, 너무 많은 노이즈가 발생할 수 있으며 너무 작으면 정보 부족으로 인해 올바르게 분류하기 어려울 수 있다.
- [0192] 변수의 개수는 너무 적은 수의 변수를 고려한다면, 예측 모델의 복잡성은 떨어지겠지만, 많은 수의 변수를 고려했을 때에 비해 정확도는 떨어질 수 있다.
- [0193] 변수의 개수가 줄어들면 줄어들수록 G-mean이 상승하는 특징이 있다.
- [0194] 클래스 개수(Number of classes)는 클래스가 갖고 있는 차원으로, 바이너리형태의 클래스라면 각 클래스의 기대 확률은 0.5이고, 클래스 수가 4개인 경우는 0.25일 것이다.
- [0195] 또한, 확률적으로 바이너리 형태인 경우와 클래스가 4개인 경우의 기대확률이 다르고 클래스가 많을수록 기대확률이 떨어지기 때문에 클래스의 정확도와 밀접한 관련이 있을 수 있다.
- [0196] 결측 값은 다양한 분야에서 발생하고 있다. 결측 값은 데이터 마이닝, 기계 학습 및 기타 정보 시스템에서 좋지 않은 영향을 미칠 수 있다.
- [0197] 결측 값은 일반적으로 센서 결함, 과학 실험에서의 응답 부족, 측정 결함, 디지털 시스템의 데이터 전송 문제 또는 설문 조사에 대한 응답자의 응답을 꺼리기 때문에 발생할 수 있다.
- [0198] 결측 값이 많으면 클래스를 분류하기 위한 정보가 부족하여 오분류를 발생시킬 수 있다.
- [0199] HHI는 산업의 기업 시장 집중도를 나타내는 지수 중 하나로, HHI의 장점은 전체 산업의 경쟁 상황을 시장상황을 0 내지 1로 나타내어 한번에 직관적으로 알 수 있다는 장점이 있다.
- [0200] 클래스의 균형 상태에 HHI를 적용할 수 있는데 시장의 점유율을 각 클래스의 비율로 볼 때 class의 균형 상태와 시장의 경쟁상태가 유사한 형태를 갖고 있기 때문 적용할 수 있다.
- [0201] 엔트로피(entropy)는 정보의 양과 순도를 나타내고, 주어진 데이터에 의해 발생하는 정보량에 대한 불확실성을 정량적으로 측정할 수 있는 방법으로, 발생확률이 1에 가까울수록 정보의 양은 적고, 발생확률이 적을수록 정보의 양이 많아지게 된다.
- [0202] 즉, 희귀한 정보일수록 정보량을 많이 갖고 있으며 보편적인 데이터일수록 데이터가 일관성을 갖고 있다고 할 수 있다.

