(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0054701
(43) 공개일자 2023년04월25일

(51) 국제특허분류(Int. Cl.)
G06N 20/00 (2019.01) *G06F 16/9032* (2019.01)
G06F 16/9035 (2019.01) *G06F 18/21* (2023.01)
G06F 18/2113 (2023.01) *G06F 18/214* (2023.01)
G06F 18/22 (2023.01) *G06N 3/096* (2023.01)
G06N 3/0985 (2023.01)
(52) CPC특허분류
G06N 20/00 (2021.08)
G06F 16/90328 (2019.01)
(21) 출원번호 10-2023-7009608
(22) 출원일자(국제) 2021년08월24일
심사청구일자 2023년03월20일
(85) 번역문제출일자 2023년03월20일
(86) 국제출원번호 PCT/US2021/047334
(87) 국제공개번호 WO 2022/046759
국제공개일자 2022년03월03일
(30) 우선권주장
17/002,717 2020년08월25일 미국(US)

(71) 출원인
알테릭스 인코포레이티드
미국 92618 캘리포니아 어바인 라구나 캐니언 로드 17200
(72) 발명자
블랜차드 딜런
미국 92618 캘리포니아주 어바인 라구나 캐니언 로드 17200 알테릭스 인코포레이티드
헤이늘 타일러
미국 92618 캘리포니아주 어바인 라구나 캐니언 로드 17200 알테릭스 인코포레이티드
호호무트 롤랜드 맨프레드
미국 92618 캘리포니아주 어바인 라구나 캐니언 로드 17200 알테릭스 인코포레이티드
(74) 대리인
특허법인코리아나

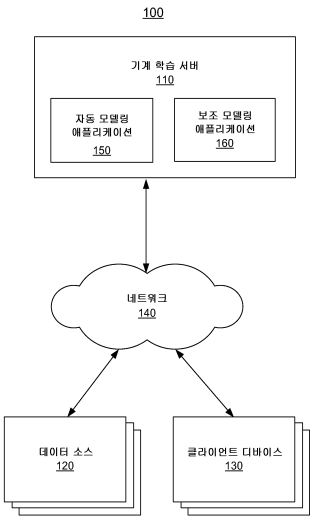
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **하이브리드 기계 학습**

(57) 요 약

모델은 하이브리드 기계 학습 프로세스를 통해 훈련된다. 하이브리드 기계 학습 프로세스에서, 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 예측을 수행하기 위한 모델을 생성한다. 자동 기계 학습 프로세스는 파이프라인을 사용하여 모델을 훈련하고 파이프라인 스텝에서 판단을 수행한다. 자동 기계 학습 프로세스를 통해 모델을 훈련시킨 후, 파이프라인의 표현이 생성되어 사용자 인터페이스에서 사용자에게 제시된다. 사용자 인터페이스는 사용자가 자동 기계 학습 프로세스에서 수행된 적어도 일부 판단을 수정할 수 있게 한다.
사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을 수신하고 훈련된 모델을 리파이닝하는데 사용된다. 새로운 데이터에 기초하여 예측을 수행하기 위해 리파이닝된 모델이 배치된다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/9035 (2019.01)
G06F 18/2113 (2023.01)
G06F 18/214 (2023.01)
G06F 18/2178 (2023.01)
G06F 18/22 (2023.01)
G06N 3/096 (2023.01)
G06N 3/0985 (2023.01)

명 세 서

청구범위

청구항 1

컴퓨터 구현된 방법으로서,

데이터세트를 수신하는 단계;

수신된 상기 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 훈련시키는 단계로서, 상기 자동 기계 학습 프로세스는 파이프라인에 기초하여 일련의 판단을 수행하는 단계를 포함하는, 상기 모델을 훈련시키는 단계;

상기 모델이 훈련된 후에, 상기 파이프라인의 표현을 생성하는 단계;

사용자 인터페이스에 디스플레이하기 위해 상기 파이프라인의 표현을 제공하는 단계로서, 상기 사용자 인터페이스는 사용자가 상기 자동 기계 학습 프로세스에서 수행된 판단 중 적어도 일부를 수정할 수 있게 하는, 상기 파이프라인의 표현을 제공하는 단계;

상기 사용자 인터페이스를 통해 상기 사용자로부터 하나 이상의 수정을 수신하는 단계; 및

상기 사용자로부터의 상기 하나 이상의 수정에 기초하여 상기 모델을 리파이닝하는 단계로서, 상기 리파이닝된 모델은 새로운 데이터에 기초하여 상기 예측을 수행하는데 사용되는, 상기 모델을 리파이닝하는 단계를 포함하는, 컴퓨터 구현된 방법.

청구항 2

제 1 항에 있어서,

각각의 판단은 상기 자동 기계 학습 프로세스 동안 상기 파이프라인의 일련의 스텝 중의 스텝에서 수행되고, 상기 파이프라인의 표현을 생성하는 단계는

상기 일련의 스텝 중 하나 이상의 스텝 각각에 대한 표현을 생성하는 단계를 포함하고, 상기 스텝의 표현은 상기 스텝에 대한 복수의 옵션을 포함하고, 상기 복수의 옵션은 상기 자동 기계 학습 프로세스 동안 상기 스텝에서 수행된 판단을 포함하는, 컴퓨터 구현된 방법.

청구항 3

제 2 항에 있어서,

상기 스텝의 표현은 상기 복수의 옵션 각각에 대한 순위 점수를 더 포함하고, 상기 옵션의 순위 점수는 상기 옵션에 대한 추천 수준을 나타내는, 컴퓨터 구현된 방법.

청구항 4

제 1 항에 있어서,

상기 파이프라인의 표현을 생성하는 단계는

상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에 사용되는 특징, 상기 특징에 대한 복수의 선택적 데이터 유형, 및 각각의 선택적 데이터 유형에 대한 순위 점수를 포함하는 데이터 유형 목록을 생성하는 단계를 포함하고, 상기 복수의 선택적 데이터 유형은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 상기 특징에 대해 선택된 제1 데이터 유형을 포함하고, 상기 선택적 데이터 유형의 상기 순위 점수는 데이터 유형이 특징의 실제 데이터 유형일 확률을 나타내는, 컴퓨터 구현된 방법.

청구항 5

제 4 항에 있어서,

상기 사용자 인터페이스를 통해 상기 사용자로부터 하나 이상의 수정을 수신하는 단계는

상기 사용자 인터페이스를 통해 상기 사용자로부터 상기 복수의 선택적 데이터 유형으로부터의 상기 특징에 대한 제2 데이터 유형의 선택을 수신하는 단계로서, 상기 제2 데이터 유형은 상기 제1 데이터 유형과는 상이한 데이터 유형인, 상기 선택을 수신하는 단계; 및

상기 제2 데이터 유형에 기초하여 특징의 값을 인코딩하는 단계를 포함하는, 컴퓨터 구현된 방법.

청구항 6

제 1 항에 있어서,

상기 파이프라인의 표현을 생성하는 단계는

특징 목록을 생성하는 단계를 포함하고, 상기 특징 목록은 복수의 특징 및 상기 특징이 예측에 얼마나 중요한지를 나타내는 각각의 특징에 대한 설명을 포함하고, 상기 복수의 특징은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 특징을 포함하는, 컴퓨터 구현된 방법.

청구항 7

제 1 항에 있어서,

상기 파이프라인의 표현을 생성하는 단계는

알고리즘 목록을 생성하는 단계를 포함하고, 상기 알고리즘 목록은 복수의 알고리즘 및 상기 모델의 훈련에 대한 상기 알고리즘의 선택 이유 또는 비선택 이유를 나타내는 각각의 알고리즘에 대한 설명을 포함하고, 상기 복수의 알고리즘은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 알고리즘을 포함하는, 컴퓨터 구현된 방법.

청구항 8

데이터 분석 시스템에서 데이터 블록을 처리하기 위한 실행가능한 컴퓨터 프로그램 명령을 저장하는 비일시적 컴퓨터 판독가능 메모리로서,

상기 명령은

데이터세트를 수신하는 것;

수신된 상기 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 훈련시키는 것으로서, 상기 자동 기계 학습 프로세스는 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함하는, 상기 모델을 훈련시키는 것;

상기 모델이 훈련된 후에, 상기 파이프라인의 표현을 생성하는 것;

사용자 인터페이스에 디스플레이하기 위해 상기 파이프라인의 표현을 제공하는 것으로서, 상기 사용자 인터페이스는 사용자가 상기 자동 기계 학습 프로세스에서 수행된 판단 중 적어도 일부를 수정할 수 있게 하는, 상기 파이프라인의 표현을 제공하는것;

상기 사용자 인터페이스를 통해 상기 사용자로부터 하나 이상의 수정을 수신하는 것; 및

상기 사용자로부터의 상기 하나 이상의 수정에 기초하여 상기 모델을 리파이닝하는 것으로서, 상기 리파이닝된 모델은 새로운 데이터에 기초하여 상기 예측을 수행하는데 사용되는, 상기 모델을 리파이닝하는 것

을 포함하는 동작을 수행하도록 실행가능한, 비일시적 컴퓨터 판독가능 메모리.

청구항 9

제 8 항에 있어서,

상기 자동 기계 학습 프로세스 동안 상기 파이프라인의 일련의 스텝 중의 스텝에서 각각의 판단이 수행되고, 상기 파이프라인의 표현을 생성하는 것은

상기 일련의 스텝 중 하나 이상의 스텝 각각에 대한 표현을 생성하는 것을 포함하고, 상기 스텝의 표현은 상기

스텝에 대한 복수의 옵션을 포함하고, 상기 복수의 옵션은 상기 자동 기계 학습 프로세스 동안 상기 스텝에서 수행된 판단을 포함하는, 비일시적 컴퓨터 판독가능 메모리.

청구항 10

제 9 항에 있어서,

상기 스텝의 표현은 상기 복수의 옵션 각각에 대한 순위 점수를 더 포함하고, 상기 옵션의 순위 점수는 상기 옵션에 대한 추천 수준을 나타내는, 비일시적 컴퓨터 판독가능 메모리.

청구항 11

제 8 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에 사용되는 특징, 상기 특징에 대한 복수의 선택적 데이터 유형, 및 각각의 선택적 데이터 유형에 대한 순위 점수를 포함하는 데이터 유형 목록을 생성하는 것을 포함하고, 상기 복수의 선택적 데이터 유형은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 상기 특징에 대해 선택된 제1 데이터 유형을 포함하고, 상기 선택적 데이터 유형의 상기 순위 점수는 데이터 유형이 특징의 실제 데이터 유형일 확률을 나타내는, 비일시적 컴퓨터 판독가능 메모리.

청구항 12

제 11 항에 있어서,

상기 사용자 인터페이스를 통해 상기 사용자로부터 하나 이상의 수정을 수신하는 것은

상기 사용자 인터페이스를 통해 상기 사용자로부터 상기 복수의 선택적 데이터 유형으로부터의 상기 특징에 대한 제2 데이터 유형의 선택을 수신하는 것으로서, 상기 제2 데이터 유형은 상기 제1 데이터 유형과는 상이한 데이터 유형인, 상기 제 2 데이터 유형의 선택을 수신하는 것; 및

상기 제2 데이터 유형에 기초하여 상기 특징의 값을 인코딩하는 것을 포함하는, 비일시적 컴퓨터 판독가능 메모리.

청구항 13

제 8 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

특징 목록을 생성하는 것을 포함하고, 상기 특징 목록은 복수의 특징 및 사용자가 예측에 얼마나 중요한지를 나타내는 각각의 특징에 대한 설명을 포함하고, 상기 복수의 특징은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 특징을 포함하는, 비일시적 컴퓨터 판독가능 메모리.

청구항 14

제 8 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

알고리즘 목록을 생성하는 것을 포함하고, 상기 알고리즘 목록은 복수의 알고리즘 및 상기 모델의 훈련에 대한 상기 알고리즘의 선택 이유 또는 비선택 이유를 나타내는 각각의 알고리즘에 대한 설명을 포함하고, 상기 복수의 알고리즘은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 알고리즘을 포함하는, 비일시적 컴퓨터 판독가능 메모리.

청구항 15

시스템으로서,

컴퓨터 프로그램 명령을 실행하기 위한 컴퓨터 프로세서; 및

동작을 수행하기 위해 상기 컴퓨터 프로세서에 의해 실행가능한 컴퓨터 프로그램 명령을 저장하는 비일시적 컴

퓨터 판독가능 메모리를 포함하고, 상기 동작은

데이터세트를 수신하는 것;

수신된 상기 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 훈련시키는 것으로서, 상기 자동 기계 학습 프로세스는 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함하는, 상기 모델을 훈련시키는 것;

상기 모델이 훈련된 후에, 상기 파이프라인의 표현을 생성하는 것;

사용자 인터페이스에 디스플레이하기 위해 상기 파이프라인의 표현을 제공하는 것으로서, 상기 사용자 인터페이스는 사용자가 상기 자동 기계 학습 프로세스에서 수행된 판단 중 적어도 일부를 수정할 수 있게 하는, 상기 파이프라인의 표현을 제공하는 것;

상기 사용자 인터페이스를 통해 상기 사용자로부터 하나 이상의 수정을 수신하는 것; 및

상기 사용자로부터의 상기 하나 이상의 수정에 기초하여 상기 모델을 리파이닝하는 것으로서, 상기 리파이닝된 모델은 새로운 데이터에 기초하여 상기 예측을 수행하는데 사용되는, 상기 모델을 리파이닝하는 것을 포함하는, 시스템.

