(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0095231
(43) 공개일자 2023년06월29일

(51) 국제특허분류(Int. Cl.)
G06F 16/242 (2019.01) *G06F 16/23* (2019.01)
G06F 16/2455 (2019.01) *G06F 16/25* (2019.01)
G06F 17/40 (2006.01) *G06F 9/451* (2018.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 16/2445 (2019.01)
G06F 16/2365 (2019.01)
(21) 출원번호 10-2021-0184499
(22) 출원일자 2021년12월22일
심사청구일자 2021년12월22일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
장윤
서울특별시 동대문구 정릉천동로 16, 103동 401호 (용두동, 용두두산위브)
홍혜인
서울특별시 중랑구 상봉로15마길 22, 203호 (면목동, 대성하이츠)
유상봉
서울특별시 광진구 군자로13길 18-5, 3층 (군자동)
(74) 대리인
특허법인엠에이피에스

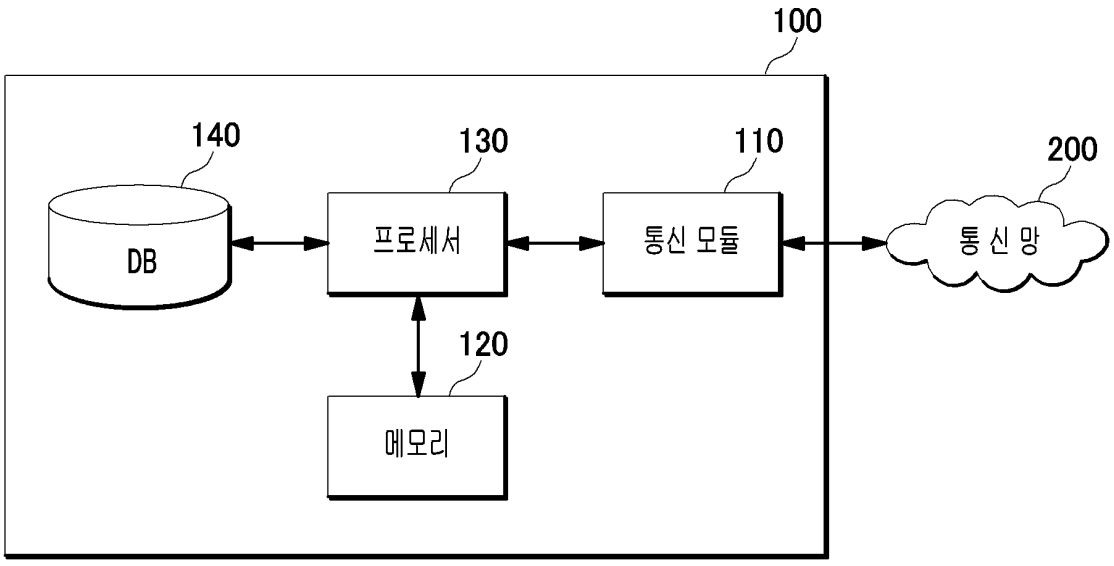
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **데이터 프로파일링 장치 및 방법**

(57) 요 약

본 발명의 일 측면에 따른 데이터 프로파일링 장치는 데이터 프로파일링 프로그램이 저장된 메모리; 상기 데이터 프로파일링 프로그램을 실행하는 프로세서를 포함하되, 상기 데이터 프로파일링 프로그램은 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신하고, 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하고, 상기 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하고, 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/2448 (2019.01)
G06F 16/24554 (2019.01)
G06F 16/24565 (2019.01)
G06F 16/252 (2019.01)
G06F 17/40 (2020.08)
G06F 9/451 (2018.02)
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711125889
과제번호 2019-0-00374-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발(R&D,정보화)
연구과제명 빅데이터 및 AI기반 에너지신산업형 분산자원(신재생에너지발전단지) 중개거래 플랫폼 개발
기 여 율 1/2
과제수행기관명 (주)엘시스
연구기간 2021.01.01 ~ 2021.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711126466
과제번호 2019-0-00242-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 SW컴퓨팅산업원천기술개발(R&D,정보화)
연구과제명 빅데이터 신뢰성과 활용성 극대화를 위한 빅데이터 증강분석 프로파일링 플랫폼 개발
기 여 율 1/2
과제수행기관명 (주)피씨엔
연구기간 2021.01.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

데이터 프로파일링 장치에 있어서,

데이터 프로파일링 프로그램이 저장된 메모리;

상기 데이터 프로파일링 프로그램을 실행하는 프로세서를 포함하되,

상기 데이터 프로파일링 프로그램은 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신하고, 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하고, 상기 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하고, 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력하는 것인, 데이터 프로파일링 장치.

청구항 2

제1항에 있어서,

상기 복수의 프로파일링 전략은 문제 데이터 감소 정도에 기반한 프로파일링 전략, 데이터 품질 향상 정도 기반 프로파일링 전략 또는 전문가 레시피 기반 프로파일링 전략을 포함하는 것인, 데이터 프로파일링 장치.

청구항 3

제1항에 있어서,

상기 데이터 프로파일링 프로그램은, 계층 구조를 가지는 프로비넌스(Provenance) 그래프를 통해 대상 데이터에 대한 선택 입력을 수신하는 제 1사용자 인터페이스를 표시하되, 상기 프로비넌스 그래프는 입력 데이터 노드와 입력 데이터 노드에 대해 상기 각 프로파일링 행동이 적용된 결과 데이터를 나타내는 복수의 자식 노드 를 포함하는 것인, 데이터 프로파일링 장치.

청구항 4

제3항에 있어서,

상기 데이터 프로파일링 프로그램은, 상기 제 1사용자 인터페이스에 인접하여 표시되는 제 2 사용자 인터페이스를 통해 상기 복수의 프로파일링 행동을 표시하고, 상기 데이터 프로파일링 예상 결과는 상기 제2 사용자 인터페이스를 통해 표시하는 것인, 데이터 프로파일링 장치.

청구항 5

제1항에 있어서,

상기 프로파일링 전략으로서 문제 데이터 감소 정도에 기반한 프로파일링 전략 이 선택된 경우, 복수의 프로파일링 행동 중 문제 데이터의 감소 개수가 가장 큰 순서에 따라 미리 설정된 개수의 프로파일링 행동을 표시하되,

상기 복수의 프로파일링 행동은 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중

앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM(expectation maximization) 알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동을 포함하는 것인, 데이터 프로파일링 장치.

청구항 6

제1항에 있어서,

상기 프로파일링 전략으로서 데이터 품질 향상 정도 기반 프로파일링 전략이 선택된 경우, 복수의 프로파일링 행동 중 데이터 품질 향상 정도가 가장 큰 순서에 따라 미리 설정된 개수의 프로파일링 행동을 표시하되,

상기 복수의 프로파일링 행동은 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM(expectation maximization) 알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동을 포함하는 것인, 데이터 프로파일링 장치.

청구항 7

제1항에 있어서,

상기 프로파일링 전략으로서 전문가 레시피 기반 프로파일링 전략이 선택된 경우, 전문가가 수행한 데이터 프로파일링 행동을 라벨링하여 학습한 복수의 학습 모델을 각각의 프로파일링 행동으로서 표시하는 것인, 데이터 프로파일링 장치.

청구항 8

데이터 프로파일링 방법에 있어서,

(a) 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신하는 단계;

(b) 상기 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하는 단계;

(c) 상기 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하는 단계; 및

(d) 상기 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력하는 단계를 포함하는, 데이터 프로파일링 방법.

청구항 9

제8항에 있어서,

상기 복수의 프로파일링 전략은 문제 데이터 감소 정도에 기반한 프로파일링 전략, 데이터 품질 향상 정도 기반 프로파일링 전략 또는 전문가 레시피 기반 프로파일링 전략을 포함하는 것인, 데이터 프로파일링 방법.

청구항 10

제8항에 있어서,

상기 (b) 단계는 계층 구조를 가지는 프로비넌스(Provenance) 그래프를 통해 대상 데이터에 대한 선택 입력을 수신하는 제 1사용자 인터페이스를 표시하되,

상기 프로비넌스 그래프는 입력 데이터 노드와 입력 데이터 노드에 대해 상기 각 프로파일링 행동이 적용된 결과 데이터를 나타내는 복수의 자식 노드 를 포함하는 것인, 데이터 프로파일링 방법.