- [0203] 따라서 엔트로피는 특이 데이터가 많지 않고, 일관된 순도 높은 데이터인지 알 수 있는 지표가 될 수 있다.
- [0204] 클래스 엔트로피는 엔트로피를 사용하여 클래스의 불균형을 표현한 것으로, 클래스 불균형이 심할수록 높은 수로 표현될 수 있다.
- [0205] Silhouette 계수는 클러스터링 분석 후 클러스터가 올바르게 구성되었는지 검증하기 위해 사용되는 지표이다.
- [0206] 산출된 클러스터내에 값들이 얼마나 밀도 있게 뭉쳐있는지, 클러스터간의 거리는 충분한지를 판단할 수 있다. Silhouette의 값이 1에 가까울수록 클러스터의 개수가 적절하다고 판단될 수 있다.
- [0207] 데이터 선형성(data nonlinearity)은 데이터의 형태가 복잡할수록 분류문제에서 어려움을 겪을 수 있고, 일부 분류 알고리즘은 선형모형을 이용하여 분류 문제를 해결한다.
- [0208] 예를 들어, 본 발명은 데이터의 선형성과 이상치를 동시에 고려하는 방법인 선형 분류기의 비선형(Nonlinearity of linear classifier) 측정 방법을 사용할 수 있다.
- [0209] 비선형 분류기(non-linear classifier)를 통해 분류하고, SVM(support vector machine)의 선형 커널 기능(linear kernel function)을 통해 분류된 결과를 비교하여, 오류(error)를 비교하는 방법이 활용될 수 있다.
- [0210] 허브 스코어(hub score)는 네트워크의 연결성을 이용하여 데이터의 응집력을 측정하는 지표이다.
- [0211] 허브 스코어는 주어진 데이터를 이용하여 네트워크를 구성하고, 노드에 연결된 숫자를 통해 측정될 수 있다.
- [0212] 변수의 겹침정도(feature overlap)은 변수 간의 중복 정도를 비율로 제시한다. 데이터가 겹치지 않을 수록 의사결정 경계가 명확해진다. 따라서 변수의 중첩이 높을 수록 Tomek link, ENN, CNN과 같은 과소 샘플링 방법이 더 유리할 수 있다.
- [0213] 이웃성(neighborhood)는 데이터가 잘못 레이블링이 된 경우 데이터의 위치가 엉뚱한 곳에 배치되기 때문에 선형 분리 문제에서 분류 성능을 떨어뜨리는 원인과 관련될 수 있다.
- [0214] 각 인스턴스에 대해 클래스 내에서 가장 가까운 이웃과의 거리와 다른 클래스의 가장 가까운 이웃까지의 거리를 계산하고, 클래스 간 거리의 합에 대한 클래스 내 거리의 합계의 비율을 데이터셋의 개수로 이용한다.
- [0215] 차원(Dimensionality)은 PCA(Principal Component Analysis)를 통해 감소된 변수의 수와 원본 데이터 변수의 수 간의 차이를 비율로 나타낼 수 있다.
- [0217] 도 7은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 생성하는 메타데이터셋을 예시하는 도면이다.
- [0218] 도 7을 참고하면, 메타데이터셋의 예시(700)는 식별번호, 정확도, G-mean 점수, F-score 점수, 알고리즘 데이터, 방법, 비율, 복수의 데이터셋 특성 등 다양한 정보를 포함할 수 있다.
- [0220] 도 8은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 분류 알고리즘과 관련하여 성능 평가한 결과를 예시하는 도면이다.
- [0221] 도 8을 참고하면, 그래프(800)는 가로축에서 분류 알고리즘의 종류를 나타내고, 세로축에서 주파수를 나타낸다.
- [0222] 분류 알고리즘의 종류는 knn(Nearest Neighbor), LR(Logistic regression), NB(naïve Bayes), RF(random forest) 및 SVM(support vector machine)을 포함한다.
- [0223] 그래프(800)는 클래스를 분류하는 알고리즘인 랜덤 포레스트(Random Forest, RF)가 다른 알고리즘에 비해 우수한 성능을 나타내고 있음을 보여준다.
- [0224] k-NN방법은 나이브 베이즈(naïve Bayes) 분류기나, SVM, 의사결정 나무 모형과 다르게 학습데이터를 이용하지 않다가 실증데이터가 주어져야 움직이는 lazy학습법이다.
- [0225] 로지스틱 회귀(Logistic regression) 모델은 로지스틱 누적 분포 함수(Cumulative Distribution Function, CDF)의 기능적 형태를 가지고 있다.