청구항 16

제 15 항에 있어서,

각각의 판단은 상기 자동 기계 학습 프로세스 동안 상기 파이프라인의 일련의 스텝 중의 스텝에서 수행되고, 상기 파이프라인의 표현을 생성하는 것은

상기 일련의 스텝 중 하나 이상의 스텝 각각에 대한 표현을 생성하는 것을 포함하고, 상기 스텝의 표현은 상기 스텝에 대한 복수의 옵션을 포함하고, 상기 복수의 옵션은 상기 자동 기계 학습 프로세스 동안 상기 스텝에서 수행된 판단을 포함하는, 시스템.

청구항 17

제 16 항에 있어서,

상기 스텝의 표현은 복수의 옵션 각각에 대한 순위 점수를 더 포함하고, 상기 옵션의 순위 점수는 상기 옵션에 대한 추천 수준을 나타내는, 시스템.

청구항 18

제 15 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에 사용되는 특징, 상기 특징에 대한 복수의 선택적 데이터 유형, 및 각각의 선택적 데이터 유형에 대한 순위 점수를 포함하는 데이터 유형 목록을 생성하는 것을 포함하고, 상기 복수의 선택적 데이터 유형은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 상기 특징에 대해 선택된 제1 데이터 유형을 포함하고, 상기 선택적 데이터 유형의 상기 순위 점수는 데이터 유형이 특징의 실제 데이터 유형일 확률을 나타내는, 시스템.

청구항 19

제 15 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

특징 목록을 생성하는 것을 포함하고, 상기 특징 목록은 복수의 특징 및 사용자가 예측에 얼마나 중요한지를 나타내는 각각의 특징에 대한 설명을 포함하고, 상기 복수의 특징은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 특징을 포함하는, 시스템.

청구항 20

제 15 항에 있어서,

상기 파이프라인의 표현을 생성하는 것은

알고리즘 목록을 생성하는 것을 포함하고, 상기 알고리즘 목록은 복수의 알고리즘 및 상기 모델의 훈련에 대한 상기 알고리즘의 선택 이유 또는 비선택 이유를 나타내는 각각의 알고리즘에 대한 설명을 포함하고, 상기 복수의 알고리즘은 상기 모델을 훈련시키기 위해 상기 자동 기계 학습 프로세스에서 사용되는 알고리즘을 포함하는, 시스템.

발명의 설명

기 술 분 야

[0001] 설명된 실시예는 일반적으로 데이터 스트림을 처리하는 것에 관련하며, 특히 하이브리드 기계 학습 기술을 사용하여 데이터 스트림에 기초하여 예측을 수행하도록 모델을 훈련하는 것에 관련한다.

배 경 기 술

[0002] 자동 기계 학습 도구는 실제 문제에 기계 학습을 적용하는 프로세스를 자동화한다. 현재, 자동 기계 학습 도구는 원시 데이터세트의 수신으로부터 배치 가능한 기계 학습 모델의 생성에 이르는 완전한 파이프라인을 커버한다. 이러한 도구는 유리하게는 간단한 해결책을 생성하고 이러한 해결책을 빠르고 효율적으로 생성할 수 있게 한다. 그러나, 자동 기계 학습 도구는 데이터세트와 관련된 분야 지식을 통합하지 않기 때문에 차선적인 해결책을 생성하는 경우가 많다. 이 도구는 데이터세트를 이해하고 있는 데이터 분석가가 기계 학습 프로세스를 제어하거나 달리 수정할 수 있는 기능을 거의 제공하지 않는다. 결과적으로, 현재 이용 가능한 자동 기계 학습 도구로 생성된 모델은, 이들이 데이터에 기초하여 예측을 수행할 수 있는 한계만큼 우수하지는 못하다.

발명의 내용

해결하려는 과제

과제의 해결 수단

[0003] 상기 및 기타 문제는 데이터 분석 시스템에서 데이터 블록을 처리하기 위한 방법, 컴퓨터 구현 데이터 분석 시스템 및 컴퓨터 판독가능 메모리에 의해 해결된다. 방법의 실시예는 데이터세트를 수신하는 단계를 포함한다. 이 방법은 수신된 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 생성하는 단계를 더 포함한다. 자동 기계 학습 프로세스는 기계 학습 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함한다. 이 방법은 훈련된 모델이 생성된 후 기계 학습 파이프라인의 표현을 생성하는 단계를 더 포함한다. 이 방법은 사용자 인터페이스에 디스플레이하기 위해 기계 학습 파이프라인의 표현을 제공하는 단계를 더 포함한다. 사용자 인터페이스는 사용자에게 자동 기계 학습 프로세스에서 수행한 판단 중 적어도 일부를 수정할 수 있게 한다. 이 방법은 사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을 수신하는 단계를 더 포함한다. 방법은 또한 사용자로부터의 하나 이상의 수정에 기초하여 모델을 리파이닝(refining)하는 단계를 포함한다. 리파이닝된 모델은 새로운 데이터에 기초하여 예측을 수행하는데 사용된다.

[0004] 컴퓨터 구현 데이터 분석 시스템의 실시예는 컴퓨터 프로그램 명령을 실행하기 위한 컴퓨터 프로세서를 포함한다. 시스템은 또한 동작을 수행하기 위해 컴퓨터 프로세서에 의해 실행가능한 컴퓨터 프로그램 명령을 저장하는 비일시적 컴퓨터 판독가능 메모리를 포함한다. 동작은 데이터세트의 수신을 포함한다. 동작은 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 생성하기 위해 수신된 데이터세트에 대해 자동 기계 학습 프로세스를 수행하는 것을 더 포함한다. 자동 기계 학습 프로세스는 기계 학습 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함한다. 동작은 훈련된 모델이 생성된 후 기계 학습 파이프라인의 표현을 생성하는 것을 더 포함한다. 동작은 사용자 인터페이스에 디스플레이하기 위해 기계 학습 파이프라인의 표현을 제공하는 것을 더 포함한다. 사용자 인터페이스는 사용자에게 자동 기계 학습 프로세스에서 수행한 판단 중 적어도 일부를 수정할 수 있게 한다. 동작은 사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을

수신하는 것을 더 포함한다. 동작은 사용자의 하나 이상의 수정에 기초하여 모델을 리파이닝하는 것을 또한 포함한다. 리파이닝된 모델은 새로운 데이터에 기초하여 예측을 수행하는데 사용된다.

[0005] 비일시적 컴퓨터 판독가능 메모리의 실시예는 실행가능한 컴퓨터 프로그램 명령을 저장한다. 명령은 동작을 수행하기 위해 실행가능하다. 동작은 데이터세트의 수신을 포함한다. 동작은 새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 생성하기 위해 수신된 데이터세트에 대해 자동 기계 학습 프로세스를 수행하는 것을 더 포함한다. 자동 기계 학습 프로세스는 기계 학습 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함한다. 동작은 훈련된 모델이 생성된 후 기계 학습 파이프라인의 표현을 생성하는 것을 더 포함한다. 동작은 사용자 인터페이스에 디스플레이하기 위해 기계 학습 파이프라인의 표현을 제공하는 것을 더 포함한다. 사용자 인터페이스는 사용자에게 자동 기계 학습 프로세스에서 수행한 판단 중 적어도 일부를 수정할 수 있게 한다. 동작은 사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을 수신하는 것을 더 포함한다. 동작은 사용자의 하나 이상의 수정에 기초하여 모델을 리파이닝하는 것을 또한 포함한다. 리파이닝된 모델은 새로운 데이터에 기초하여 예측을 수행하는데 사용된다.

도면의 간단한 설명

[0006] 도 1은 일 실시예에 따른 기계 학습 서버를 포함하는 기계 학습 환경을 나타내는 블록도이다.

도 2는 일 실시예에 따른 자동 모델링 애플리케이션을 나타내는 블록도이다.

도 3은 일 실시예에 따른 보조 모델링 애플리케이션을 예시하는 블록도이다.

도 4는 일 실시예에 따른 하이브리드 기계 학습을 위한 예시적인 사용자 인터페이스이다.

도 5는 일 실시예에 따른 하이브리드 기계 학습 프로세스를 예시하는 흐름도이다.

도 6은 실시예에 따른 도 1의 기계 학습 서버로서 사용하기 위한 전형적인 컴퓨터 시스템의 기능도를 예시하는 고수준 블록도이다.

도면은 단지 예시의 목적으로 다양한 실시예를 묘사한다. 본 기술 분야의 숙련자는 본 출원에 설명된 실시예의 원리를 벗어나지 않고 본 출원에 예시된 구조 및 방법의 대안적 실시예가 채용될 수 있음을 다음 설명으로부터 쉽게 인식할 것이다. 다양한 도면에서 유사한 참조 번호 및 명칭은 유사한 요소를 나타낸다.

발명을 실시하기 위한 구체적인 내용

[0007] 도 1은 일 실시예에 따른 기계 학습 서버(110)를 포함하는 기계 학습 환경(100)을 나타내는 블록도이다. 환경(100)은 네트워크(140)를 통해 기계 학습 서버(110)에 연결된 다수의 데이터 소스(120) 및 클라이언트 디바이스(130)를 더 포함한다. 예시된 환경(100)은 다수의 데이터 소스(120) 및 클라이언트 디바이스(130)에 결합된 단 하나의 기계 학습 서버(110)를 포함하지만, 실시예는 다수의 기계 학습 서버, 단일 데이터 소스 및 단일 클라이언트 디바이스 또는 이들의 다른 변형을 가질 수 있다.

[0008] 기계 학습 서버(110)는 기계 학습 모델을 구성하고 모델을 배치하여 데이터에 기초하여 예측을 수행하는 데 이용되는 컴퓨터 기반 시스템이다. 예측의 예에는 고객이 일정 기간 내에 거래를 수행할 것인지 여부, 거래가 가짜인지 여부, 사용자가 컴퓨터 기반 상호작용을 수행할 것인지 여부 등을 포함한다. 데이터는 네트워크(140)를 통해 다수의 데이터 소스(120) 중 하나 이상 또는 다수의 클라이언트 디바이스(130) 중 하나 이상으로부터 수집, 취합 또는 달리 액세스된다. 기계 학습 서버(110)는 다양한 데이터 소스(120) 또는 클라이언트 디바이스(130)로부터의 데이터를 액세스, 준비, 혼합 및 분석하는 데 채용되는 확장 가능한 소프트웨어 도구 및 하드웨어 자원을 구현할 수 있다.

[0009] 일부 실시예에서, 기계 학습 서버(110)는 하이브리드 기계 학습을 구현하는 컴퓨터 시스템이다. 기계 학습 서버(110)는 자동 모델링 애플리케이션(150) 및 보조 모델링 애플리케이션(160)을 포함한다. 자동 모델링 애플리케이션(150)은 데이터세트에 대해 자동 기계 학습 프로세스를 수행하여 모델을 훈련한다. 대조적으로, 보조 모델링 애플리케이션(160)은 모델을 훈련시키기 위해 데이터세트 및 사용자 입력 모두를 사용하여 보조 기계 학습 프로세스를 수행한다. 사용자 입력은 보조 모델링 애플리케이션(160)이 모델을 훈련시키기 위해 데이터세트를 처리하는 방식을 지정한다. 예를 들어, 사용자 입력은 데이터 유형, 데이터 대치 방법, 특징 또는 알고리즘을 선택하거나, 하이퍼파라미터를 조절하거나, 기계 학습 프로세스에 대한 다른 지침을 제공하거나, 이들의 일부 조합을 제공할 수 있다.

[0010] 두 애플리케이션은 별도로 또는 함께 실행하여 모델을 훈련할 수 있다. 2개의 애플리케이션이 함께 실행될 때, 이들은 자동 모델링 애플리케이션(150)에 의해 생성된 모델을 제한하거나 달리 리파이닝하기 위해 보조 모델링 애플리케이션(160)이 사용되는 하이브리드 기계 학습 프로세스를 통해 모델을 훈련한다.