청구항 11

제10항에 있어서,

상기 (c) 단계는 상기 제 1사용자 인터페이스에 인접하여 표시되는 제 2 사용자 인터페이스를 통해 상기 복수의 프로파일링 행동을 표시하는 단계를 포함하고,

상기 (d) 단계는 상기 제2 사용자 인터페이스를 통해, 각 프로파일링 행동에 의한 프로파일링 예상 결과를 표시하는 것인, 데이터 프로파일링 방법.

청구항 12

제8항에 있어서,

상기 프로파일링 전략으로서 문제 데이터 감소 정도에 기반한 프로파일링 전략 이 선택된 경우, 상기 (c) 단계는 복수의 프로파일링 행동 중 문제 데이터의 감소 개수가 가장 큰 순서에 따라 미리 설정된 개수의 프로파일링 행동을 표시하되,

상기 복수의 프로파일링 행동은 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM(expectation maximization) 알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동을 포함하는 것인, 데이터 프로파일링 방법.

청구항 13

제8항에 있어서,

상기 프로파일링 전략으로서 데이터 품질 향상 정도 기반 프로파일링 전략이 선택된 경우, 상기 (c) 단계는 복수의 프로파일링 행동 중 데이터 품질 향상 정도가 가장 큰 순서에 따라 미리 설정된 개수의 프로파일링 행동을 표시하되,

상기 복수의 프로파일링 행동은 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM(expectation maximization) 알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동을 포함하는 것인, 데이터 프로파일링 방법.

청구항 14

제8항에 있어서,

상기 프로파일링 전략으로서 전문가 레시피 기반 프로파일링 전략이 선택된 경우, 상기 (c) 단계는 전문가가 수행한 데이터 프로파일링 행동을 라벨링하여 학습한 복수의 학습 모델을 각각의 프로파일링 행동으로서 표시하는 것인, 데이터 프로파일링 방법.

청구항 15

제 8 항 내지 제 14 항 중 어느 한 항에 따르는 데이터 프로파일링 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 비일시적 컴퓨터 판독가능 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공 신경망 학습 모델의 구축에 사용되는 각종 데이터를 대상으로 데이터 프로파일링을 수행하는 데이터 프로파일링 장치 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 인공 신경망 기술이나 빅데이터 처리 기술이 다양한 산업 분야에 확산됨에 따라, 인공 신경망 학습 모델의 구축 과정에 사용되는 데이터의 품질에 대한 관심이 높아지고 있다.

[0003] 초기에는 다량의 데이터를 수집하는데 관심이 집중되었다면, 최근에는 데이터의 품질을 높은 수준으로 유지하는 기술에 대한 관심이 증가하고 있으며, 이러한 데이터 품질 관리를 위한 방법으로서 데이터 프로파일링에 대한 연구가 진행되고 있다. 데이터 프로파일링은 데이터를 수집하는 과정뿐만 아니라, 수집된 데이터의 오류를 찾아내고 이를 최소화하는 과정을 포함한다. 보다 구체적으로는 수집된 데이터에 대하여 통계적 분석을 수행하고 데이터 품질을 확인하는 과정을 수행한다. 이를 위해, 종래의 기술에 따르면 메타데이터의 수집 및 분석 과정과 데이터 현상을 분석하는 과정을 수행하고 있다. 메타데이터는 운영 중인 데이터베이스의 테이블 명, 컬럼 명, 제약조건 등의 정보를 포함하는 것으로서, 메타데이터 분석을 통해 테이블 및 컬럼에 대한 정보와 데이터 관리 문서를 매핑하여 불일치 사항을 분석한다. 일 예로서, 데이터 프로파일링 수행을 위하여 컬럼 속성 분석을 통해 일차적인 품질 현황을 파악하는데, 컬럼의 총 건수, 유일값 수, NULL값 수, 공백값 수, 최댓값, 최솟값, 최대 빈도, 최소빈도 등 기초 집계값을 산출하여 값이 유효한 범위 내에 있는지 판단한다.

[0004] 이러한 종래의 데이터 프로파일링 기술은 사용자의 데이터 사용 목적과는 관계없이 획일적인 규칙에 따라 고정된 데이터 프로파일링 보고서를 생성한다는 점에서 단점이 있다. 또한, 데이터 프로파일링을 위해서는 작업자의 전문적인 기술을 요하고 있다는 점에서도 개선이 필요한 상태이다. 나아가, 각각의 데이터 프로파일링 프로세서를 진행하는 과정에서 데이터의 품질이 어느 정도로 향상되고 있는지를 파악하기 어려운 점도 있다.

[0005] 본 발명은 이러한 종래 기술의 문제점을 해소하기 위한 것으로, 사용자 편의성을 향상시킨 새로운 데이터 프로파일링 기술을 제공하고자 한다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 문제점을 해결하기 위하여, 데이터 프로파일링을 위한 다양한 방법을 편리하게 적용할 수 있는 데이터 프로파일링 장치 및 방법을 제공하고자 한다.

[0007] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0008] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 데이터 프로파일링 장치는, 데이터 프로파일링 프로그램이 저장된 메모리; 상기 데이터 프로파일링 프로그램을 실행하는 프로세서를 포함하되, 상기 데이터 프로파일링 프로그램은 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신하고, 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하고, 상기 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하고, 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력하는 것이다.

[0009] 본 발명의 다른 측면에 따른 데이터 프로파일링 방법은, 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일

링 전략에 대한 사용자의 선택 입력을 수신하는 단계; 상기 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하는 단계; 상기 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하는 단계; 및 상기 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력하는 단계를 포함한다.

발명의 효과

[0010] 본 발명은 전술한 종래 기술의 문제점을 해결하기 위한 것으로서, 전문적인 기술이 없는 작업자도 편리하게 데이터 프로파일링을 수행할 수 있도록 한다. 특히, 여러 프로파일링 전략과 그에 따른 복수의 프로파일링 행동을 편리하게 선택할 수 있도록 하여, 작업자가 자신의 의도에 맞는 데이터 프로파일링을 수행할 수 있도록 한다.

[0011] 또한, 각각의 프로파일링 행동의 선택에 도움이 될 수 있도록, 프로파일링 행동에 대한 예상 결과를 출력함으로써, 작업자가 정확하고 신속한 의사 결정을 할 수 있도록 한다.

도면의 간단한 설명

[0012] 도 1은 본 발명의 일 실시예에 따른 데이터 프로파일링 장치의 구성을 도시한 블록도이다.

도 2는 본 발명의 일 실시예에 데이터 프로파일링 방법을 도시한 순서도이다.

도 3은 본 발명의 일 실시예에 따른 데이터 프로파일링 장치에서 프로파일링 전략을 선택하는 과정을 도시한 것이다.

도 4 및 도 5는 본 발명의 일 실시예에 따른 데이터 프로파일링 장치에서 프로파일링 행동을 선택하는 과정을 도시한 것이다.

도 6은 본 발명의 일실 시예에 따른 전문가 레시피 기반의 프로파일링 행동을 설명하기 위한 도면이다.

도 7은 본 발명의 일 실시예에 따른 데이터 프로파일링 결과의 시각화 기능을 도시한 도면이다.

도 8은 본 발명의 일 실시예에 따른 시각화 그래프의 예를 도시한 것이다.

발명을 실시하기 위한 구체적인 내용

[0013] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0014] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "전기적으로 연결"되어 있는 경우도 포함한다. 또한 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0015] 본 명세서에 있어서 '부(部)'란, 하드웨어에 의해 실현되는 유닛(unit), 소프트웨어에 의해 실현되는 유닛, 양방을 이용하여 실현되는 유닛을 포함한다. 또한, 1 개의 유닛이 2 개 이상의 하드웨어를 이용하여 실현되어도 되고, 2 개 이상의 유닛이 1 개의 하드웨어에 의해 실현되어도 된다. 한편, '~부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니며, '~부'는 어드레싱 할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '~부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들 및 변수들을 포함한다. 구성요소들과 '~부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '~부'들로 결합되거나 추가적인 구성요소들과 '~부'들로 더 분리될 수 있다. 뿐만 아니라, 구성요소들 및 '~부'들은 디바이스 또는 보안 멀티미디어카드 내의 하나 또는 그 이상의 CPU들을 재생시키도록 구현될 수도 있다.