- [0226] 나이브 베이즈 분류기는 지도학습 중 확률적으로 접근하는 방법으로, 베이즈 정리를 이용한다. 나이브 베이즈는 모형이 비교적 단순하며, 계산과정이 복잡하지 않지만, 우수한 성능을 나타내는 것으로 알려져 있다.
- [0227] 랜덤 포레스트(random forest, RF)는 의사결정 나무모형을 기반으로 한 분류 알고리즘의 하나로, 여러 개의 의사결정 나무모형을 생성하여 다수결 방식의 앙상블 기법이다.
- [0228] SVM(Support Vector Machine)은 n 차원의 데이터에서 $n-1$ 차원의 초평면을 이용하여 데이터를 분류할 때 사용하는 방법이다.
- [0230] 도 9은 본 발명의 일실시예에 따른 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치가 샘플링 방법과 관련하여 성능 평가한 결과를 제시하는 도면이다.
- [0231] 도 9를 참고하면, 그래프(900)는 가로축에서 샘플링 방법의 종류를 나타내고, 세로축에서 주파수를 나타낸다.
- [0232] 샘플링 방법의 종류는 Adasyn(Adaptive Synthetic Sampling Approach for Imbalanced Learning), ENN(Edited Nearest Neighbors), NCL(Neighbourhood Cleaning Rule), ROS(Random Over Sampling), RUS(Random Under Sampling), SMOTE(Synthetic Minority Oversampling TEchnique) 및 Tomek을 포함한다.
- [0233] 그래프(900)는 Adasyn과 SMOTE가 우수한 성능을 나타내는 것을 확인시켜준다.
- [0234] 과대 표집 방법에는 ROS, SMOTE, ADASYN을 포함하고, 과소 표집법은 RUS, ENN, Tomek link 방법, CNN 및 NCL을 포함한다.
- [0235] ROS는 임의 과대 추출 방법은 소수의 클래스를 다수 클래스의 데이터 크기와 같아질 때까지 데이터를 무작위로 선택하여 반복 복원 추출하는 방법이다.
- [0236] SMOTE는 소수 클래스의 임의의 데이터를 선정하고 최 근접 이웃 k 개(k -Nearest Neighbor, NN)사이에 새로운 인공 데이터를 생성하는 방법이다.
- [0237] ADASYN(Adaptive Synthetic Sampling)은 SMOTE를 기반으로 소수 클래스의 밀도 분포를 고려하여 데이터를 생성하는 방법이다.
- [0238] RUS는 다수 클래스(Majority Class)를 임의로 삭제시켜, 소수의 클래스(Minority Class)의 비율에 맞추는 표집 방법으로, RUS는 ROS와 같이 사용하기 편리한 장점이 있고, 대규모 데이터는 데이터 수를 줄여 비용을 줄일 수 있고, 임의로 데이터를 줄이기 때문에 중요한 정보를 손실할 가능성이 높다.
- [0239] CNN방법은 Hart(1968)에 제안된 방법으로 훈련데이터 중 임의 데이터를 선정하여 집합 X 에 저장하고, 또 다른 하나의 데이터를 선정하여 집합 X 에 포함시킨다.
- [0240] 집합 X 의 데이터를 NN(Nearest Neighbors)규칙을 사용하여 분류하여, 오분류 되었을 때 집합 X 에 저장하고, 다시 임의 데이터를 선정하여 집합 X 에 포함시켜 NN규칙을 사용한다.
- [0241] CNN방법은 오분류 된 X 집합을 제외한, 훈련데이터가 모두 잘 분류가 되었을 때까지 반복하고, 서로 다른 클래스의 경계(boundary)가 명확한 데이터를 남기는 방법이다.
- [0242] ENN은 CNN의 변형으로, CNN과 달리 집합 X 에 포함된 값이 오분류인 경우 X 에서 제외시킬 수 있다.
- [0243] Tomek link는 CNN샘플링 방법을 기본으로 하여 의사 결정 경계 근처에 있는 내부 데이터를 제거하는 방법이다.
- [0244] NCL은 CNN(Condensed Nearest Neighbour)과 ENN(Edited Nearest Neighbours)을 혼합한 방법이다.
- [0246] 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 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0247] 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. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0248] 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 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embodiment)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 매체에 저장될 수 있다.

[0249] 이상과 같이 실시예들이 비록 한정된 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.

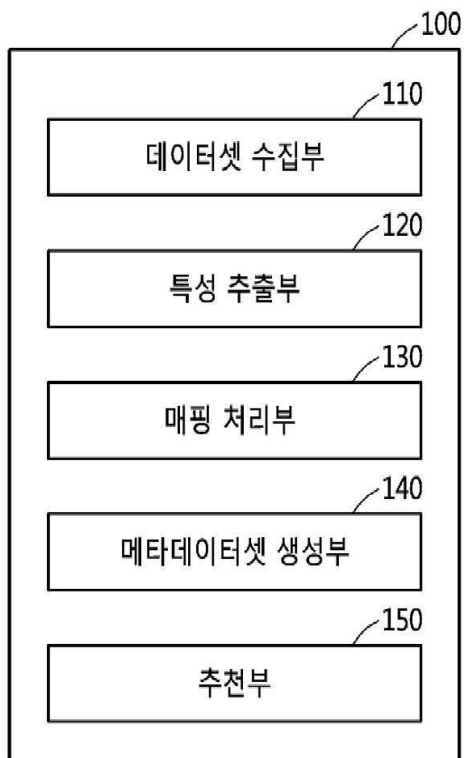
[0250] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