[0011] 하이브리드 기계 학습 프로세스가 자동 기계 학습 프로세스로 시작하는 실시예에서, 자동 모델링 애플리케이션(150)은 자동 기계 학습 프로세스를 수행하고 모델을 생성하기 위해 기계 학습 파이프라인("파이프라인"이라고도 지칭됨)을 따른다. 파이프라인은 자동 기계 학습 프로세스의 워크플로이며 모델을 훈련하는 일련의 스텝을 지정한다. 하나의 예에서, 파이프라인의 스텝에는 데이터 준비, 특징 가공, 모델 훈련, 모델 검증 및 모델 배치가 포함된다. 스텝은 하위 스텝을 포함할 수 있다. 예를 들어, 데이터 준비 스텝은 데이터 유형 설정, 데이터 인코딩 및 데이터 대치를 포함할 수 있으며, 특징 가공 스텝은 특징 선택 및 특징 순위화를 포함할 수 있고, 모델 훈련 스텝은 하이퍼파라미터 튜닝 및 알고리즘 선택을 포함할 수 있다. 일부 실시예에서, 파이프라인은 본 출원에 설명된 것과 상이한 순서 및/또는 더 많거나 더 적거나 다른 스텝의 스텝을 포함한다. 파이프라인은 사용자 또는 자동 모델링 애플리케이션(150)에 의해 생성될 수 있다.

[0012] 자동 모델링 애플리케이션(150)은 파이프라인의 스텝을 수행할 때 일련의 판단을 수행한다. 자동 모델링 애플리케이션(150)은 각각의 스텝에서 하나 이상의 판단을 수행할 수 있다. 일부 실시예에서, 자동 모델링 애플리케이션(150)은 이러한 스텝에서 수행된 하나 이상의 판단을 최적화하기 위해 파이프라인의 일부 스텝을 반복적으로 처리한다. 자동 모델링 애플리케이션(150)은 비순차적으로 파이프라인의 스텝을 다룰 수 있다. 예를 들어, 자동 모델링 애플리케이션(150)은 특징에 대한 데이터 유형을 설정하기 전에 특징을 선택할 수 있다. 자동 모델링 애플리케이션(150)은 스텝에서 수행된 초기 판단을 최적화하기 위해 후속 스텝을 다룬 후에 해당 스텝을 재방문할 수 있다. 예를 들어, 자동 모델링 애플리케이션(150)은 먼저 변수에 대한 데이터 유형을 선택한 다음 변수에서 특징을 추출하고, 그 후, 변수에 대한 데이터 유형을 변경하기 위해 복귀할 수 있다.

[0013] 훈련된 모델이 생성된 후, 하이브리드 기계 학습 프로세스는 보조 기계 학습 프로세스로 이동한다. 보조 모델링 애플리케이션(160)은 예를 들어 자동 기계 학습 프로세스로부터 자동 모델링 애플리케이션(150)에 의해 수행된 판단 및 최적화를 추출함으로써 자동 모델링 애플리케이션(150)에 의해 사용되는 기계 학습 파이프라인의 표현을 생성한다. 기계 학습 파이프라인의 표현은 파이프라인의 일부 또는 모든 스텝의 표현을 포함할 수 있다. 스텝의 표현은 스텝에 대한 대안 옵션뿐만 아니라 자동 기계 학습 프로세스 동안 자동 모델링 애플리케이션(150)에 의해 수행된 판단을 포함한다. 판단 및 대안 옵션은 이하 "옵션"으로 통칭한다. 스텝의 표현은 옵션에 대한 추천 수준을 나타내는 각각의 옵션에 대한 추천 점수, 사용자가 수정을 수행할 수 있게 도울 수 있는 옵션에 대한 설명 등과 같은 다른 정보를 포함할 수 있다. 일부 실시예에서, 추천 점수는 자동 기계 학습 프로세스에서 자동 모델링 애플리케이션(150)에 의해 결정되고, 보조 모델링 애플리케이션(160)은 자동 기계 학습 프로세스로부터 추천 점수를 도출한다. 일부 실시예에서, 보조 모델링 애플리케이션(160)은 자동 기계 학습 프로세스에 기초하여 추천 점수를 결정한다.

[0014] 보조 모델링 애플리케이션(160)은 예를 들어 GUI에서 사용자에게 디스플레이하기 위한 파이프라인의 표현을 제공한다. 사용자는 파이프라인의 스텝을 검토 및/또는 자동 기계 학습 프로세스에서 수행한 판단을 수정할 수 있다. 보조 모델링 애플리케이션(160)은 사용자로부터 수정을 수신하고 수정에 기초하여 훈련된 모델을 리파이닝할 수 있다. 그 후, 리파이닝된 훈련된 모델을 배치하여 새로운 데이터에 기초하여 예측을 수행한다.

[0015] 일부 실시예에서, 하이브리드 기계 학습 프로세스는 보조 기계 학습 프로세스로 시작한다. 보조 모델링 애플리케이션(160)은 사용자 입력을 수신하고 사용자 입력에 기초하여 파이프라인 세트를 생성한다. 예를 들어, 보조 모델링 애플리케이션(160)은 사용자 인터페이스에서 추천(예를 들어, 데이터 유형, 변환기, 특징, 알고리즘 및/또는 하이퍼파라미터의 추천)을 제시할 수 있고 사용자는 사용자 인터페이스를 통해 추천에 기초하여 사용자 입력을 제공할 수 있다. 보조 모델링 애플리케이션(160)은 파이프라인 세트를 자동 모델링 애플리케이션(150)으로 발신하고 자동 모델링 애플리케이션(150)은 파이프라인 세트에 기초하여 자동 기계 학습 프로세스를 수행한다. 순수한 자동 기계 학습 프로세스와 비교할 때, 하이브리드 기계 학습 프로세스는 자동 모델링 애플리케이션(150)이 그 검색/최적화를 파이프라인 세트로 제한할 수 있기 때문에 더 적은 시간 및/또는 컴퓨팅 자원을 소비한다. 또한, 파이프라인 세트를 생성하기 위해 사용자 입력을 사용함으로써, 자동 모델링 애플리케이션(150)이 일반적으로 가지고 있지 않은 사용자의 도메인 지식을 이용한다.

[0016] 데이터 소스(120)는 기계 학습 서버(110)에 전자 데이터를 제공한다. 데이터 소스(120)는 저장 디바이스,

예컨대, 하드 디스크 드라이브(HDD) 또는 솔리드 스테이트 드라이브(SSD), 다수의 저장 디바이스에 대한 액세스를 관리 및 제공하는 컴퓨터, 저장 영역 네트워크(SAN), 데이터베이스 또는 클라우드 저장소 시스템일 수 있다.

데이터 소스(120)는 또한 다른 소스로부터 데이터를 검색할 수 있는 컴퓨터 시스템일 수 있다. 데이터 소스(120)는 기계 학습 서버(110)로부터 원격일 수 있고 네트워크(140)를 통해 데이터를 제공할 수 있다. 또한, 일부 또는 모든 데이터 소스(120)는 데이터 분석 시스템에 직접 결합될 수 있고 네트워크(140)를 통한 데이터 전달 없이 데이터를 제공할 수 있다.

[0017] 데이터 소스(120)에 의해 제공된 데이터는 데이터 레코드(예를 들어, 행)로 구성될 수 있다. 각각의 데이터 레코드에는 하나 이상의 값이 포함된다. 예를 들어, 데이터 소스(120)에 의해 제공되는 데이터 레코드는 일련의 쉼표로 구분된 값을 포함할 수 있다. 데이터는 데이터 분석 시스템(110)을 사용하여 기업과 관련된 정보를 설명한다. 예를 들어, 데이터 소스(120)로부터의 데이터는 웹사이트에서 액세스 가능한 콘텐츠 및/또는 애플리케이션과의 컴퓨터 기반 상호작용(예를 들어, 클릭 추적 데이터)을 설명할 수 있다. 다른 예로서, 데이터 소스(120)로부터의 데이터는 온라인 및/또는 매장에서의 고객 거래를 설명할 수 있다. 기업은 제조, 판매, 금융 및 은행업무와 같은 다양한 산업 중 하나 이상에 속할 수 있다.

[0018] 클라이언트 디바이스(130)는 사용자 입력을 수신할 수 있을 뿐만 아니라 네트워크(140)를 통해 데이터를 송신 및/또는 수신할 수 있는 하나 이상의 컴퓨팅 디바이스이다. 일 실시예에서, 클라이언트 디바이스(130)는 데스크탑 또는 랩톱 컴퓨터와 같은 종래의 컴퓨터 시스템이다. 대안적으로, 클라이언트 디바이스(130)는 PDA(personal digital assistant), 이동 전화, 스마트폰 또는 다른 적절한 디바이스와 같은 컴퓨터 기능을 갖는 디바이스일 수 있다. 클라이언트 디바이스(130)는 네트워크(140)를 통해 하나 이상의 데이터 소스(120) 및 기계 학습 서버(110)와 통신하도록 구성된다. 일 실시예에서, 클라이언트 디바이스(130)는 클라이언트 디바이스(130)의 사용자가 기계 학습 서버(110)와 상호작용할 수 있게 하는 애플리케이션을 실행한다. 예를 들어, 클라이언트 디바이스(130)는 예를 들어 기계 학습 서버(110)에 의해 지원되는 GUI를 운영하는 것을 통해 네트워크(140)를 통해 클라이언트 디바이스(130)와 기계 학습 서버(110) 사이의 상호작용을 가능하게 하도록 애플리케이션을 실행한다. 클라이언트 디바이스(130)는 GUI를 디스플레이하는 디스플레이 디바이스를 포함하거나 이와 달리 연관되어 있다. 클라이언트 디바이스(130)는 또한 사용자가 GUI에 입력을 제공하는 것과 같이 GUI와 상호작용할 수 있게 하는 입력 디바이스, 예를 들어 키보드, 마우스 등과 연관되어 있다. 다른 실시예에서, 클라이언트 디바이스(130)는 IOS® 또는 ANDROID™와 같은 클라이언트 디바이스(130)의 네이티브 운영 체제에서 실행되는 애플리케이션 프로그래밍 인터페이스(API)를 통해 기계 학습 서버(110)와 상호작용한다. 클라이언트 디바이스(130)는 하나 이상의 데이터 소스(120)와 상호작용하여 데이터를 데이터 소스(120)로 송신하거나 데이터 소스(120)로부터 데이터를 획득할 수 있다.

[0019] 네트워크(140)는 기계 학습 서버(110)와 데이터 소스(120) 사이의 통신 경로를 나타낸다. 일 실시예에서, 네트워크(140)는 인터넷이고 표준 통신 기술 및/또는 프로토콜을 사용한다. 따라서, 네트워크(140)는 이더넷, 802.11, WiMAX(worldwide interoperability for microwave access), 3G, LTE(Long Term Evolution), DSL(digital subscriber line), ATM(asynchronous transfer mode), InfiniBand, PCI 익스프레스 어드밴스드 스위칭 등 같은 기술을 사용하는 링크를 포함할 수 있다. 마찬가지로, 네트워크(140)에서 사용되는 네트워킹 프로토콜은 MPLS(multiprotocol label switching), TCP/IP(transmission control protocol/Internet protocol), UDP(User Datagram Protocol), HTTP(hypertext transport protocol), SMTP(simple mail transfer protocol), FTP(file transfer protocol) 등을 포함할 수 있다.

[0020] 네트워크(140)를 통해 교환되는 데이터는 HTML(hypertext markup language), XML(extensible markup language) 등을 포함하는 기술 및/또는 형식을 사용하여 표현될 수 있다. 또한, SSL(secure sockets layer), TLS(transport layer security), VPN(virtual private network), IPsec(internet protocol security) 등과 같은 종래의 암호화 기술을 사용하여 링크의 전부 또는 일부를 암호화할 수 있다. 다른 실시예에서, 엔티티는 앞서 설명한 것 대신에 또는 이에 더하여 맞춤형 및/또는 전용 데이터 통신 기술을 사용할 수 있다.

[0021] 도 2는 일 실시예에 따른 자동 모델링 애플리케이션(200)을 나타내는 블록도이다. 자동 모델링 애플리케이션(200)은 도 1의 자동 모델링 애플리케이션(150)의 실시예이다. 자동 모델링 애플리케이션(200)은 파이프라인을 사용하여 모델을 훈련시키기 위해 데이터세트에 대한 자동 기계 학습 프로세스를 수행한다. 훈련된 모델은 사용자가 정의할 수 있는 목표 변수를 예측을 수행하는데 사용된다. 자동 모델링 애플리케이션(200)은 데이터 준비 모듈(210), 특징 가공 모듈(220), 모델 훈련 모듈(230), 모델 검증 모듈(240) 및 데이터베이스(250)를 포함한다. 본 기술 분야의 숙련자는 다른 실시예가 여기서 설명된 것과 상이한 및/또는 다른 컴포넌트를 가질 수 있고 기능이 다른 방식으로 컴포넌트 사이에 분배될 수 있음을 인식할 것이다. 자동 모델

링 애플리케이션(200)의 컴포넌트는 파이프라인의 스텝에서 일련의 판단을 수행함으로써 자동 기계 학습 프로세스를 수행하기 위해 함께 동작한다.