[0016] 이하에서 언급되는 데이터 프로파일링 장치는 네트워크를 통해 서버나 타 단말에 접속할 수 있는 컴퓨터나 휴대용 단말기로 구현될 수 있다. 여기서, 컴퓨터는 예를 들어, 웹 브라우저(WEB Browser)가 탑재된 노트북, 데스크

톱(desktop), 랩톱(laptop) 등을 포함하고, 휴대용 단말기는 예를 들어, 휴대성과 이동성이 보장되는 무선 통신 장치로서, 각종 이동 통신 기반 단말, 스마트폰, 태블릿 PC 등과 같은 모든 종류의 핸드헬드(Handheld) 기반의 무선 통신 장치를 포함할 수 있다. 또한, "네트워크"는 근거리 통신망(Local Area Network; LAN), 광역 통신망(Wide Area Network; WAN) 또는 부가가치 통신망(Value Added Network; VAN) 등과 같은 유선 네트워크나 이동 통신망(mobile radio communication network) 또는 위성 통신망 등과 같은 모든 종류의 무선 네트워크로 구현될 수 있다.

[0017] 도 1은 본 발명의 일 실시예에 따른 데이터 프로파일링 장치의 구성을 도시한 블록도이다.

[0018] 데이터 프로파일링 장치(100)는 메모리(110), 통신모듈(120), 프로세서(130) 및 디스플레이(140)를 포함한다.

[0019] 메모리(110)는 데이터 프로파일링 프로그램이 저장된다. 또한, 메모리(110)에는 데이터 프로파일링 장치(100)의 구동을 위한 운영 체제나 데이터 프로파일링 프로그램의 실행 과정에서 발생되는 여러 종류가 데이터가 저장된다. 데이터 프로파일링 프로그램은 입력 데이터에 대하여 적용될 수 있는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신하는 단계; 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신하는 단계; 선택된 대상 데이터에 대하여 상기 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시하는 단계; 및 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력하는 단계를 각각 수행하도록 구성된다. 데이터 프로파일링 프로그램에 의해 구현되는 데이터 프로파일링 과정에 대해서는 추후 상세히 설명하기로 한다.

[0020] 이때, 메모리(110)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다.

[0021] 또한, 메모리(110)는 프로세서(130)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 여기서, 메모리(110)는 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치 외에 자기 저장 매체(magnetic storage media) 또는 플래시 저장 매체(flash storage media)를 포함할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0022] 통신모듈(120)은 데이터 프로파일링 장치(100)의 결과를 외부 컴퓨팅 장치에 전송하거나, 외부 컴퓨팅 장치로부터 프로파일링 대상이 되는 여러 종류의 데이터를 수신하게 하는 하나 이상의 구성요소를 포함할 수 있다. 예를 들어, 통신모듈(120)은 다른 네트워크 장치와 유무선 연결을 통해 제어 신호 또는 데이터 신호와 같은 신호를 송수신하기 위해 필요한 하드웨어 및 소프트웨어를 포함하는 장치일 수 있다.

[0023] 프로세서(130)는 메모리(110)에 저장된 프로그램을 실행하되, 데이터 프로파일링 프로그램의 실행에 따르는 전체 과정을 제어한다. 프로세서(130)가 수행하는 각각의 동작에 대해서는 추후 보다 상세히 살펴보기로 한다.

[0024] 이러한 프로세서(130)는 데이터를 처리할 수 있는 모든 종류의 장치를 포함할 수 있다. 예를 들어 프로그램 내에 포함된 코드 또는 명령으로 표현된 기능을 수행하기 위해 물리적으로 구조화된 회로를 갖는, 하드웨어에 내장된 데이터 처리 장치를 의미할 수 있다. 이와 같이 하드웨어에 내장된 데이터 처리 장치의 일 예로써, 마이크로프로세서(microprocessor), 중앙처리장치(central processing unit: CPU), 프로세서 코어(processor core), 멀티프로세서(multiprocessor), ASIC(application-specific integrated circuit), FPGA(field programmable gate array) 등의 처리 장치를 망라할 수 있으나, 본 발명의 범위가 이에 한정되는 것은 아니다.

[0025] 데이터베이스(140)는 데이터 프로파일링의 대상이 되는 입력 데이터가 저장되거나, 데이터 프로파일링 프로그램의 실행과정에서 생성되는 여러 데이터가 저장될 수 있다.

[0026] 도 2는 본 발명의 일 실시예에 데이터 프로파일링 방법을 도시한 순서도이고, 도 3 내지 도 7은 본 발명의 일 실시예에 따른 데이터 프로파일링 과정을 설명하기 위한 참고도면이다.

[0027] 먼저, 데이터 프로파일링의 대상이 되는 입력 데이터에 대하여, 본 발명에서 적용하고자 하는 복수의 프로파일링 전략에 대한 사용자의 선택 입력을 수신한다(S210).

[0028] 데이터 프로파일링 장치(100)가 제공하는 프로파일링 전략은 문제 데이터 감소 정도에 기반한 프로파일링 전략, 데이터 품질 향상 정도 기반 프로파일링 전략 또는 전문가 레시피 기반 프로파일링 전략을 포함한다.

[0029] 도 3은 본 발명의 일 실시예에 따른 데이터 프로파일링 장치에서 프로파일링 전략을 선택하는 과정을 도시한 것이다.

[0030] 사용자 인터페이스(310)를 통해 3가지 프로파일링 전략이 표시되는데, 문제 데이터 감소 정도에 기반한 프로파

일링 전략(Based on the degree of problem reduction), 데이터 품질 향상 정도 기반 프로파일링 전략(Based on the degree of quality improvement) 또는 전문가 레시피 기반 프로파일링 전략(Based on expert recipes)이 각각 표시되고, 사용자는 희망하는 프로파일링 전략을 선택할 수 있다.

[0031] 각각의 데이터 프로파일링 전략은 데이터의 품질에 부정적인 영향을 주는 문제 데이터를 식별하고, 이를 감소시키는 구체적인 행동을 나타내는 프로파일링 행동이 실행되도록 한다. 이때, 문제 데이터로는 불완전한 데이터 수집, 수정 등 다양한 이유로 인한 누락된 값을 의미하는 결측값, 다른 대부분의 값에 비해 크게 벗어난 값을 의미하는 극단값, 값이 전, 후 값 사이에 존재하지 않는 값을 의미하는 비 일관값등이 있을 수 있다.

[0032] 또한, 이와 같은 데이터 프로파일링을 실행한 후의 데이터의 품질을 평가하기 위해 데이터 갖는 정규 분포 특성을 이용한다. 데이터의 품질이 높을수록 정규분포 특성을 갖는 다는 가정에 따라, KS(kolomogorov smironov) 테스트, 왜도, 첨도, 정보 엔트로피(information entropy)등의 계산 방법을 사용할 수 있다.

[0033] KS(Kolomogorov smironov) 테스트는 정규 분포와 데이터 분포의 경험적 분포 함수 차이를 의미하며 값이 작을수록 데이터의 품질이 높음을 의미한다. 왜도/첨도는 정규 분포와 비교했을 때 데이터 분포의 대칭 정도/뾰족한 정도를 의미하며 값이 0에 가까울수록 데이터 품질이 높음을 의미한다. 정보 엔트로피는 정보량을 의미하며 값이 작을수록 데이터 품질이높음을 의미한다.

[0034] 다시 도 2를 참조하면, 프로파일링 전략이 적용될 대상 데이터에 대한 사용자의 선택 입력을 수신한다(S220).

[0035] 본 발명에서는 도 3에 도시된 바와 같이, 사용자 인터페이스(320)를 통해 계층 구조를 가지는 프로비넌스(Provenance) 그래프를 표시하고, 이를 이용하여 프로파일링 대상이 되는 대상 데이터를 선택하는 사용자의 선택 입력을 손쉽게 수신할 수 있도록 한다.