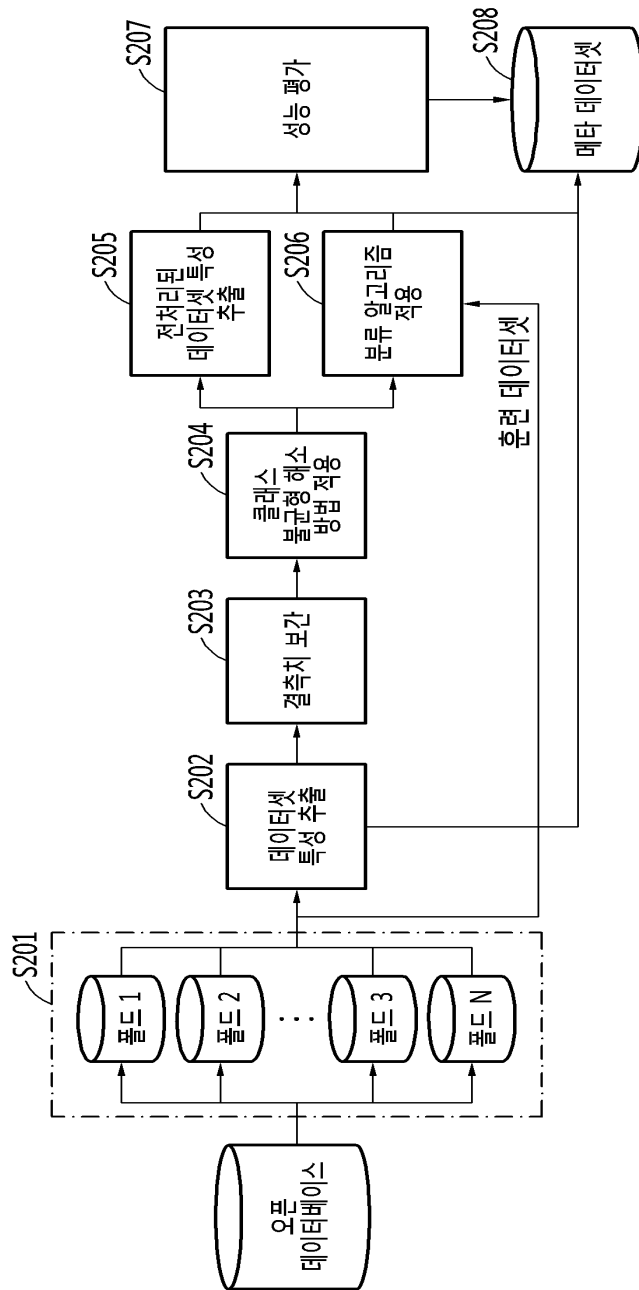
[0251] 100: 메타데이터셋을 이용한 샘플링 방법 및 분류 알고리즘 추천 장치
110: 데이터셋 수집부 120: 특성 추출부
130: 매핑처리부 140: 메타데이터셋 생성부
150: 추천부