[0022] 데이터 준비 모듈(210)은 데이터세트의 데이터를 처리하여 모델 훈련을 위한 훈련 데이터세트를 준비한다. 데이터 준비 모듈(210)은 데이터세트 내에 연관된 변수에 대한 데이터 유형을 결정한다. 데이터세트와 연관된 변수는 데이터세트의 변수이거나 데이터세트의 하나 이상의 변수로부터 변환될 수 있다. 일부 실시예에서, 데이터세트와 연관된 변수는 예측 변수, 즉, 특징이다. 일부 실시예에서, 데이터 준비 모듈(210)은 변수가 수치 데이터 유형인지, 범주형 데이터 유형인지, 시계열 데이터 유형인지, 우편번호 데이터 유형인지 또는 텍스트 데이터 유형인지 여부와 같은 변수의 데이터 유형을 선택적 데이터 유형의 풀에서 선택한다. 일부 실시예에서, 데이터 준비 모듈(210)은 규칙 기반 분석을 통해 변수에 대한 하나 이상의 데이터 유형을 결정한다. 결정은 데이터 준비 모듈(210)에 의해 유지되는 규칙에 기초한다. 하나의 예에서, 규칙은 변수의 데이터 값에 기초하여 변수에 대한 데이터 유형을 지정하며, 예를 들어 수치 값을 포함하는 변수는 수치 데이터 유형을 갖고 텍스트 값을 포함하는 변수는 텍스트 데이터 유형을 갖는다. 데이터 준비 모듈(210)이 변수가 숫자라고 결정한 경우, 이는 규칙에 따라 변수가 정수인지 실수인지 여부를 추가로 결정할 수 있다. 정수인 경우, 데이터 준비 모듈(210)은 그 후 변수의 고유한 정수 값의 수를 결정할 수 있다. 데이터 준비 모듈(210)이 변수의 고유 정수 값의 수가 임계값 미만이라고 결정하는 경우, 규칙은 변수의 데이터 유형이 범주형임을 나타낸다. 규칙은 또한 변수의 데이터 유형이 숫자일 수 있음을 제안할 수 있다. 또 다른 예에서 규칙은 변수에 대한 설명에 기초하여 변수에 대한 데이터 유형을 지정하고, 예를 들어 데이터세트에서 "사용자 ID"라는 명칭의 변수는 ID 데이터 유형을 갖고 "생년월일"이라는 명칭의 변수는 시계열 데이터 유형을 가진다.

[0023] 데이터 준비 모듈(210)은 변수의 각각의 선택적 데이터 유형별로 순위 점수를 결정할 수 있다. 데이터 유형의 순위 점수는 선택적 데이터 유형이 변수의 실제 데이터 유형일 확률을 나타낸다. 일부 실시예에서, 데이터 준비 모듈(210)에 의해 유지되는 규칙은 변수에 대한 다양한 데이터 유형의 확률을 나타낼 수 있다. 앞서 설명된 예에서, 데이터 준비 모듈(210)이 변수의 고유한 정수 값의 수가 임계값 미만이라고 결정하는 경우, 규칙은 범주형 데이터 유형이 수치 데이터 유형보다 더 높은 확률을 가짐을 나타낼 수 있다. 데이터 준비 모듈(210)은 또한 정수 값이 우편번호인지 여부를 결정하고 규칙에 따라 우편번호에 대한 확률을 결정할 수 있다. 또 다른 예에서 규칙은 데이터세트에서 "우편번호"라는 명칭의 변수의 경우 범주형 데이터 유형 및 수치 데이터 유형에 대한 확률이 시계열 데이터 유형 및 텍스트 데이터 유형에 대한 확률보다 더 높다는 것을 나타낸다. 또 다른 예에서 규칙은 텍스트 값을 포함하는 변수의 경우, 범주형 데이터 유형 및 텍스트 데이터 유형에 대한 확률이 시계열 데이터 유형 및 수치 데이터 유형에 대한 확률보다 더 높다는 것을 나타낸다.

[0024] 일부 실시예에서, 데이터 준비 모듈(210)은 데이터 유형을 사용하여 훈련된 모델의 성능을 평가하여 각각의 선택적 데이터 유형의 순위 점수를 결정할 수 있다. 예를 들어, 데이터 준비 모듈(210)은 선택적 데이터 유형을 사용하여 모델을 훈련시키고 모델의 성능에 기초하여 선택적 데이터 유형의 순위 점수를 결정할 수 있다. 데이터 준비 모듈(210)은 다양한 데이터 유형을 사용하여 훈련된 모델의 성능을 비교하고(모델의 기계 학습 프로세스에서 다른 판단은 동일할 수 있음) 비교에 기초하여 데이터 유형의 순위 점수를 결정할 수 있다. 예를 들어, 데이터 준비 모듈(210)은 해당 데이터 유형을 사용하여 훈련한 모델이 다른 데이터 유형을 사용하여 훈련한 모델보다 더 나은 성능을 나타내는 경우 해당 데이터 유형의 순위 점수가 다른 데이터 유형보다 더 높은 것으로 결정할 수 있다. 순위 점수/확률은 특징, 알고리즘, 하이퍼파라미터 등에 대한 검색과 같은 자동 기계 학습 프로세스에서 자동 모델링 애플리케이션(200)에 의해 수행되는 추가 검색을 제한하는데 사용될 수 있다.

[0025] 데이터 준비 모듈(210)은 순위 점수에 기초하여 4가지 데이터 유형 중 하나를 변수의 데이터 유형으로 선택한다. 예를 들어, 데이터 준비 모듈(210)은 확률이 가장 높은 데이터 유형을 선택한다.

[0026] 데이터 준비 모듈(210)은 결정된 데이터 유형에 기초하여 데이터세트의 데이터를 인코딩할 수 있다. 일부 데이터 유형의 값(예를 들어, 범주형 값)은 모델 훈련에 더 적절한 다른 표현으로 인코딩된다. 일부 실시예에서, 데이터 준비 모듈(210)은 범주형 변수, 시계열 변수 및/또는 텍스트 변수의 값을 인코딩하여 값을 이진 값으로 변환한다. 데이터 준비 모듈(210)은 텍스트를 수치 값으로 변환, 라벨 인코딩, 원 핫 인코딩, 커스텀 이진 인코딩, 후향 차분 인코딩, 다항식 인코딩 등과 같은 다양한 방법을 사용하여 날짜를 인코딩할 수 있다.

[0027] 데이터 준비 모듈(210)은 데이터세트에서 누락된 값을 검출하고 데이터 대치를 수행하여 값을 공급한다. 일

부 실시예에서, 데이터 준비 모듈(210)은 누락된 값을 대체하기 위해 데이터세트의 현재 값에 기초하여 새로운 값을 결정한다. 예를 들어, 데이터 준비 모듈(210)은 누락된 값이 있는 각각의 열에 대해 해당 열의 누락된 값을 해당 열의 현재 값의 평균 또는 중앙값으로, 해당 열에서 가장 빈도가 높은 값으로, 또는 데이터세트에 없는 새로운 샘플로부터의 값으로 대체한다. 데이터 준비 모듈(210)은 kNN(k-Nearest Neighbor) 대치, 핫 데크 대치, 콜드 데크 대치, 회귀 대치, 확률적 회귀 대치, 외삽 및 보간, 단일 대치, 다중 대치, MICE(Multivariate Imputation by Chained Equation), 심층 신경망을 사용한 대치 등 같은 다른 대치 방법을 사용할 수 있다. 데이터 준비 모듈(210)은 다수의 대치 방법을 식별하고 대응 변수의 데이터 유형에 기초하여 식별된 대치 방법을 순위화할 수 있다. 하나의 예에서, 수치 변수에 대해, 데이터 준비 모듈(210)은 데이터세트에 이상값이 있는지를 결정하고, 만약 있다면, 평균의 대치 방법보다 중앙값의 대치 방법을 더 높게 순위화한다.

[0028] 특징 가공 모듈(220)은 데이터세트로부터 특징을 추출한다. 특징 가공 모듈(220)은 데이터세트의 변수를 특징으로 추출할 수 있고 및/또는 변환기를 사용하여 데이터세트의 변수를 특징으로 변환할 수 있다. 변환기가 변수 값에 적용되면, 이는 특징의 값을 생성한다. 일부 실시예에서, 특징 가공 모듈(220)은 데이터세트의 변수, 목표 변수, 목표 변수와 관련된 비즈니스 문제 등과 같은 하나 이상의 요인에 기초하여 모델 훈련 모듈(230) 변환기의 풀에서 변환기를 선택한다.

[0029] 특징 가공 모듈(220)은 특징을 순위화하고 각각의 특징에 대한 순위 점수를 결정한다. 특징의 순위 점수는 목표 변수를 예측을 수행하는 데 특징이 얼마나 중요한지, 달리 말해서, 특징이 예측자로서 얼마나 우수한지를 나타낸다. 일부 실시예에서, 특징 가공 모듈(220)은 특징 및 데이터세트에 기초하여 랜덤 포레스트를 구성한다. 특징 가공 모듈(220)은 랜덤 포레스트의 각각의 결정 트리에 기초하여 특징의 순위 점수를 결정하고 개별 순위 점수의 평균을 특징의 순위 점수로서 획득한다. 특징 가공 모듈(220)은 특징이 전체 예측 모델에 얼마나 기여하는 지를 측정하기 위해 각각의 결정 트리의 일부로서 GINI 불순도를 사용할 수 있다. 랜덤 포레스트를 사용하여 결정된 특징의 순위 점수는 해당 특징이 다른 특징에 비교하여 상대적으로 얼마나 중요한지를 나타내는 것으로 "상대 순위 점수"라 지칭된다. 하나의 예에서, 순위 모듈(330)은 가장 높은 순위의 선택된 특징의 상대 순위 점수가 1이라고 결정한다. 순위 모듈(330)은 그 후, 나머지 특징 각각의 순위 점수와 가장 높은 순위 특징의 순위 점수의 비율을 대응하는 선택된 특징의 상대 순위 점수로 결정한다.

[0030] 특징 가공 모듈(220)은 예를 들어 GKT(Goodman-Kruskal Tau) 측정에 기초하여 각각의 선택된 특징에 대한 절대 순위 점수를 결정할 수 있다. GKT 측정은 로컬 또는 절대적인 연관성의 척도이며 특징이 목표를 얼마나 잘 예측하는지 나타낸다. 특징 가공 모듈(220)은 모델을 훈련시키기 위한 특징으로서 그 상대 순위 점수 및/또는 절대 순위 점수에 기초하여 특징 그룹의 부분집합을 선택할 수 있다.

[0031] 모델 훈련 모듈(230)은 특징 가공 모듈(220)에 의해 결정된 특징 및 그의 순위 점수에 기초하여 모델을 훈련한다. 일부 실시예에서, 모델 훈련 모듈(230)은 후보 알고리즘의 풀로부터 알고리즘("추정기"라고도 지칭됨)을 선택한다. 후보 알고리즘의 예는 예를 들어 결정 트리, 로지스틱 회귀, 랜덤 포레스트, XGBoost, 선형 SVM(linear support vector machine), AdaBoost, 신경망, 나이브 베이즈, 메모리 기반 학습, 랜덤 포레스트, 배깅 트리, 부스트 트리, 부스트 스텀프 등을 포함한다. 모델 훈련 모듈(230)은 선택된 알고리즘을 사용하여 모델을 훈련시킨다. 일부 실시예에서, 모델 훈련 모듈(230)은 이용 가능한 정보, 예를 들어, 모델 훈련을 위한 시간 제한, 연산 자원 제한(예를 들어, 프로세서 제한, 메모리 사용 제한 등), 해결해야 할 예측 문제, 데이터세트의 특성, 선택한 특징 등에 기초하여 풀 내의 후보 알고리즘의 수를 제한할 수 있다. 모델 훈련 모듈(230)은 각각의 후보 알고리즘을 테스트하고 최상의 것을 선택할 수 있다. 모델 훈련 모듈(230)은 후보 알고리즘의 성능을 평가하기 위해 성능 측정(예를 들어, 분류 정확도)과 연관된 테스트 하네스를 정의할 수 있다. 예를 들어, 모델 훈련 모듈(230)은 후보 알고리즘으로 훈련된 모델을 검증 데이터세트(모델 훈련에 사용된 데이터세트와 다른 데이터세트)에 적용하여 훈련된 모델의 정확도를 정량화한다. 정확도 측정에 적용되는 일반적인 메트릭은 정밀도 = TP/(TP + FP) 및 재현율 = TP/(TP + FN)이며, 여기서, 정밀도는 모델이 예측한 전체 결과(TP + FP 또는 위양성) 중에서 모델이 올바르게 예측한 결과(TP 또는 진양성)가 얼마나 많은지이고, 재현율은 실제로 발생한 총 수(TP + FN 또는 위음성) 중에서 모델이 올바르게 예측한 결과(TP)가 얼마나 많은지이다. F 점수(F-score = 2 * PR/(P + R))는 정밀도와 재현율을 단일 척도로 통합한다.