[0036] 예를 들어, 원본 데이터는 최상위 노드(1)에 배치시킨 상태에서, 사용자 인터페이스(310)를 통해 선택된 프로파일링 전략에 의해 실행되는 프로파일링 행동별로 그 적용 결과가 반영된 데이터를 복수의 자식 노드(2, 3, 4)의 형태로 표시할 수 있다. 그리고, 사용자는 특정 자식 노드를 선택하면, 해당 노드에 대응되는 프로파일링 행동이 적용된 결과를 확인할 수 있다. 또한, 해당 자식 노드에 대하여 추가적으로 프로파일링 행동이 적용될 수 있으며, 그 프로파일링 행동별로 적용 결과가 반영된 또 다른 자식 노드(5, 6, 7)의 형태로 각각의 데이터 프로파일링 결과가 표시될 수 있다.

[0037] 다음으로, 앞선 단계(S220)에서 선택된 대상 데이터에 대하여 앞선 단계(S210)에서 선택된 프로파일링 전략에 의해 실행될 수 있는 복수의 프로파일링 행동을 표시한다(S230).

[0038] 앞서 설명한 각각의 프로파일링 전략에는 각 전략을 구현하는 복수의 프로파일링 행동이 실행될 수 있다.

[0039] 데이터 프로파일링 장치(100)는 문제 데이터 감소 정도에 기반한 프로파일링 전략이 선택된 경우, 결측 값이 있는 행(column)을 삭제하는 프로파일링 행동이나, 결측값을 다른 값으로 대체하는 프로파일링 행동이 표시되도록 한다. 보다 구체적으로는, 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값(LOCF(last observation carried forward))으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM(expectation maximization) 알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동을 실행하도록 한다. EM은 관측되지 않은 잠재변수에 의존하는 확률 모델에서 최대가능도나 최대사후확률을 갖는 모수의 추정값을 추정하여 결측값으로 대체하는 방법이다. 이때, 문제 데이터 감소 정도에 기반한 프로파일링 전략이 선택된 경우에는 복수의 프로파일링 행동 중 문제 데이터 감소 정도가 가장 높은 순서대로 프로파일링 행동을 추천한다. 바람직하게는, 문제 데이터 감소 정도가 가장 높은 3개의 프로파일링 행동을 추천한다. 이는, 대상 데이터에 따라 문제 데이터 감소 정도가 상이해지므로, 각 케이스에 따라 문제 데이터 감소 개수를 확인 후 그 개수가 높은 순서에 따라 프로파일링 행동을 추천하게 된다.

[0040] 도 4를 참조하면, 사용자 인터페이스(320)와 인접하여 표시되는 사용자 인터페이스(330)을 통해 복수의 프로파일링 행동이 표시될 수 있다.

[0041] 도시된 바와 같이, 노드(4)에 대응하는 데이터는 원본 데이터(1)에 대하여 사용자에 의해 선택된 프로파일링 행동에 의해 프로파일링이 실행된 데이터를 나타낸다. 그리고, 노드 (4)에 대응하는 데이터가 사용자에 의해 추가로 선택되면, 도시된 바와 같이, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동(331), 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동(332), 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행

동(333)이 각각 표시되고, 그중 어느 한 프로파일링 행동을 사용자가 선택할 수 있도록 한다. 이때, 노드(4)에 대응하는 데이터의 품질에 대해서는 막대 그래프를 이용하여 도시하고 있다. 즉, 데이터의 컬럼별(습도, 온도, pm25, pm10)로 결측(missing) 데이터나, 극단(extreme)의 개수를 막대 그래프로 표시하여, 한눈에 파악할 수 있도록 한다.

[0042] 다음으로, 복수의 프로파일링 행동에 의한 데이터 프로파일링 예상 결과를 출력한다(S240).

[0043] 도 5에 도시된 바와 같이, 사용자가 어느 한 프로파일링 행동을 선택한 경우, 사용자 인터페이스(330)를 통해 각 프로파일링 행동에 의한 프로파일링 예상 결과를 표시할 수 있다. 특히, 각 프로파일링 행동에 의해, 데이터의 품질이 향상되는 것으로 판단한 근거를 표시하도록 한다. 구체적으로 살펴보면, 프로파일링 예상 결과를 표시하는 화면(340)에서 제 1 그래프(341)와 제 3 그래프(343)는 프로파일링 행동을 수행하기 전의 데이터를 기초로 생성된 것이고, 제 2 그래프(342)와 제 4 그래프(344)는 프로파일링 행동을 수행한 후의 데이터를 기초로 생성된 것으로서, 각 그래프의 비교를 통해 데이터 품질의 향상 정도를 손쉽게 확인할 수 있도록 한다. 예를 들어, 히트 맵을 통해 결측 데이터의 개수를 표시하는 제 1 그래프(341)와 제 2 그래프(342)를 보면, 히트 맵에서는 결측값의 개수가 많을수록 색이 진해진다. 따라서, 제 1 그래프(341)는 pm10 컬럼의 초기에 결측 값이 많은 것을 확인할 수 있는데, 결측값을 평균값으로 대체하는 프로파일링 행동에 의하면, 제 2 그래프(342)에서와 같이, pm10 컬럼에서 결측값이 모두 사라진 것을 확인할 수 있다.

[0044] 또한, 제 3 그래프(343)와 제 4 그래프(344)는 극단값 분포를 나타내는 히스토그램을 도시한 것이다. 각 그래프에서 빈도수가 48이상인 값은 극단값에 해당하는데, 결측값을 평균값으로 대체하는 프로파일링 행동에 의하면 제 3 그래프(343)와는 달리 제 4 그래프(344)에서는 극단값의 개수가 감소함을 확인할 수 있다. 이와 같이, 각 프로파일링 행동의 선택시 발생되는 프로파일링 예상 결과를 미리 표시함으로써, 사용자가 적절한 프로파일링 행동을 선택할 수 있도록 도움을 줄 수 있다.

[0045] 한편, 앞서 살펴본 데이터 품질 향상 정도 기반 프로파일링 전략이 선택된 경우에는, 복수의 프로파일링 행동 중 데이터 품질 향상 정도가 큰 프로파일링 행동을 우선적으로 추천한다. 즉, 앞서 설명한 결측값이 있는 행을 삭제하는 프로파일링 행동, 결측값이 있는 행을 최솟값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 최댓값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 중앙값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 평균값으로 대체하는 프로파일링 행동, 결측값이 있는 행을 마지막으로 관찰된 값으로 대체하는 프로파일링 행동 또는 결측값이 있는 행을EM알고리즘에 의해 산출된 값으로 대체하는 프로파일링 행동 중 데이터 품질 향상 정도가 가장 큰 순서에 따라 프로파일링 행동을 추천한다. 이때, 데이터 품질 향상 정도를 산출하는 기준으로는 앞서 설명한 KS(kolomogorov smironov) 테스트, 왜도, 첨도, 정보 엔트로피(information entropy)등의 계산 방법을 사용하며, 이를 이용하여 데이터 품질 향상 정도에 대한 순위를 결정할 수 있다.

[0046] 또한, 프로파일링 전략으로서 전문가 레시피 기반 프로파일링 전략이 선택된 경우에는, 전문가가 수행했던 프로파일링 프로세스를 학습시킨 복수의 학습 모델이 각각의 프로파일링 행동으로서 사용된다.

[0047] 도 6은 본 발명의 일실 시예에 따른 전문가 레시피 기반의 프로파일링 행동을 설명하기 위한 도면이다.

[0048] 본 발명에서는 전문가가 수행한 데이터 프로파일링 행동을 라벨링하여 학습한 복수의 학습 모델을 각각의 프로파일링 행동으로서 제공할 수 있다. 이러한 학습 모델의 구축을 위해, 우선 프로파일링 과정에 대한 전문가 로그 데이터를 수집한다. 전문가의 행동으로는 결측 값이 있는 행을 삭제하는 행동이나, 결측 값을 다른 값(최댓값, 최솟값, 중앙값, 평균값, EM, LOCF)으로 대체하는 여러 가지 행동이 수집될 수 있다.