도면

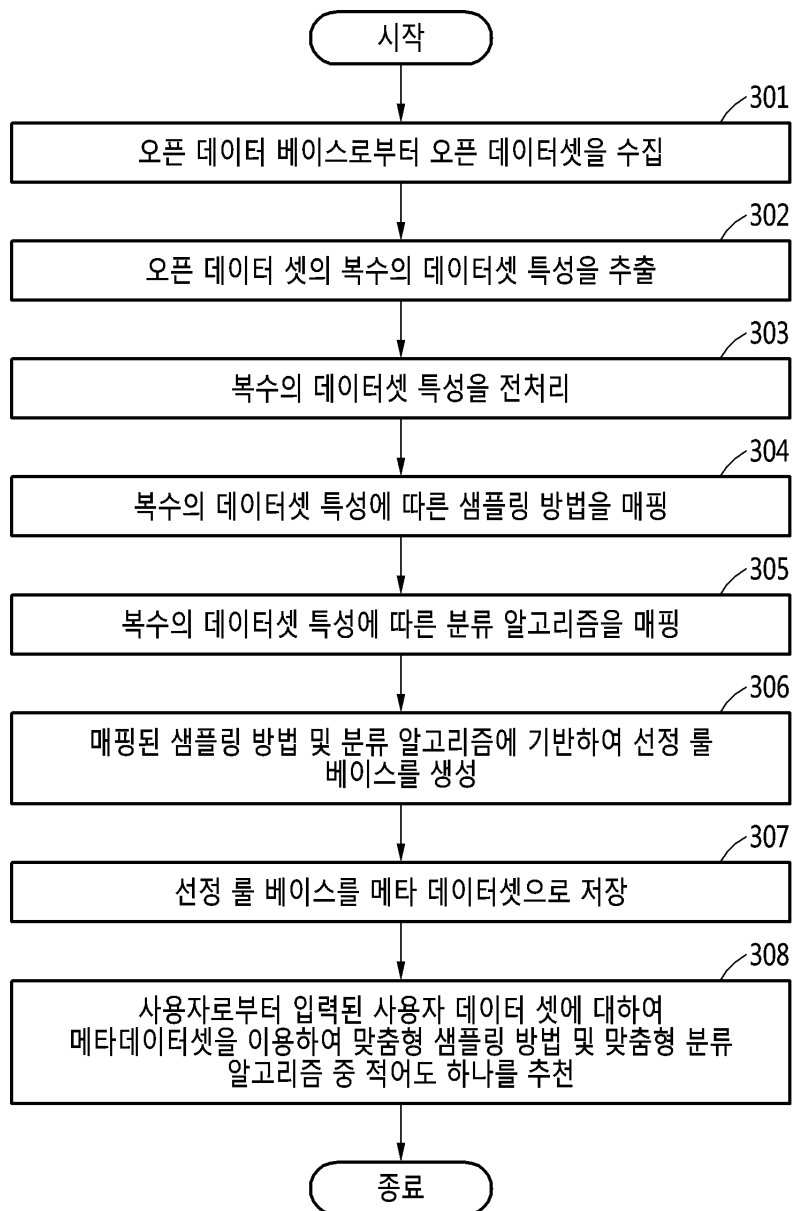
도면1



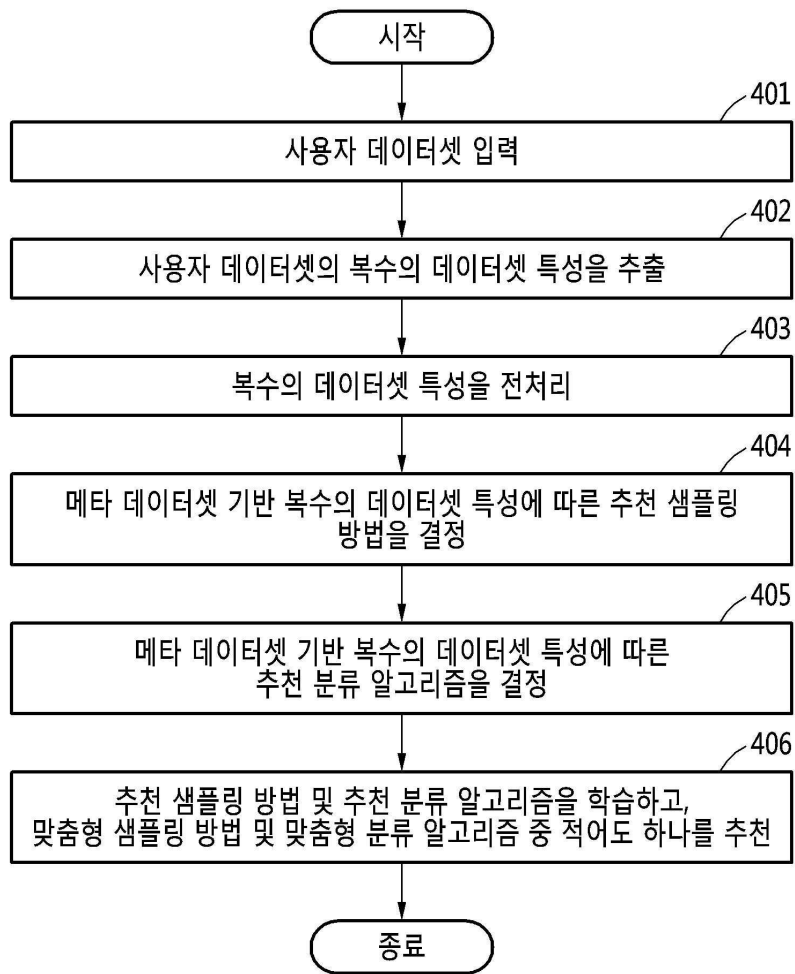
도면2



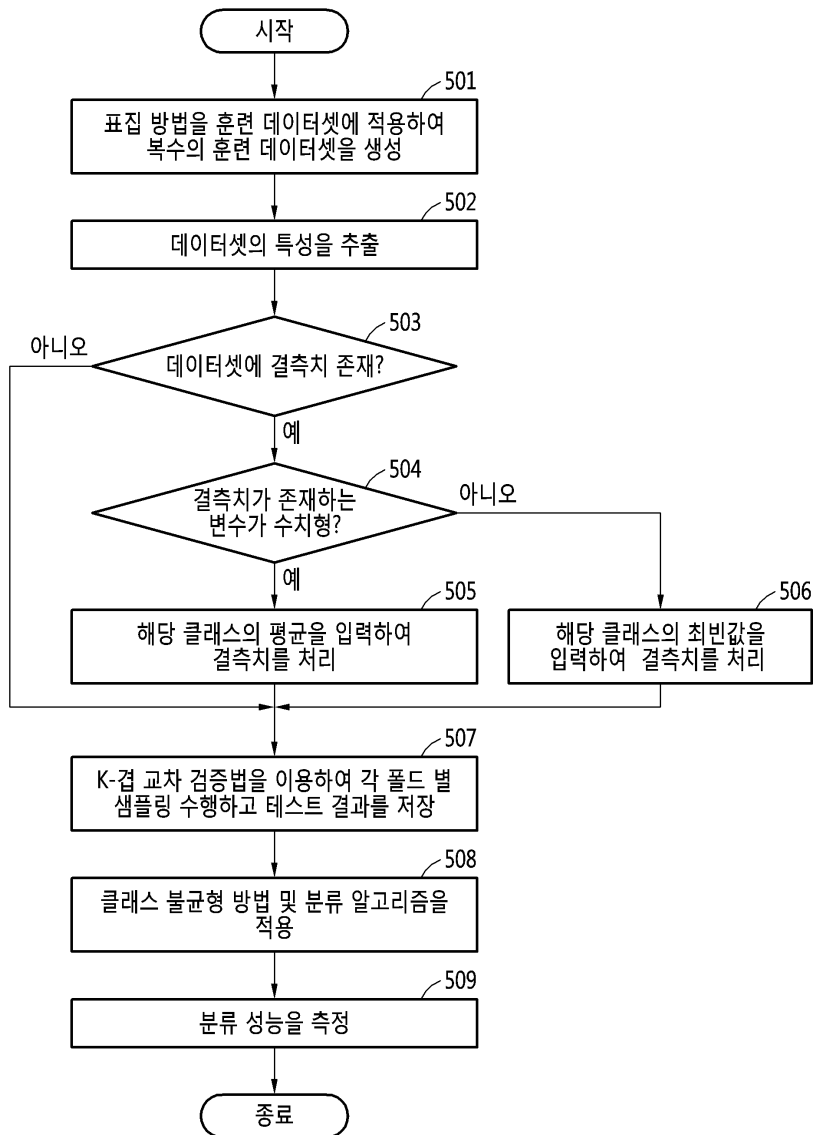
도면3



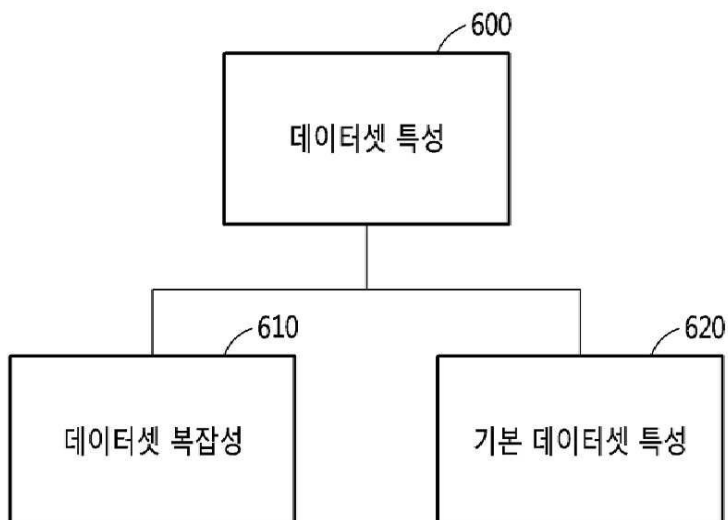
도면4



도면5



도면6

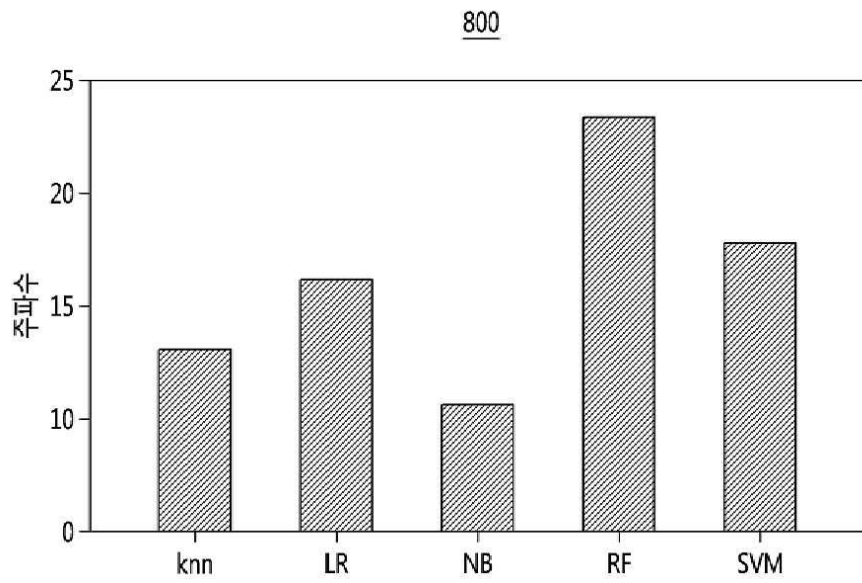


도면7

700

id	Accuracy	G-mean	F-score	Elapsed	Algorithm data	method	HHI	ratio	RQ	centropy	ventropy	num_class	instance	num_feat	num_num	factnum	Dimension	Linear	Net	QL	NN	sl	mean	mean	mean
cmc_fold_1	0.992747	0.995943	1	2.95 RF	cmc	ROC	0.1	100	0.836195	3.321928	12.89531	10	10286	17	16	0	0	0.5625	0.000875	0.937642	0.496484	0.231407	0.403925	0.368877	0.368877
cmc_fold_1	0.993654	0.996451	1	1.41 knn	cmc	ROC	0.1	100	0.836195	3.321928	12.89531	10	10286	17	16	0	0	0.5625	0.000875	0.937642	0.496484	0.231407	0.403925	0.368877	0.368877
cmc_fold_1	0.842248	0.865329	0.895652	0.02 NB	cmc	ROC	0.1	100	0.836195	3.321928	12.89531	10	10286	17	16	0	0	0.5625	0.000875	0.937642	0.496484	0.231407	0.403925	0.368877	0.368877