[0032] 테스트 하네스에 대해 후보 알고리즘을 테스트한 결과는 후보 알고리즘이 성능 측정에 대해 예측 문제에서 기능하는 방식을 추정한다. 모델 훈련 모듈(230)은 최상의 성능을 갖는 후보 알고리즘을 선택한다. 일부 실시예에서, 모델 훈련 모듈(230)은 선택된 후보 알고리즘을 더 최적화한다. 모델 훈련 모듈(230)은 후보 알

고리즘의 테스트 결과 또는 다른 정보에 기초하여 후보 알고리즘의 장단점을 분석할 수 있다.

[0033] 모델 훈련 모듈(230)은 또한 선택된 알고리즘에 의해 요구되는 하이퍼파라미터를 결정할 수 있다. 하이퍼파라미터는 기계 학습 프로세스를 제어하는데 사용되는 값을 가진 파라미터이다. 하이퍼파라미터는 훈련 프로세스에서 사용되며 훈련 프로세스의 속도나 품질에 영향을 미칠 수 있다. 하이퍼파라미터의 예는 학습률, 미니 배치 크기, 서포트 벡터 머신을 위한 C 및 시그마 하이퍼파라미터, 신경망 크기, 신경망 토폴로지를 포함한다. 모델 훈련 모듈(230)은 그리드 검색, 랜덤 검색 또는 다른 방법으로 하이퍼파라미터를 결정할 수 있다. 일부 실시예에서, 모델 훈련 모듈(230)은 자동 기계 학습 또는 다른 기술에 대해 지정된 디폴트 값을 사용하여 과거 훈련 프로세스로부터 하나 이상의 하이퍼파라미터를 유도함으로써 하이퍼파라미터를 획득한다.

데이터베이스(240)는 자동 모델링 애플리케이션(200)에 의해 수신, 사용 및 생성된 데이터와 같은 자동 모델링 애플리케이션(200)과 연관된 데이터를 저장한다. 예를 들어, 데이터베이스(240)는 데이터세트, 파이프라인, 파이프라인 스텝에서 수행된 판단, 훈련 데이터세트, 특징, 변환기, 알고리즘, 하이퍼파라미터, 훈련된 모델 등을 저장한다.

[0034] 도 3은 일 실시예에 따른 보조 모델링 애플리케이션(300)을 예시하는 블록도이다. 보조 모델링 애플리케이션(300)은 도 1의 보조 모델링 애플리케이션(160)의 실시예이다. 보조 모델링 애플리케이션(300)은 사용자 입력에 기초하여 자동 모델링 애플리케이션(200)에 의해 훈련된 모델을 리파이닝한다. 보조 모델링 애플리케이션(300)은 파이프라인 표현 모듈(310), 사용자 인터페이스 모듈(320), 모델 리파이닝 모듈(330) 및 데이터베이스(340)를 포함한다. 본 기술 분야의 숙련자는 다른 실시예가 여기서 설명된 것과 상이한 및/또는 다른 컴포넌트를 가질 수 있고 기능이 다른 방식으로 컴포넌트 사이에 분배될 수 있음을 인식할 것이다.

[0035] 파이프라인 표현 모듈(310)은 기계 학습 파이프라인의 표현을 생성한다. 일부 실시예에서, 파이프라인 표현 모듈(310)은 자동 모델링 애플리케이션(150)과 같은 자동 모델링 애플리케이션으로부터 기계 학습 파이프라인을 획득한다. 예를 들어, 파이프라인 표현 모듈(310)은 자동 모델링 애플리케이션에 의해 수행되는 자동 기계 학습 프로세스로부터 기계 학습 파이프라인을 검색한다. 일부 실시예에서, 파이프라인 표현 모듈(310)은 자동 기계 학습 프로세스에서 사용되고 생성된 데이터와 같은 자동 모델링 애플리케이션(200)의 데이터베이스(240)에 저장된 데이터를 검색한다. 파이프라인 표현 모듈(310)은 데이터를 분석하고 데이터를 파이프라인의 스텝과 각각의 스텝의 옵션에 맵핑한다. 파이프라인 표현 모듈(310)은 분석 및 맵핑에 기초하여 파이프라인의 표현을 생성한다. 일부 실시예에서, 파이프라인 표현 모듈(310)은 하나보다 더 많은 파이프라인에 대해 자동 모델링 애플리케이션(200)에 쿼리한다. 예를 들어, 파이프라인 표현 모듈(310)은 자동 모델링 애플리케이션(200)에 의해 식별된 미리 결정된 수의 파이프라인에 대해 쿼리한다. 쿼리에 응답하여, 자동 모델링 애플리케이션(200)은 그 검색에서 미리 결정된 수의 최상의 파이프라인을 파이프라인 표현 모듈(310)로 발신한다. 파이프라인 표현 모듈(310)은 모델 정확도와 같은 파이프라인을 사용하여 훈련된 모델의 성능 측정에 기초하여 자동 모델링 애플리케이션(200)으로부터 수신된 파이프라인 중 하나를 선택할 수 있다. 파이프라인 표현 모듈(310)은 그 후 선택된 파이프라인 및 파이프라인의 스텝/컴포넌트를 변환하여 파이프라인의 표현을 생성한다.

[0036] 파이프라인 표현 모듈(310)은 파이프라인의 하나 이상의 스텝에 대한 표현을 생성한다. 스텝의 표현("스텝 표현"이라고도 지칭됨)에는 스텝에 대한 옵션이 포함된다. 옵션 중 하나는 자동 기계 학습 프로세스 중 스텝에서 수행한 판단이다. 표현은 또한 각각의 옵션에 대한 순위 점수 또는 추천 지표를 포함할 수 있다. 순위 점수 또는 추천 지표는 옵션에 대한 추천 수준을 나타낸다. 일부 스텝의 경우, 표현에 각각의 옵션에 대한 설명이 포함된다. 설명에는 예를 들어 옵션의 기능에 대한 설명, 옵션의 사정(예를 들어, 장단점), 옵션의 평가, 옵션을 선택 및/또는 비선택 이유에 대한 설명 등이 포함된다. 순위 점수 및/또는 설명은 자동 모델 애플리케이션(200) 또는 파이프라인 표현 모듈(310)에 의해, 예를 들어 도 2와 관련하여 앞서 설명한 기술을 사용하여 결정될 수 있다. 일부 실시예에서, 순위 점수 및/또는 옵션의 설명은 옵션을 사용하여 훈련된 모델의 성능 측정(예를 들어, 예측 정확도)에 기초하여 결정된다. 예를 들어, 옵션을 사용하여 훈련된 모델의 성능이 더 좋다고 결정되면 옵션의 추천 점수가 더 높다.

[0037] 하나의 예에서, 파이프라인 표현 모듈(310)은 데이터 유형 설정 스텝의 표현으로서 데이터 유형 목록을 생성한다. 데이터 유형 목록은 하나 이상의 변수(예를 들어, 자동 기계 학습 프로세스에 사용되는 특징)와 연관된다. 여기에는 각각의 특징에 대한 복수의 선택적 데이터 유형이 포함된다. 데이터 유형 목록에는 각각의 선택적 데이터 유형에 대한 순위 점수도 포함된다. 순위 점수는 선택적 데이터 유형이 특징의 실제 데이터 유형일 확률을 나타낸다.

[0038] 다른 예에서, 파이프라인 표현 모듈(310)은 데이터 대치 스텝의 표현으로서 데이터 대치 목록을 생성한다. 데이터 대치 목록에는 자동 기계 학습 프로세스에서 대치된 누락 값이 있는 하나 이상의 특징이 포함된다. 각각의 특징에 대해, 데이터 대치 목록에는 여러 선택적 대치 방법이 포함되어 있으며, 여기에는 특징의 값을 대치하기 위해 자동 기계 학습 프로세스에서 사용되는 대치 방법이 포함된다. 데이터 대치 목록은 특징의 전체 행에서 누락 값이 있는 행의 백분율을 나타내는 각각의 특징에 대한 백분율을 포함할 수 있다. 파이프라인 표현 모듈(310)은 각각의 선택적인 대치 방법에 기초하여 대체 값을 계산할 수 있고, 데이터 대치 스텝의 표현에 대체 값을 포함할 수 있다.

[0039] 또 다른 예에서, 파이프라인 표현 모듈(310)은 특징 가공 스텝의 표현으로서 특징 목록을 생성한다. 특징 목록에는 자동 기계 학습 프로세스에서 사용되는 특징의 일부 또는 전부가 포함되며 자동 기계 학습 프로세스에서 사용되지 않은 특징도 포함될 수 있다. 특징 목록은 특징이 예측에 얼마나 중요한지를 나타내는 각각의 특징의 순위 점수, 각각의 특징에 대한 설명, 특징에 대한 평가 또는 이들의 일부 조합을 포함할 수 있다. 특징 목록에는 특징 추출에 사용되는 변환기도 포함될 수 있다.

[0040] 또 다른 예에서, 파이프라인 표현 모듈(310)은 알고리즘 선택 스텝의 표현으로서 알고리즘 목록을 생성한다. 알고리즘 목록에는 모델을 훈련시키기 위해 사용되는 선택적 알고리즘이 포함된다. 선택적 알고리즘 중 하나는 자동 기계 학습 프로세스에서 선택되고 사용되는 알고리즘이며, 파이프라인 표현 모듈(310)은 이를 알고리즘 목록에서 "추천" 알고리즘으로 라벨링할 수 있다. 각각의 선택적 알고리즘에 대해 알고리즘 목록은 선택적 알고리즘의 장단점을 보여주는 설명을 포함할 수 있다. 장단점은 리파이닝할 모델에 특정할 수 있다. 파이프라인 표현 모듈(310)은 또한 예를 들어, 자동 기계 학습 프로세스에서 선택적 알고리즘의 성능의 테스트 결과에 기초하여 각각의 선택적 알고리즘에 대한 순위 점수를 결정할 수 있다. 하나의 예에서, 파이프라인 표현 모듈(310)은 선택적 알고리즘을 사용하여 훈련된 모델의 측정된 정확도에 기초하여 순위 점수를 결정한다.

[0041] 사용자 인터페이스 모듈(320)은 GUI를 지원하고 GUI를 통한 제시를 위한 파이프라인의 표현을 제공한다. GUI는 사용자가 파이프라인의 표현을 보고, 파이프라인에 기초하여 하는 자동 기계 학습 프로세스에서 수행한 판단을 검토하고, 적어도 일부 판단의 수정을 수행할 수 있게 한다. 하나의 예에서, GUI는 사용자가 자동 기계 학습 프로세스에 사용되는 변수의 데이터 유형을 상이한 데이터 유형으로 변경할 수 있게 한다. 또 다른 예에서, GUI는 사용자가 변환기 변경, 자동 기계 학습 프로세스에서 선택된 특징 제거, 자동 기계 학습 프로세스에서 선택되지 않은 특징 추가 또는 자동 기계 학습 프로세스에서 결정된 특징 순위의 변경과 같은 특징 편집을 허용한다. 예를 들어, 사용자는 자동 모델링 애플리케이션(200)의 분야 지식 부족으로 자동 기계 학습 프로세스에서 목표 누설을 초래함에도 불구하고 선택되었던 특징을 제거하기 위해 자신의 분야 지식을 사용할 수 있다. 또한, GUI는 사용자가 알고리즘 및/또는 하이퍼파라미터를 변경할 수 있게 한다.

[0042] 일부 실시예에서, GUI는 사용자가 파이프라인의 스텝을 선택하고 수정하기 위해 상호작용할 수 있는 복수의 제어 요소를 포함한다. 제어 요소의 예는 체크박스, 버튼, 탭, 아이콘, 드롭다운 목록, 목록 상자, 라디오 버튼, 토글, 텍스트 필드, 날짜 필드를 포함한다. 예를 들어, GUI에는 각각의 스텝에 대한 탭이 포함되어 있으며 사용자는 탭을 클릭하여 스텝에 대한 판단 및 대안 옵션에 액세스할 수 있다. 각각의 스텝의 옵션은 사용자가 한 번에 하나 이상의 옵션을 선택할 수 있게 하는 드롭다운 목록 또는 체크박스로 제시될 수 있다. 대응 스텝에 대한 판단은 대안 옵션과 구별될 수 있으며, 예를 들어 판단은 GUI에서 "추천"으로 표시될 수 있거나 강조 표시될 수 있다. GUI는 사용자가 판단을 유지하거나(예를 들어, 아무것도 하지 않거나 판단을 나타내는 아이콘을 클릭하여) 대안 옵션을 선택(예를 들어, 대안 옵션을 나타내는 아이콘을 클릭)할 수 있게 한다.