[0049] 다음으로, 이와 같이 수집된 데이터를 전처리하여 학습 데이터를 생성한다. 학습에 사용되는 데이터는 전문가의 프로파일링 행동전 데이터를 대상으로 문제 데이터(결측값, 극단값, 비 일관값) 의 개수를 산출하거나, 데이터의 품질을 산출한다. 그리고, 전문가의 프로파일링 행동에 대한 로그 정보와 해당 프로파일링 행동 이후 문제 데이터가 감소한 개수 또는 데이터의 품질 향상 정도를 라벨링하여 학습 데이터를 생성한다.

[0050] 이와 같이, 수집된 학습 데이터를 이용하여 기계 학습 모델을 구축한다. 학습 과정에서 다양한 전문가의 프로파일링 행동에 기초한 여러 학습 데이터가 구축되므로, 해당 학습 모델을 이용한 추론 과정에서는 입력되는 대상 데이터에 대해, 각 전문가의 프로파일링 행동별로 문제 데이터의 감소 개수 또는 데이터 품질의 향상 정도를 산출할 수 있게 되며, 이를 이용하여, 문제 데이터의 감소 개수 크거나 데이터 품질 향상 정도가 큰 전문가의 프로파일링 레시피를 프로파일링 행동으로서 추천할 수 있게 된다.

[0051] 이와 같이, 본 발명에 따른 데이터 프로파일링 장치(100)를 이용하면, 사용자는 사용자 인터페이스(320)에 표시되는 프로비넌스 그래프에서 특정 노드를 선택하는 과정만으로도, 각 프로파일링 행동에 따른 결과를 미리 확인

할 수 있어, 다양한 프로파일링 행동을 미리 적용하고, 최적의 프로파일링 행동을 선택할 수 있게 된다.

[0052] 이와 같은 과정에 따라 프로파일링 행동이 선택되면, 선택된 프로파일링 행동에 따라 데이터 프로파일링을 실행한다(S250).

[0053] 이때, 입력된 원본 데이터는 원본 상태를 유지하면서, 사용자 인터페이스(320)을 통해 사용자에 의해 순차적으로 선택된 프로파일링 행동이 통합적으로 적용된 결과가 별도의 데이터로서 생성될 수 있다.

[0054] 도 7은 본 발명의 일 실시예에 따른 데이터 프로파일링 결과의 시각화 기능을 도시한 도면이다.

[0055] 먼저, 사용자는 쿼리 입력창(350)을 통해, 자연어 쿼리를 입력할 수 있고, 데이터 프로파일링 장치(100)는 사용자가 입력한 쿼리에 따라 시각화 동작을 수행한다. 사용자는 시각화를 위한 쿼리로서 시각화 대상 데이터의 컬럼을 지정하는 속성(attribute), 시각화 대상 데이터에 대하여 수행하는 작업 종류를 나타내는 태스크(task), 시각화 표현 방법을 타나내는 시각화 유형(Vis type)을 입력할 수 있다.

[0056] 도시된 바와 같이, 사용자는 속성으로서 pm10 데이터와 pm25 데이터를 각각 나타내는 데이터 컬럼을 입력할 수 있다. 또한 사용자는 선택된 데이터들에 대하여 태스크로서 상관 관계(correlation)를 나타내도록 하거나, 필터링(filter)을 수행하게 하거나, 추세(trend)분석을 수행하도록 하는 등을 설정할 수 있다. 또한, 사용자는 해당 태스크를 표현하는 시각화 유형으로서, 막대 그래프, 파이 그래프, 산포도, 히스토그램 등 다양한 시각화 방식을 선택할 수 있다. 도면에서는 pm10 데이터와 pm25 데이터를 선택하고, 그들의 상관 관계를 산포도의 형태로 표시하고 있다. 이러한 시각화 그래프를 통해 최소 값, 최대 값 등의 통계적인 정보를 제공할 수 있다.

[0057] 도 8은 본 발명의 일 실시예에 따른 시각화 그래프의 예를 도시한 것이다.

[0058] 본 발명에서는 데이터 품질을 효과적으로 확인할 수 있도록 하기 위해, 다양한 형태의 그래프를 사용한다.

[0059] 막대 그래프(800)를 통해서는 데이터의 컬럼 별 품질 정도를 분석할 수 있도록 한다. 또한, 히트맵 그래프(810)를 통해 데이터의 결측값 개수를 분석할 수 있는데, 색이 진한 구간이 많을수록 결측값 개수가 많음을 의미한다. 또한, 히스토그램 그래프(820)를 통해 데이터의 극단값 개수를 분석할 수 있는데, 극단값의 높이가 높을수록 극단값 개수가 많음을 의미한다. 또한, 라인 차트(Line chart)를 통해 데이터의 비일관 값 개수를 분석할 수 있으며, 하이라이팅 된 부분이 많을수록 비일관 값 개수가 많음을 의미한다. 또한, 경험적 누적 분포 함수(ECDF, 830)를 통해 정규 분포와 데이터 분포의 경험적 분포 함수 차이를 비교할 수 있으며 두 ECDF의 형태가 다를수록 KS(kolomogorov smironov) 값이 높음을 의미한다. 또한, 밀도분포(Density plot, 840)을 통해 정규 분포와 데이터 분포의 대칭 정도/뾰족한 정도를 비교할 수 있으며 두 밀도 분포의 형태가 다를수록 왜도/첨도가 높음을 의미한다. 또한, 확률 히스트그램을 통해 데이터 분포를 비교할 수 있으며 확률 히스토그램 높이가 균등할수록 정보 엔트로피 값이 높음을 의미한다.

[0060] 본 발명의 일 실시예는 컴퓨터에 의해 실행되는 프로그램 모듈과 같은 컴퓨터에 의해 실행가능한 명령어를 포함하는 기록 매체의 형태로도 구현될 수 있다. 컴퓨터 판독 가능 매체는 컴퓨터에 의해 액세스될 수 있는 임의의 가용 매체일 수 있고, 휘발성 및 비휘발성 매체, 분리형 및 비분리형 매체를 모두 포함한다. 또한, 컴퓨터 판독 가능 매체는 컴퓨터 저장 매체를 포함할 수 있다. 컴퓨터 저장 매체는 컴퓨터 판독가능 명령어, 데이터 구조, 프로그램 모듈 또는 기타 데이터와 같은 정보의 저장을 위한 임의의 방법 또는 기술로 구현된 휘발성 및 비휘발성, 분리형 및 비분리형 매체를 모두 포함한다.

[0061] 본 발명의 방법 및 시스템은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.

[0062] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[0063] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