cmc_fold_1	0.87942	0.897806	0.883333	5.93 LR	cmc	ROC	0.1	100	0.836195	3.321928	12.89531	10	10286	17	16	0	0	0.5625	0.000875	0.937642	0.496484	0.231407	0.403925	0.368877	0.368877
cmc_fold_1	0.978241	0.97246	0.965217	0.57 SVM	cmc	ROC	0.1	100	0.836195	3.321928	12.89531	10	10286	17	16	0	0	0.5625	0.000875	0.937642	0.496484	0.231407	0.403925	0.368877	0.368877
cmc_fold_1	0.992747	0.991772	0.995261	2.99 RF	cmc	ROC	0.100001	90	0.831132	3.321918	12.8896	10	10246	17	16	0	0	0.5625	0.000877	0.937405	0.495207	0.233257	0.442855	0.328017	0.328017
cmc_fold_1	0.993654	0.992773	0.985915	1.41 knn	cmc	ROC	0.100001	90	0.831132	3.321918	12.8896	10	10246	17	16	0	0	0.5625	0.000877	0.937405	0.495207	0.233257	0.442855	0.328017	0.328017
cmc_fold_1	0.843155	0.870398	0.876712	0.01 NB	cmc	ROC	0.100001	90	0.831132	3.321918	12.8896	10	10246	17	16	0	0	0.5625	0.000877	0.937405	0.495207	0.233257	0.442855	0.328017	0.328017