[0043] 일부 실시예에서, GUI는 사용자가 파이프라인의 스텝을 비순차적으로 탐색할 수 있게 한다. 예를 들어, 사용자가 데이터 유형 설정 스텝을 검토하기 전에 사용자는 특징 선택 스텝을 검토할 수 있다. GUI는 사용자가 스텝을 재방문하고 및/또는 스텝을 다수의 수정을 수행할 수 있게 한다.

[0044] GUI는 다른 유형의 사용자 입력을 허용할 수 있다. 예를 들어, GUI는 사용자가 목표 변수 수정, 데이터세트 편집, 모델 리파이닝을 위한 자동화 수준 선택 등을 수행할 수 있게 한다.

[0045] 모델 리파이닝 모듈(330)은 GUI를 통해 사용자로부터 수신한 수정에 기초하여 훈련된 모델을 리파이닝한다. 일부 실시예에서, 모델 리파이닝 모듈(330)은 자동 모델링 애플리케이션(200)에 수정을 발신하고 자동 모델링 애플리케이션(200)은 수정을 사용하여 자동 기계 학습 프로세스의 일부 또는 전부를 튜닝한다. 자동 모델링 애플리케이션(200)은 수정에 기초하여 파이프라인에서 그 판단을 업데이트하고 업데이트된 판단에 기초하여 새

로운 모델을 훈련할 수 있다. 예를 들어, 자동 모델링 애플리케이션(200)은 사용자의 수정을 사용하여 후보 알고리즘 또는 하이퍼파라미터에 대한 그 검색을 제한한다. 하나의 예에서, GUI는 모델을 리파이닝하는데 사용될 하나 이상의 하이퍼파라미터에 대한 사용자의 명세를 수신하고, 모델 리파이닝 모듈(330)은 하나 이상의 하이퍼파라미터를 자동 모델링 애플리케이션(200)으로 발신하고, 자동 모델링 애플리케이션(200)은 이 하나 이상의 하이퍼파라미터를 사용하여 새로운 파이프라인과 알고리즘을 검색한다. 자동 모델링 애플리케이션(200)은 그 후 새로운 파이프라인 및 알고리즘을 사용하여 모델을 재훈련한다.

[0046] 일부 실시예에서, 모델 리파이닝 모듈(330)은 예를 들어, 도 2와 관련하여 앞서 설명한 기계 학습 기술을 사용함으로써 사용자의 수정에 기초하여 모델을 재훈련한다. 모델 리파이닝 모듈(330)은 훈련된 모델을 리파이닝하기 위해 파이프라인의 일부 또는 전부를 수행할 수 있으며, 이는 자동 기계 학습 프로세스와 유사한 판단을 수행하거나 새로운 판단을 수행하는 것을 수반한다. 예를 들어, 수치에서 범주형으로의 특징의 데이터 유형의 사용자 수정을 수신한 이후, 모델 리파이닝 모듈(330)은 특징의 범주형 값을 수치 특징으로 변환하기 위한 데이터 인코딩 방법을 선택하여 새로운 판단을 수행한다. 모델 리파이닝 모듈(330)은 또한 리파이닝된 모델을 검증하고 검증 후 새로운 데이터에 기초하여 예측을 수행하기 위해 리파이닝된 모델을 배치할 수 있다.

[0047] 데이터베이스(340)는 보조 모델링 애플리케이션(300)에 의해 수신, 사용 및 생성된 데이터와 같은 보조 모델링 애플리케이션(300)과 연관된 데이터를 저장한다. 예를 들어, 데이터베이스(240)는 파이프라인의 표현, 자동 모델링 애플리케이션(200)으로부터의 데이터, GUI용 데이터, 사용자 입력, 모델 리파이닝과 연관된 데이터 등을 저장한다.

[0048] 도 4는 일 실시예에 따른 하이브리드 기계 학습을 위한 예시적인 사용자 인터페이스(400)이다. 사용자 인터페이스(400)는 앞서 설명한 바와 같이 보조 모델링 애플리케이션(300)에 의해 생성된 GUI의 실시예이다. 사용자 인터페이스(400)는 파이프라인의 표현을 제시한다. 이는 파이프라인의 4개 스텝과 4개의 탭(410A-D)(집합적으로 "탭(410)"이라 지칭됨)의 표현을 포함한다. 사용자 인터페이스(400)는 사용자가 4개의 탭(410) 중 하나를 클릭하여 대응하는 스텝을 선택하고 탐색할 수 있게 한다. 도 4에 예시된 바와 같이, 스텝 1이 사용자에 의해 선택된다. 스텝 1은 데이터 유형 설정에 대한 것이다. 스텝 1에 대한 탭(410A)의 사용자 클릭 수신에 응답하여, 사용자 인터페이스는 데이터 유형 목록(420)을 사용자에게 제시한다. 데이터 유형 목록(420)은 특징 목록(430), 특징에 대한 선택적 데이터 유형(440) 및 세부사항(450)을 포함한다. 사용자 인터페이스는 사용자가 탭을 클릭함으로써 특징을 선택할 수 있게 하도록 각각의 특징에 대한 탭(435)을 갖는다. 사용자 인터페이스는 또한 삼각형 버튼(445)과 각각의 특징의 선택적 데이터 유형에 대한 드롭 목록(447)을 가지고 있다. 드롭 목록(447)에 제시된 디폴트 데이터 유형은 자동 기계 학습 프로세스에서 사용되는 데이터 유형이며 "추천"으로 표시된다. 사용자가 삼각형 버튼(445)을 클릭하는 것에 응답하여, 사용자 인터페이스는 드롭 목록에 다른 선택적 데이터 유형을 제시한다. 도 3에 예시된 바와 같이, 사용자는 우편번호 특징을 선택하고 추천 데이터 유형은 숫자이다. 사용자가 삼각형 버튼(445)을 클릭한 후, 사용자 인터페이스는 사용자에게 세 가지 다른 데이터 유형을 제시하고 사용자가 세 가지 중 하나를 선택하여 추천 데이터 유형을 대체할 수 있게 한다. 도 4에서 우편번호 특징에 대한 추천 데이터 유형은 자동 기계 학습 프로세스에서 실수로 식별되었지만 사용자는 실수를 검토하고 수정할 수 있는 기회가 있다. 사용자는 드롭 목록(447)에서 우편번호를 선택하여 우편번호 특징의 데이터 유형을 우편번호로 변경할 수 있다.

[0049] 사용자 인터페이스는 또한 데이터 유형의 추가 세부사항(450)을 제시한다. 세부사항에는 확률을 포함하고, 이들 각각은 선택적 데이터 유형에 대응하며 특징의 데이터 유형이 대응 선택적 데이터 유형일 가능성이 얼마나 높은지를 나타낸다.

[0050] 도 4에는 도시되지 않았지만, 사용자 인터페이스(400)는 사용자가 다른 스텝으로 전환하거나 사용자 인터페이스에 다른 입력을 수행함에 따라 다른 정보를 제시한다. 예를 들어, 사용자의 탭(401C) 클릭을 수신한 후, 사용자 인터페이스는 선택적 특징 및/또는 변환기를 사용자에게 제시한다. 다른 예로서, 사용자 인터페이스는 사용자의 대치 방법 선택을 수신한 후 해당 대치 방법을 사용하여 계산된 대체 값을 사용자에게 제시하고 대체 값은 사용자가 다른 대치 방법을 선택함에 따라 변경된다. 또한, 사용자 인터페이스(400)는 다른 파이프라인 스텝의 표현을 제시할 수 있다.

[0051] 도 5는 일 실시예에 따른 하이브리드 기계 학습 모델 프로세스(500)를 예시하는 흐름도이다. 일부 실시예에서, 방법은 기계 학습 서버(110)에 의해 수행되지만, 방법의 동작 중 일부 또는 전부가 다른 실시예에서 다른 엔티티에 의해 수행될 수 있다. 일부 실시예에서, 흐름도의 동작은 상이한 순서로 수행되고 상이한 및/또는 추가 스텝을 포함한다.

[0052] 기계 학습 서버(110)는 데이터세트를 수신(510)한다. 일부 실시예에서, 기계 학습 서버(110)는 기업과 연관된 데이터 소스(예를 들어, 데이터 소스(120))로부터 데이터세트를 수신한다. 기업은 제조, 판매, 금융 및 은행업무와 같은 다양한 산업 중 하나 이상에 속할 수 있다. 일부 실시예에서, 기계 학습 서버(110)는 클라이언트 디바이스(130)와 같은 클라이언트 디바이스로부터 데이터세트를 수신한다.

[0053] 기계 학습 서버(110)는 수신된 데이터세트에 대해 자동 기계 학습 프로세스를 수행(520)하여 새로운 데이터에 기초하여 예측을 수행하는 모델을 훈련시킨다. 자동 기계 학습 프로세스에는 파이프라인에 기초하여 일련의 판단을 수행하는 동작이 포함된다. 일부 실시예에서, 파이프라인은 데이터 유형 설정, 데이터 인코딩, 데이터 대치, 특징 선택, 특징 순위화, 알고리즘 선택, 하이퍼파라미터 튜닝, 모델 훈련 및 모델 검증과 같은 일련의 스텝을 포함한다. 일부 다른 실시예에서, 파이프라인은 더 적거나, 더 많거나, 다른 스텝을 포함할 수 있다. 일련의 판단 중 각각의 판단은 파이프라인의 스텝에서 수행된다.

[0054] 훈련된 모델이 생성된 후, 기계 학습 서버(110)는 파이프라인의 표현을 생성한다(530). 기계 학습 서버(110)는 파이프라인의 일부 또는 모든 스텝을 식별할 수 있고 각각의 식별된 스텝의 표현을 생성할 수 있다. 스텝의 표현은 자동 기계 학습 프로세스 동안 스텝에서 수행된 판단을 포함하는 스텝에 대한 복수의 옵션을 포함한다. 스텝의 표현은 또한 사정, 설명 등과 같은 각각의 옵션의 다른 정보 또는 옵션에 대한 추천 수준을 나타내는 복수의 옵션 각각에 대한 순위 점수를 포함할 수 있다.

[0055] 일부 실시예에서, 기계 학습 서버(110)는 모델을 훈련시키기 위해 자동 기계 학습 프로세스에서 사용되는 특징, 특징에 대한 복수의 선택적 데이터 유형, 및 데이터 유형이 특징의 실제 데이터 유형일 확률을 나타내는 각각의 선택적 데이터 유형에 대한 순위 점수를 포함하는 데이터 유형 목록을 생성한다. 복수의 선택적 데이터 유형은 모델을 훈련시키기 위해 자동 기계 학습 프로세스의 특징에 대해 선택된 제1 데이터 유형을 포함한다. 기계 학습 서버(110)는 사용자가 예측에 얼마나 중요한지를 나타내는 복수의 특징 및 각각의 특징에 대한 설명을 포함하는 특징 목록을 생성할 수 있다. 복수의 특징은 모델을 훈련시키기 위해 자동 기계 학습 프로세스에서 사용되는 특징을 포함한다. 기계 학습 서버(110)는 모델 훈련을 위에 대한 해당 알고리즘의 선택 이유 또는 비선택 이유를 나타내는 복수의 알고리즘 및 각각의 알고리즘에 대한 설명을 포함하는 알고리즘 목록을 생성할 수 있다. 알고리즘은 모델을 훈련시키기 위해 자동 기계 학습 프로세스에 사용되는 알고리즘을 포함한다.

[0056] 기계 학습 서버(110)는 사용자 인터페이스에 디스플레이하기 위한 기계 학습 파이프라인의 표현을 제공한다(540). 사용자 인터페이스는 사용자에게 자동 기계 학습 프로세스에서 수행한 판단 중 적어도 일부를 수정할 수 있게 한다. 예를 들어, 사용자 인터페이스는 사용자가 특징에 대한 새로운 데이터 유형을 선택하고, 특징의 누락 값을 대체하기 위해 새로운 데이터 대치 방법을 선택하고, 자동 기계 학습 프로세스에서 수행되는 특징 선택을 편집(예를 들어, 순위 추가, 제거 또는 순위 변경)하고, 다른 알고리즘을 선택하고, 하이퍼파라미터를 조절하는 등을 수행할 수 있게 한다.

[0057] 기계 학습 서버(110)는 사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을 수신한다(550). 기계 학습 서버(110)는 사용자로부터의 하나 이상의 수정에 기초하여 모델을 리파이닝(560)한다. 리파이닝된 모델은 새로운 데이터에 기초하여 예측을 수행하는데 사용된다. 예를 들어, 기계 학습 서버(110)는 특징에 대한 상이한 데이터 유형의 선택을 수신하고 상이한 데이터 유형에 기초하여 특징의 값을 인코딩할 수 있다.

[0058] 도 6은 실시예에 따라 도 1의 기계 학습 서버(110)로서 사용하기 위한 전형적인 컴퓨터 시스템(600)의 기능도를 예시하는 고수준 블록도이다.