부호의 설명

[0064] 100: 데이터 프로파일링 장치

110: 메모리

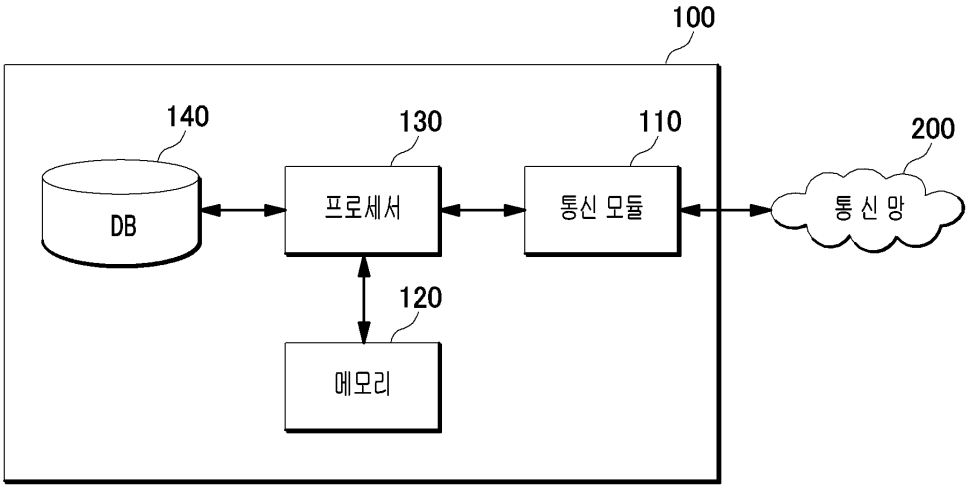
120: 통신 모듈

130: 프로세서

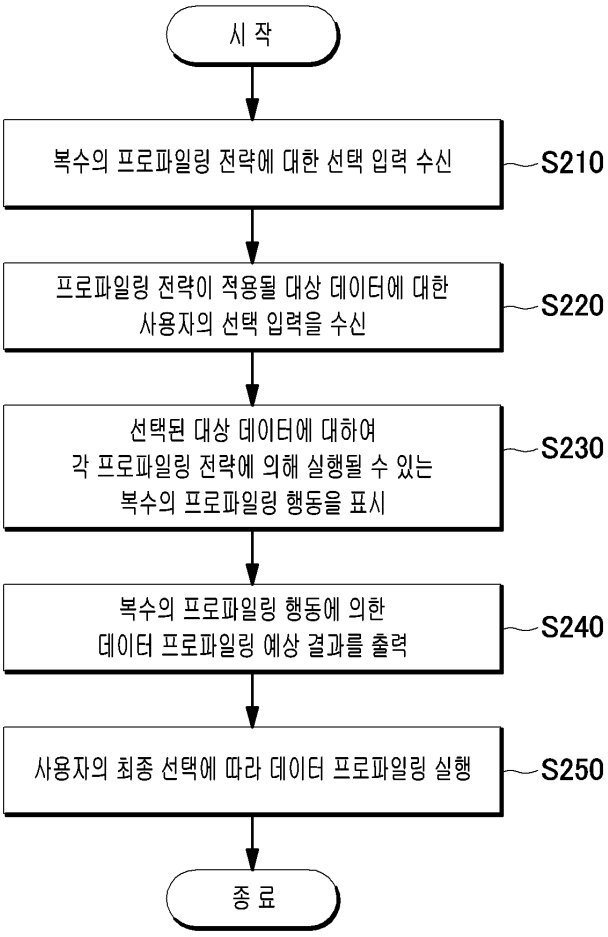
140: 데이터베이스

도면

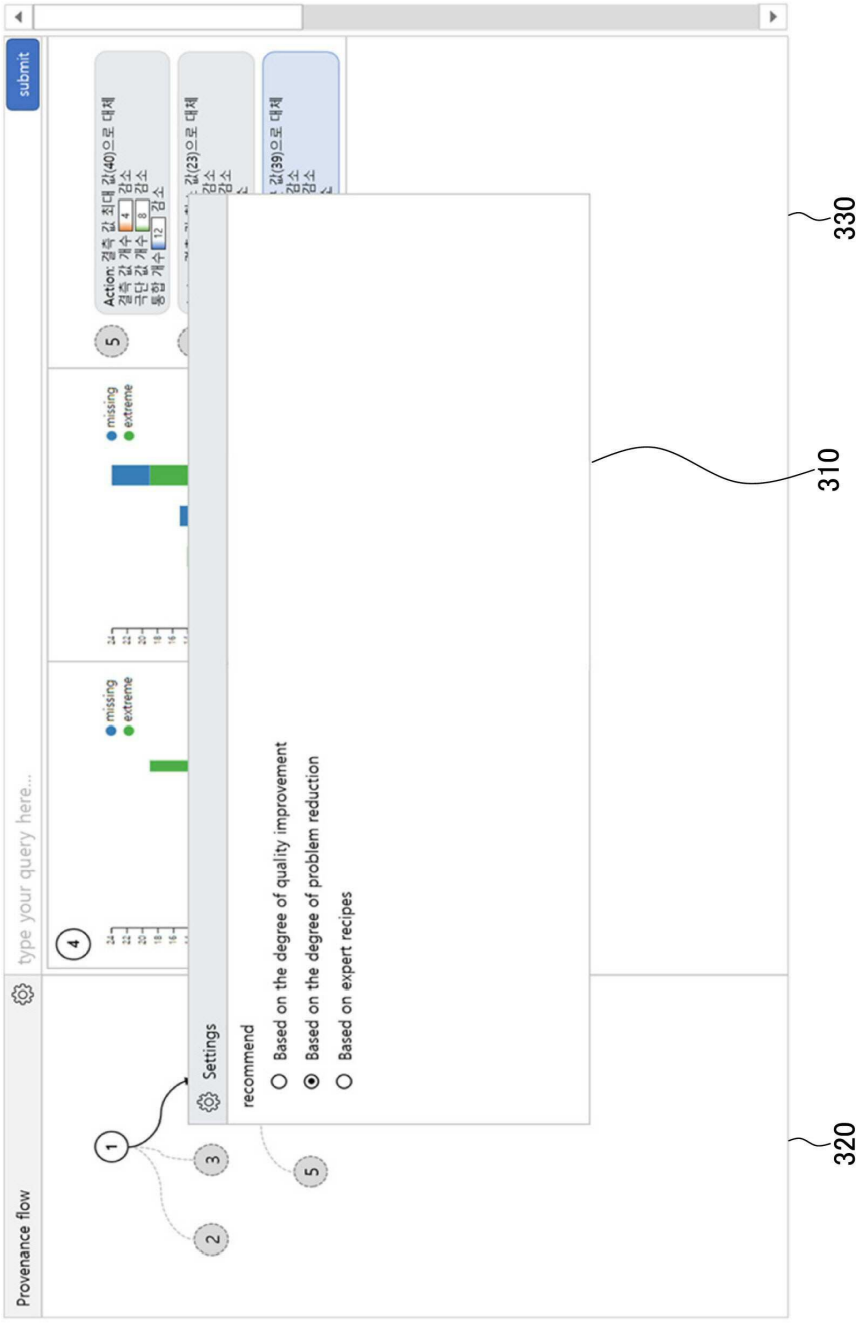
도면1



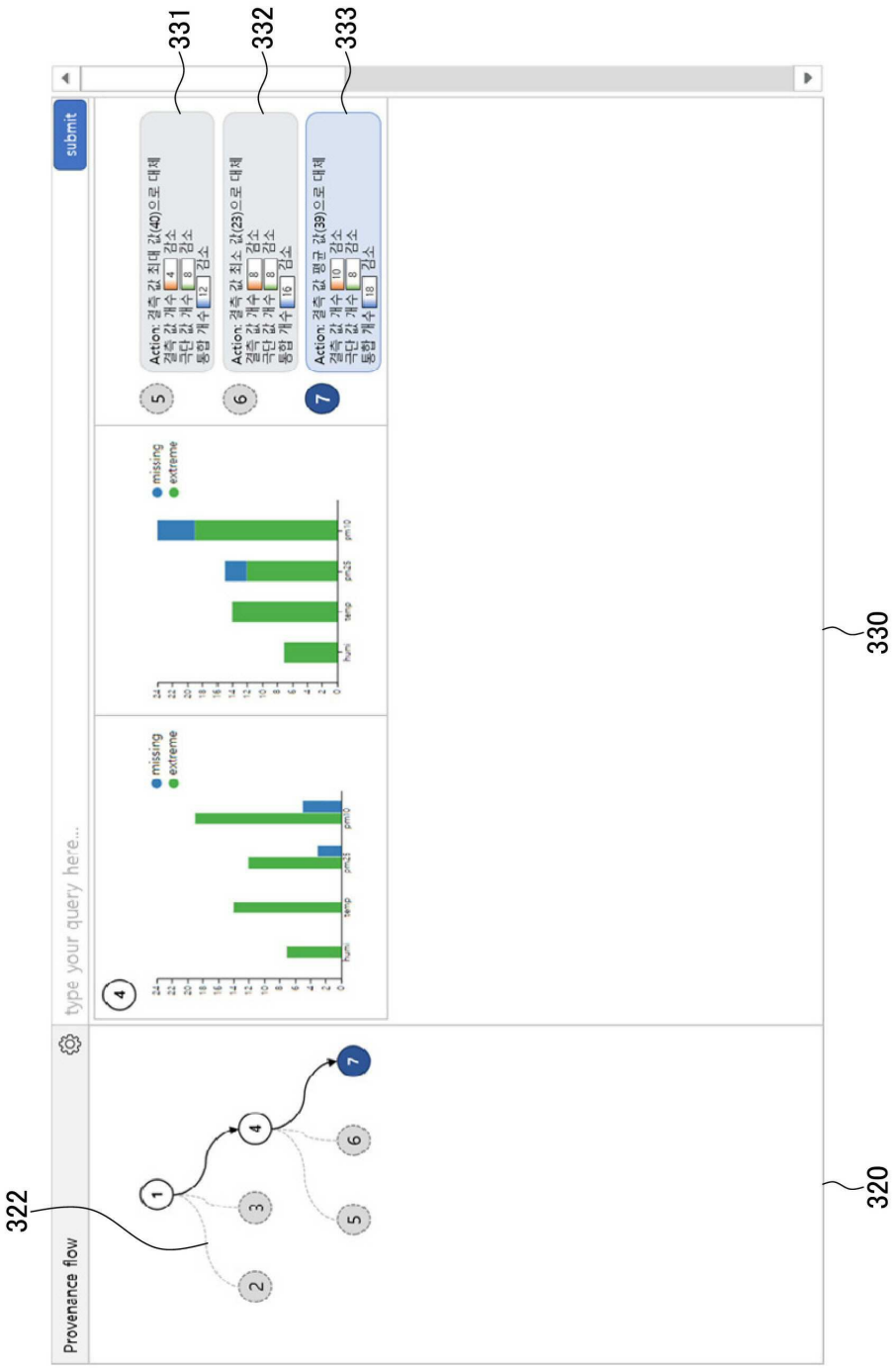
도면2


시 작
복수의 프로파일링 전략에 대한 선택 입력 수신
S210
프로파일링 전략이 적용될 대상 데이터에 대한
사용자의 선택 입력을 수신
S220
선택된 대상 데이터에 대하여