cmc_fold_1	0.887579	0.920024	0.93578	5.89 LR	cmc	ROC	0.100001	90	0.831132	3.321918	12.8896	10	10246	17	16	0	0	0.5625	0.000877	0.937405	0.495207	0.233257	0.442855	0.328017	0.328017
cmc_fold_1	0.978241	0.983725	0.985915	0.54 SVM	cmc	ROC	0.100001	90	0.831132	3.321918	12.8896	10	10246	17	16	0	0	0.5625	0.000877	0.937405	0.495207	0.233257	0.442855	0.328017	0.328017
cmc_fold_1	0.99454	0.992763	0.990566	2.84 RF	cmc	Tonek	0.100119	Tonek	0.865163	3.320777	12.83847	10	9891	17	16	0	0	0.5625	0.000893	0.935251	0.49433	0.251198	0.447164	0.324777	0.324777
cmc_fold_1	0.99545	0.993265	0.985915	1.39 knn	cmc	Tonek	0.100119	Tonek	0.865163	3.320777	12.83847	10	9891	17	16	0	0	0.5625	0.000893	0.935251	0.49433	0.251198	0.447164	0.324777	0.324777
cmc_fold_1	0.857143	0.894393	0.921659	0.01 NB	cmc	Tonek	0.100119	Tonek	0.865163	3.320777	12.83847	10	9891	17	16	0	0	0.5625	0.000893	0.935251	0.49433	0.251198	0.447164	0.324777	0.324777
cmc_fold_1	0.903549	0.933224	0.971698	5.65 LR	cmc	Tonek	0.100119	Tonek	0.865163	3.320777	12.83847	10	9891	17	16	0	0	0.5625	0.000893	0.935251	0.49433	0.251198	0.447164	0.324777	0.324777