[0059] 예시된 컴퓨터 시스템은 칩셋(604)에 결합된 적어도 하나의 프로세서(602)를 포함한다. 프로세서(602)는 동일한 다이에 다수의 프로세서 코어를 포함할 수 있다. 칩셋(604)은 메모리 제어기 허브(620) 및 입력/출력(I/O) 제어기 허브(622)를 포함한다. 메모리(606) 및 그래픽 어댑터(612)는 메모리 제어기 허브(620)에 결합되고 디스플레이(618)는 그래픽 어댑터(612)에 결합된다. 저장 디바이스(608), 키보드(610), 포인팅 디바이스(614) 및 네트워크 어댑터(616)는 I/O 제어기 허브(622)에 결합될 수 있다. 일부 다른 실시예에서, 컴퓨터 시스템(600)은 추가적인, 더 적은 또는 상이한 컴포넌트를 가질 수 있고 컴포넌트는 상이하게 결합될 수 있다. 예를 들어, 컴퓨터 시스템(600)의 실시예에는 디스플레이 및/또는 키보드가 없을 수 있다. 또한, 컴퓨터 시스템(600)은 일부 실시예에서 랙 장착형 블레이드 서버 또는 클라우드 서버 인스턴스로 실체화될 수 있다.

[0060] 메모리(606)는 프로세서(602)에 의해 사용되는 명령 및 데이터를 보유한다. 일부 실시예에서, 메모리(606)

는 랜덤 액세스 메모리이다. 저장 디바이스(608)는 비일시적 컴퓨터 판독가능 저장 매체이다. 저장 디바이스(608)는 HDD, SSD 또는 다른 유형의 비일시적 컴퓨터 판독가능 저장 매체일 수 있다. 기계 학습 서버(110)에 의해 처리되고 분석된 데이터는 메모리(606) 및/또는 저장 디바이스(608)에 저장될 수 있다.

[0061] 포인팅 디바이스(614)는 마우스, 트랙볼 또는 다른 유형의 포인팅 디바이스일 수 있으며, 컴퓨터 시스템(600)에 데이터를 입력하기 위해 키보드(610)와 조합하여 사용된다. 그래픽 어댑터(612)는 디스플레이(618)에 이미지 및 다른 정보를 디스플레이한다. 일부 실시예에서, 디스플레이(618)는 사용자 입력 및 선택을 수신하기 위한 터치스크린 기능을 포함한다. 네트워크 어댑터(616)는 컴퓨터 시스템(600)을 네트워크(160)에 결합한다.

[0062] 컴퓨터 시스템(600)은 본 출원에 설명된 기능을 제공하기 위한 컴퓨터 모듈을 실행하도록 적응된다. 본 출원에 사용될 때, "모듈"이라는 용어는 특정 기능을 제공하기 위한 컴퓨터 프로그램 명령 및 기타 로직을 의미한다. 모듈은 하드웨어, 펌웨어 및/또는 소프트웨어로 구현될 수 있다. 모듈은 하나 이상의 프로세스를 포함하거나 및/또는 프로세스의 단지 일부에 의해서만 제공될 수 있다. 모듈은 전형적으로 저장 디바이스(608)에 저장되고 메모리(606)에 로딩되며 프로세서(602)에 의해 실행된다.

[0063] 컴포넌트의 특정 명명, 대문자 용어 표기, 속성, 데이터 구조, 또는 임의의 다른 프로그래밍 또는 구조적 양태는 의무적이거나 중요하지 않으며 설명된 실시예를 구현하는 메커니즘은 다른 이름, 형식 또는 프로토콜을 가질 수 있다. 또한, 시스템은 설명된 바와 같이 하드웨어와 소프트웨어의 조합을 통해 구현되거나 전체적으로 하드웨어 요소로 구현될 수 있다. 또한, 본 출원에 설명된 다양한 시스템 컴포넌트 사이의 특정 기능 분할은 단지 예시일 뿐이며 의무적인 것은 아니며; 단일 시스템 컴포넌트에 의해 수행되는 기능은 다수의 컴포넌트에 의해 대신 수행될 수 있고, 다수의 컴포넌트에 의해 수행되는 기능은 대신 단일 컴포넌트에 의해 수행될 수 있다.

[0064] 위의 설명 중 일부는 정보에 대한 동작의 알고리즘 및 상징적 표현 측면에서 특징을 제시한다. 이러한 알고리즘 설명 및 표현은 데이터 처리 분야의 숙련자가 그들의 연구의 내용을 다른 본 기술 분야의 숙련자에게 가장 효과적으로 전달하기 위해 사용하는 수단이다. 이러한 동작은 기능적으로 또는 논리적으로 설명되지만 컴퓨터 프로그램에 의해 구현되는 것으로 이해된다. 더욱이, 일반성을 잃지 않고 이러한 동작 배열을 모듈 또는 기능 이름으로 참조하는 것이 때때로 편리한 것으로 또한 입증되어 있다.

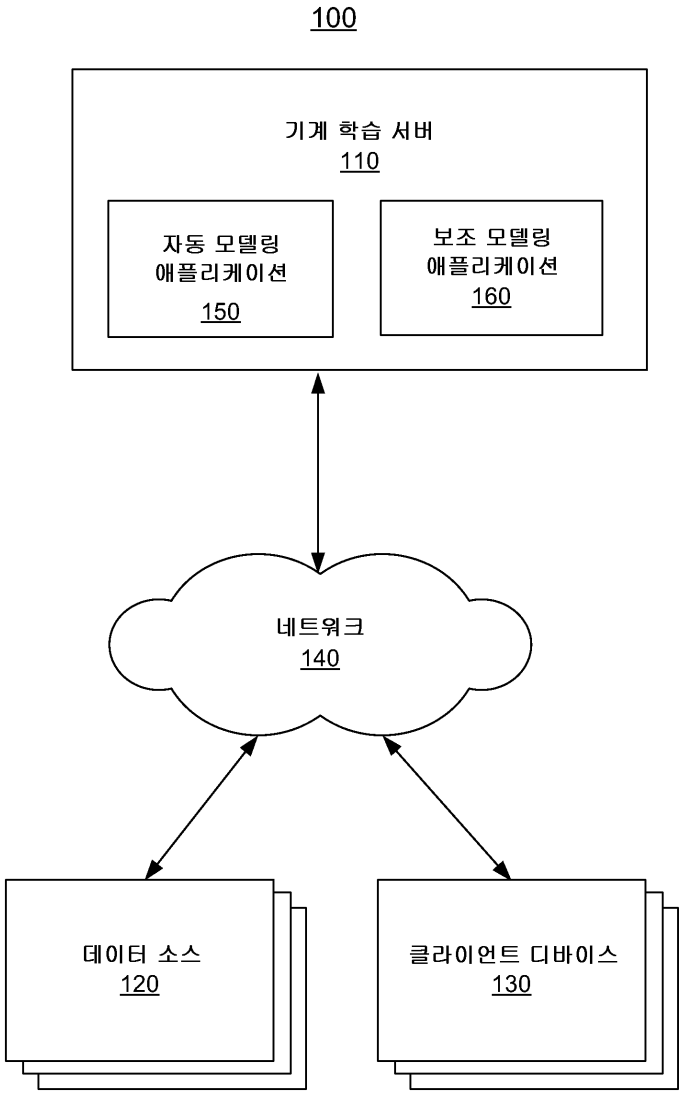
[0065] 위의 설명에서 명백히 달리 구체적으로 언급되지 않는 한, 설명 전체에서 "처리" 또는 "컴퓨팅" 또는 "계산" 또는 "결정" 또는 "디스플레이" 등과 같은 용어를 이용하는 설명은 컴퓨터 시스템 메모리 또는 레지스터 또는 다른 이러한 정보 저장, 송신 또는 디스플레이 디바이스 내에서 물리적(전자적) 양으로 표현되는 데이터를 조작하고 변환하는 컴퓨터 시스템 또는 유사한 전자 컴퓨팅 디바이스의 작용 및 프로세스를 지칭하는 것으로 이해된다.

[0066] 본 출원에 설명된 특정 실시예는 알고리즘의 형태로 설명된 프로세스 스텝 및 명령을 포함한다. 실시예의 프로세스 스텝 및 명령은 소프트웨어, 펌웨어 또는 하드웨어로 구현될 수 있고, 소프트웨어로 구현될 때 실시간 네트워크 운영 체제에 의해 사용되는 다른 플랫폼에 상주하고 동작되도록 다운로드될 수 있다는 점에 유의해야 한다.

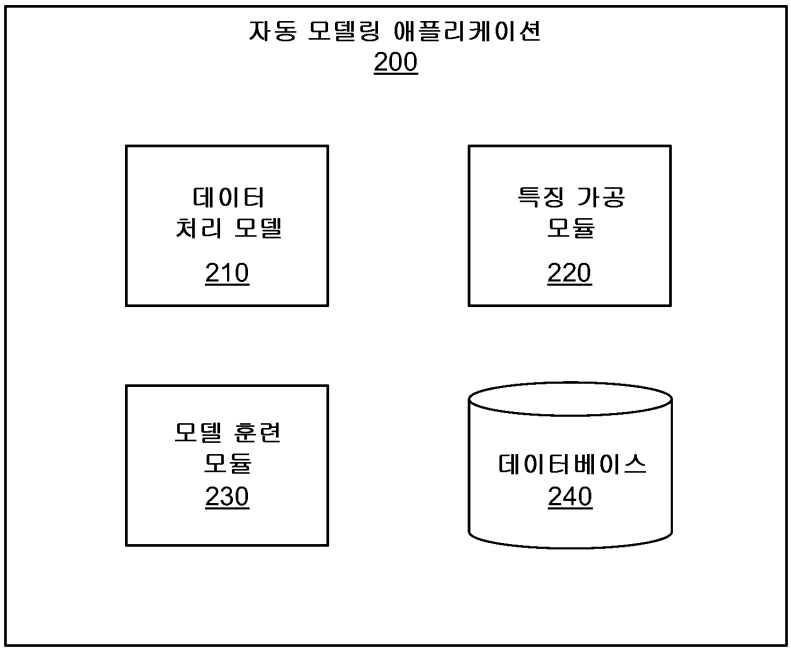
[0067] 마지막으로, 본 명세서에 사용된 언어는 주로 가독성 및 교육 목적을 위해 선택되었으며, 발명 주제를 설명하거나 제한하기 위해 선택된 것이 아닐 수 있음을 유의해야 한다. 따라서, 실시예의 개시는 예시를 위한 것이지 제한을 의도하지 않는다.

도면

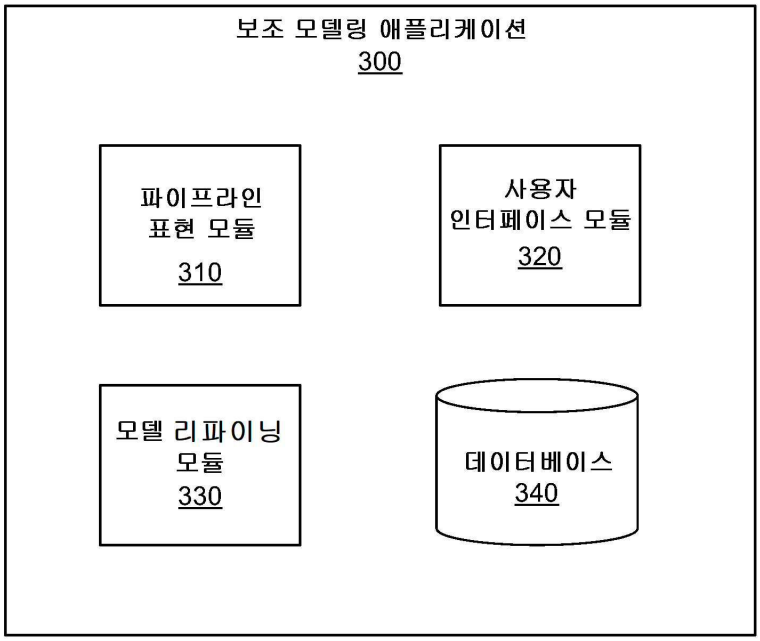
도면1


100
기계 학습 서버
110
자동 모델링
애플리케이션
150
보조 모델링
애플리케이션
160
네트워크
140
데이터 소스
120
클라이언트 디바이스
130