각 프로파일링 전략에 의해 실행될 수 있는
복수의 프로파일링 행동을 표시
S230
복수의 프로파일링 행동에 의한
데이터 프로파일링 예상 결과를 출력
S240
사용자의 최종 선택에 따라 데이터 프로파일링 실행
S250
종 료

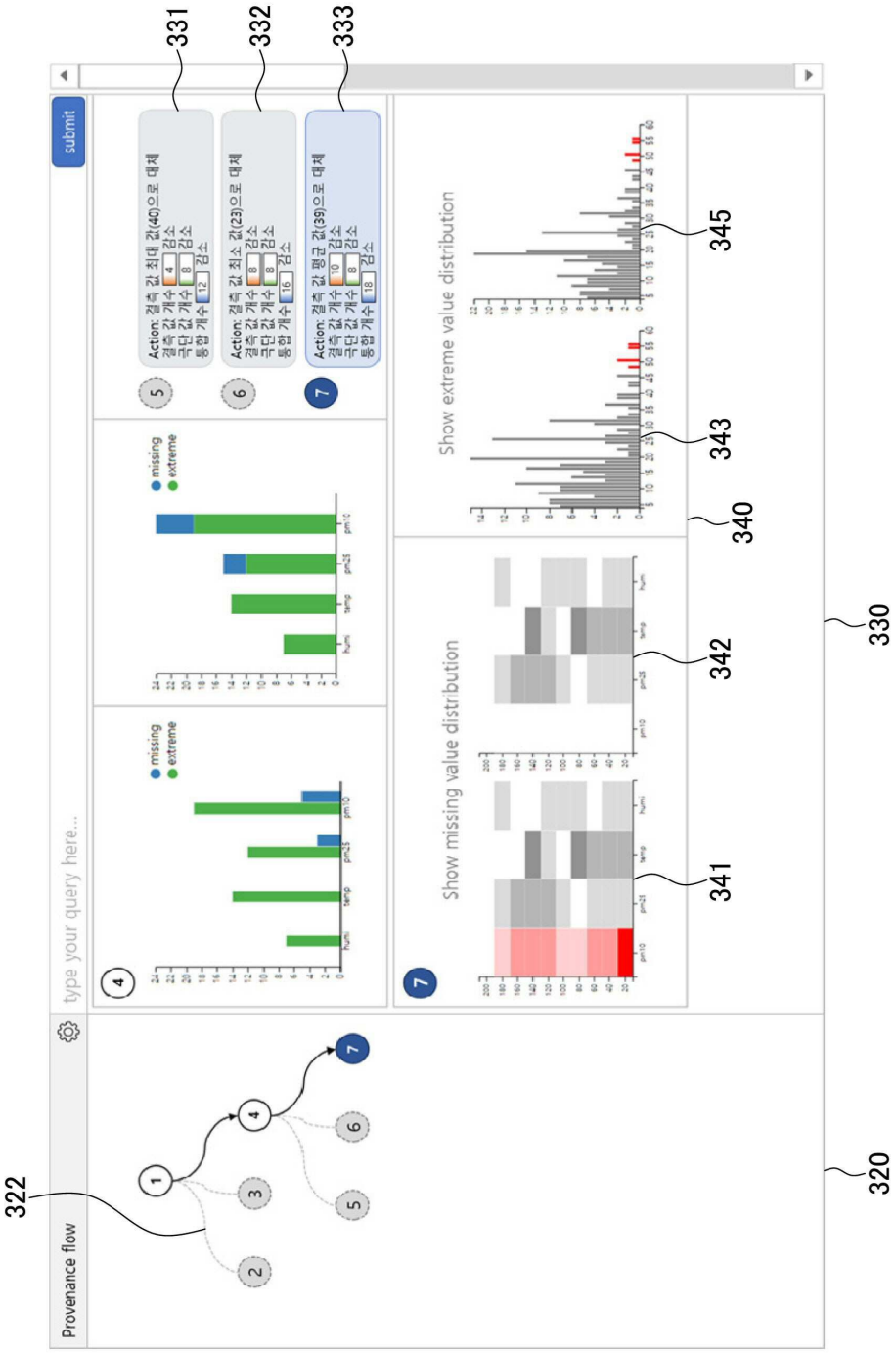
도면3


submit
type your query here...
Provenance flow
Settings
recommend
Based on the degree of quality improvement
Based on the degree of problem reduction
Based on expert recipes
missing
extreme
330
310
320