cmc_fold_1	0.987261	0.988736	0.995261	0.63 SVM	cmc	Tonek	0.100119	Tonek	0.865163	3.320777	12.83847	10	9891	17	16	0	0	0.5625	0.000893	0.935251	0.49433	0.251198	0.447164	0.324777	0.324777
new_thyro	0.952381	0.971825	0.857143	0.01 RF	new_thyro	SMOTE	0.418881	20	0.924935	1.41267	7.262317	3	236	6	5	0	0	0.6	0.007192	0.895799	0.490341	0.216344	0.652637	0.237821	0.237821
new_thyro	1	1	1	0 knn	new_thyro	SMOTE	0.418881	20	0.924935	1.41267	7.262317	3	236	6	5	0	0	0.6	0.007192	0.895799	0.490341	0.216344	0.652637	0.237821	0.237821
new_thyro	0.952381	0.971825	0.857143	0 NB	new_thyro	SMOTE	0.418881	20	0.924935	1.41267	7.262317	3	236	6	5	0	0	0.6	0.007192	0.895799	0.490341	0.216344	0.652637	0.237821	0.237821
new_thyro	0.952381	0.971825	0.857143	0 LR	new_thyro	SMOTE	0.418881	20	0.924935	1.41267	7.262317	3	236	6	5	0	0	0.6	0.007192	0.895799	0.490341	0.216344	0.652637	0.237821	0.237821

도면8



도면9