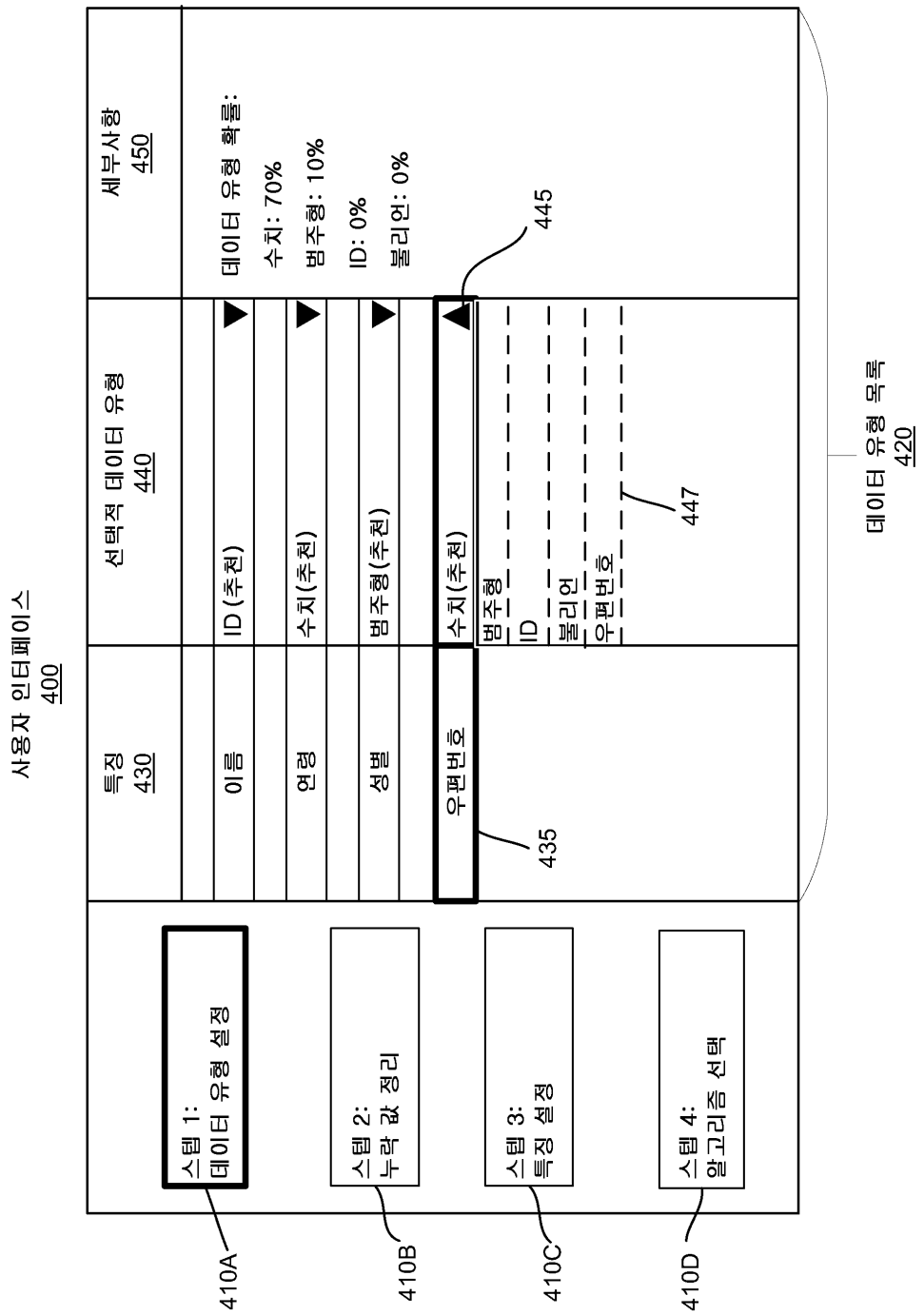
도면2


자동 모델링 애플리케이션
200
데이터
처리 모델
210
특징 가공
모듈
220
모델 훈련
모듈
230
데이터베이스
240

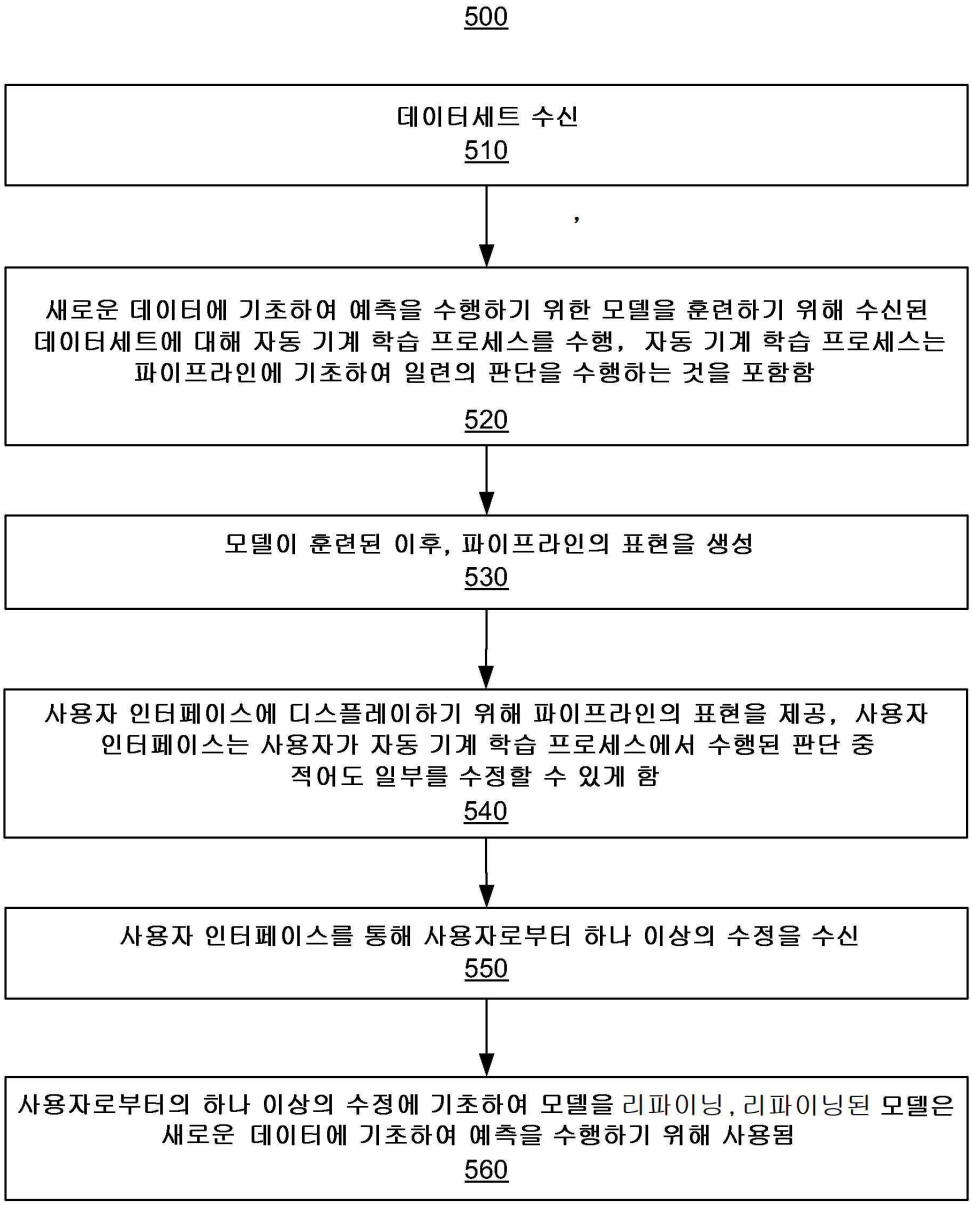
도면3


보조 모델링 애플리케이션
300
파이프라인
표현 모듈
310
사용자
인터페이스 모듈
320
모델 리파이닝
모듈
330
데이터베이스
340

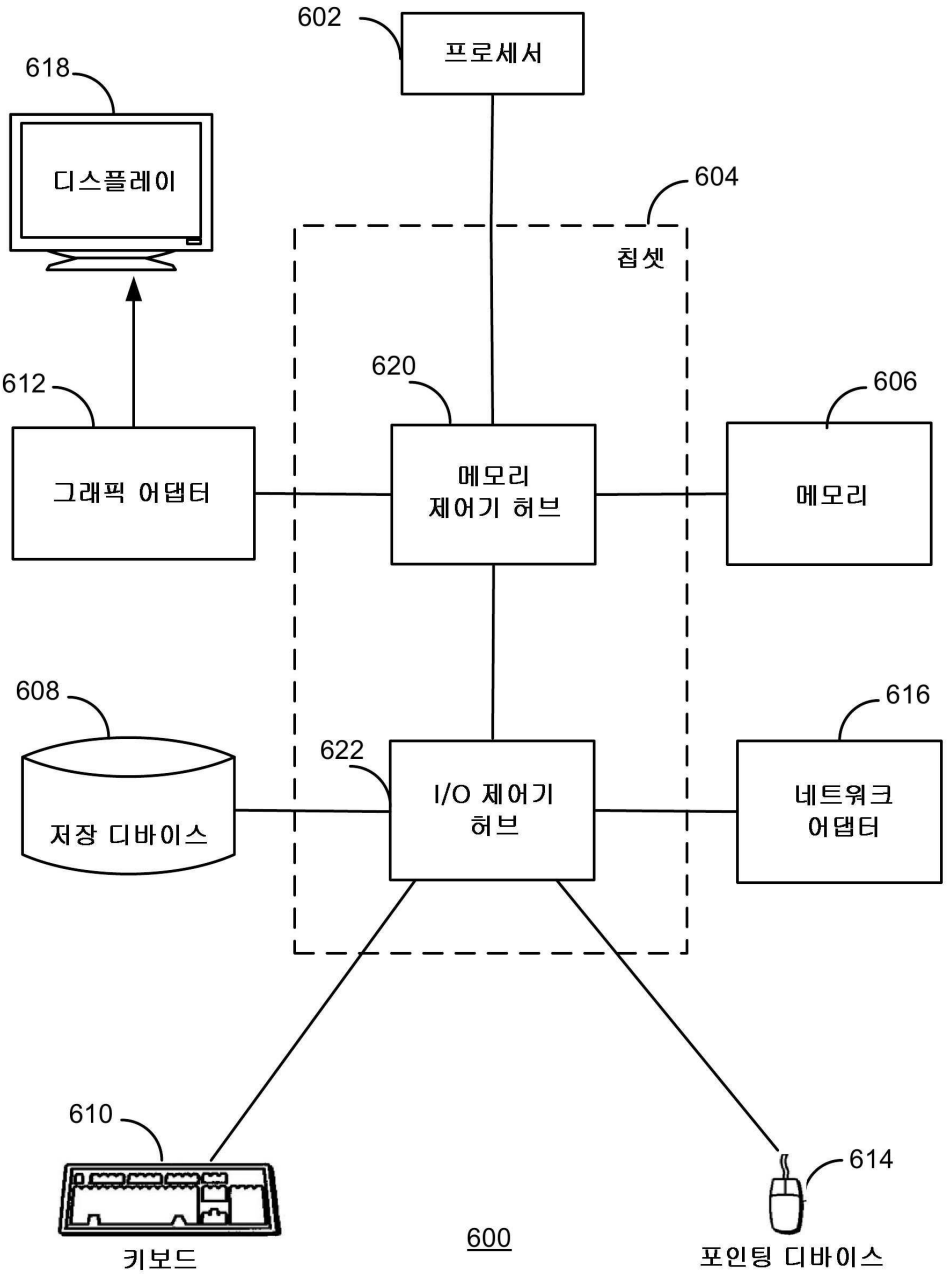
도면4


사용자 인터페이스
400
스텝 1:
데이터 유형 설정
410A
스텝 2:
누락 값 정리
410B
스텝 3:
특징 설정
410C
스텝 4:
알고리즘 선택
410D
특징
430
이름
연령
성별
우편번호
435
선택적 데이터 유형
440
ID (추천)
수치(추천)
범주형(추천)
수치(추천)
범주형
ID
불리언
우편번호
445
447
세부사항
450
데이터 유형 확률:
수치: 70%
범주형: 10%
ID: 0%
불리언: 0%
데이터 유형 목록
420

도면5


500
데이터세트 수신
510
새로운 데이터에 기초하여 예측을 수행하기 위한 모델을 훈련하기 위해 수신된 데이터세트에 대해 자동 기계 학습 프로세스를 수행, 자동 기계 학습 프로세스는 파이프라인에 기초하여 일련의 판단을 수행하는 것을 포함함
520
모델이 훈련된 이후, 파이프라인의 표현을 생성
530
사용자 인터페이스에 디스플레이하기 위해 파이프라인의 표현을 제공, 사용자 인터페이스는 사용자가 자동 기계 학습 프로세스에서 수행된 판단 중 적어도 일부를 수정할 수 있게 함
540
사용자 인터페이스를 통해 사용자로부터 하나 이상의 수정을 수신
550
사용자로부터의 하나 이상의 수정에 기초하여 모델을 리파이닝, 리파이닝된 모델은 새로운 데이터에 기초하여 예측을 수행하기 위해 사용됨
560

도면6


602
프로세서
618
디스플레이
604
칩셋
612
620
606
그래픽 어댑터
메모리
제어기 허브
메모리
608
616
622
저장 디바이스
I/O 제어기
허브
네트워크
어댑터
610
614
600
키보드
포인팅 디바이스