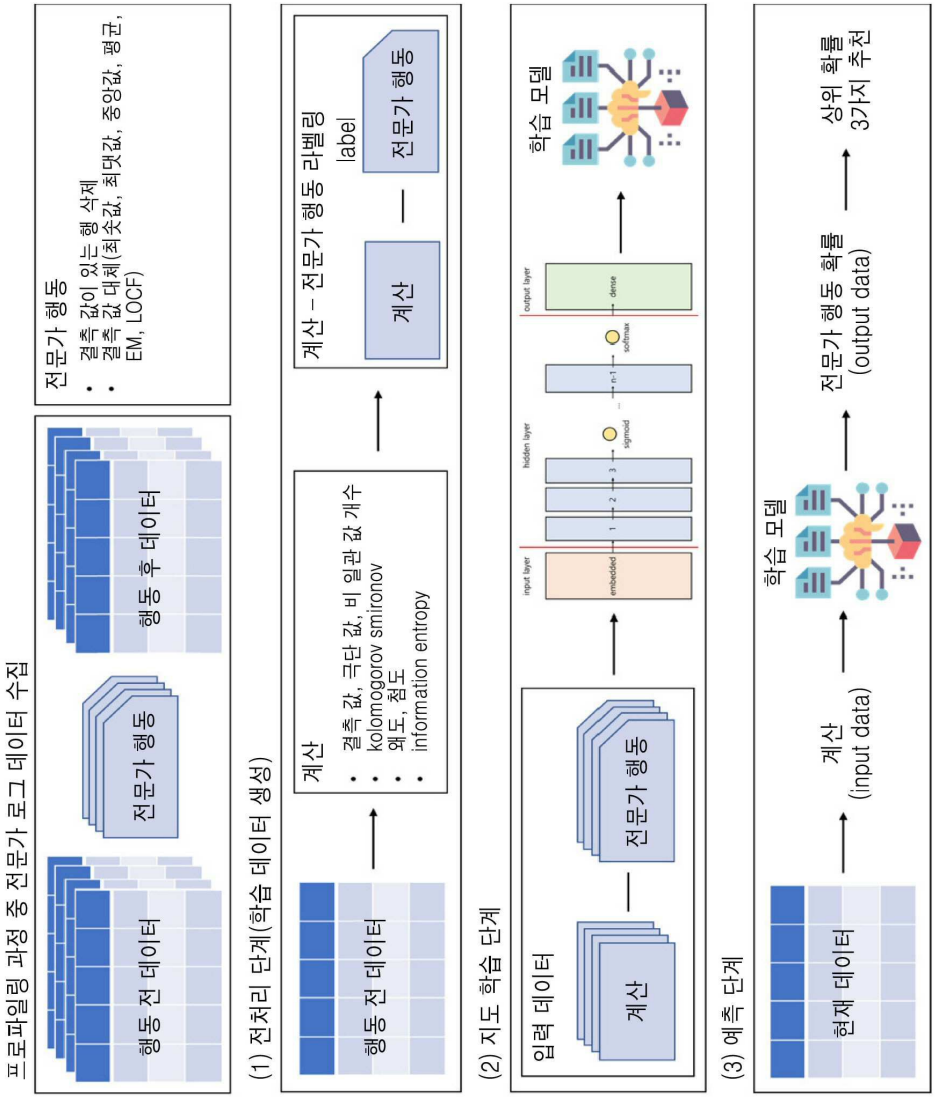
도면4


331
332
333
submit
type your query here...
Provenance flow
322
320
330
Action: 결측 값 최대 값(40)으로 대체
결측 값 개수 4 감소
극단 값 개수 8 감소
동형 개수 12 감소
Action: 결측 값 최소 값(23)으로 대체
결측 값 개수 8 감소
극단 값 개수 8 감소
동형 개수 16 감소
Action: 결측 값 평균 값(39)으로 대체
결측 값 개수 10 감소
극단 값 개수 8 감소
동형 개수 18 감소
missing
extreme

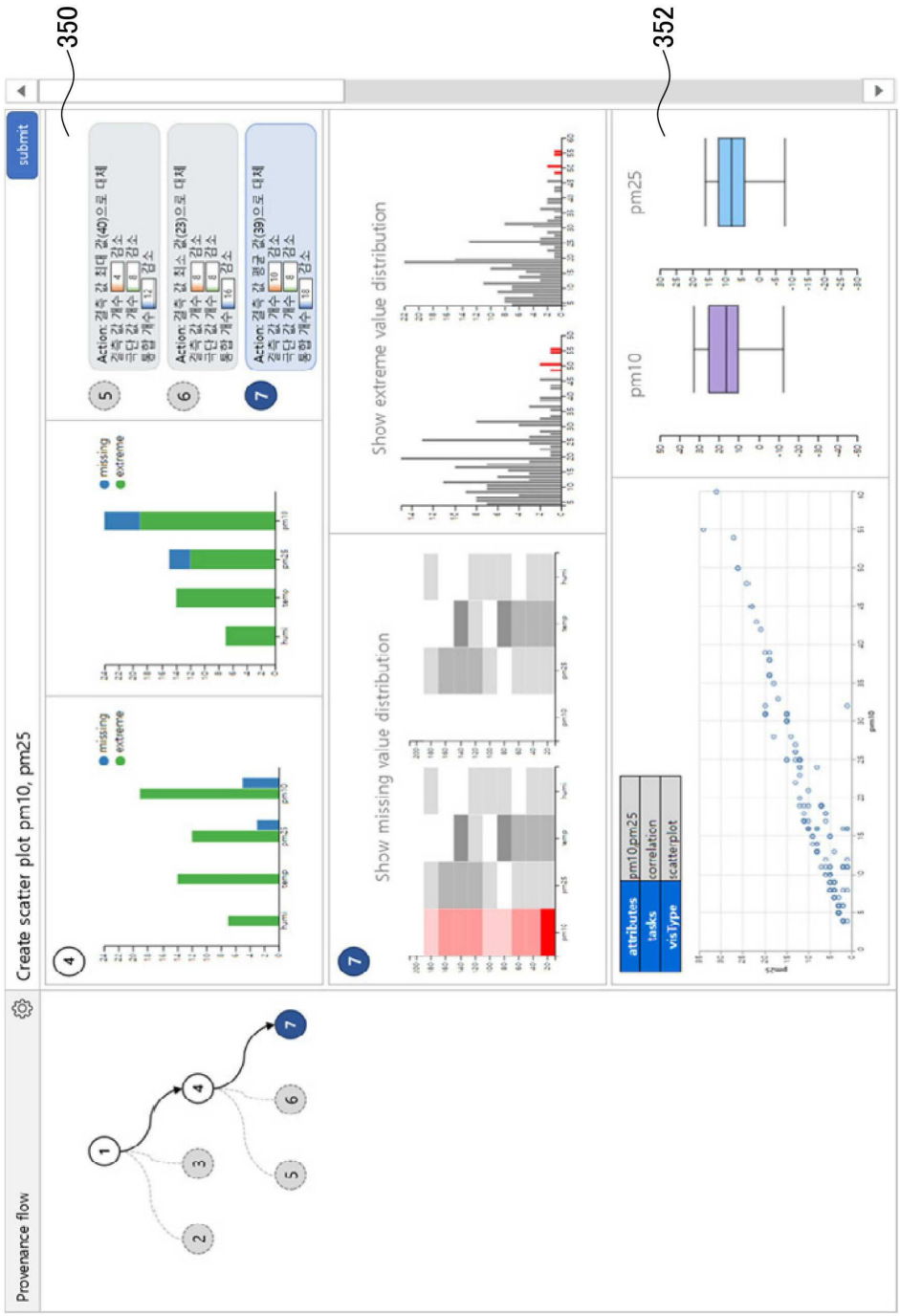
도면5


331
332
333
submit
type your query here...
Provenance flow
Action: 결측 값 최대 값(40)으로 대체
결측 값 개수 4 감소
극단 값 개수 8 감소
통합 개수 12 감소
Action: 결측 값 최소 값(23)으로 대체
결측 값 개수 8 감소
극단 값 개수 8 감소
통합 개수 16 감소
Action: 결측 값 평균 값(39)으로 대체
결측 값 개수 10 감소
극단 값 개수 8 감소
통합 개수 18 감소
missing
extreme
Show extreme value distribution
Show missing value distribution
322
320
330
340
341
342
343
345

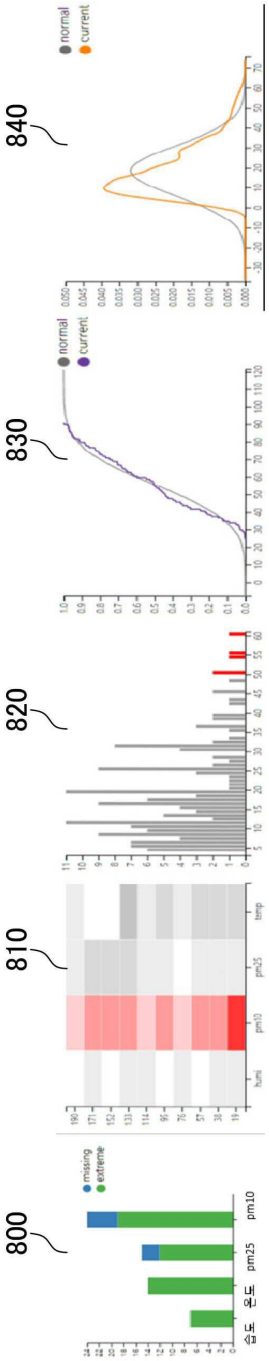
도면6


프로파일링 과정 중 전문가 로그 데이터 수집
행동 전 데이터
전문가 행동
행동 후 데이터
전문가 행동
결측 값이 있는 행 삭제
결측 값 대체(최솟값, 최댓값, 중앙값, 평균, EM, LOCF)
(1) 전처리 단계(학습 데이터 생성)
행동 전 데이터
계산
결측 값, 극단 값, 비 일관 값 개수
kolomogorov smironov
왜도, 첨도
information entropy
계산 – 전문가 행동 라벨링
label
계산
전문가 행동
(2) 지도 학습 단계
입력 데이터
계산
전문가 행동
input layer
embedded
hidden layer
1
2
3
sigmoid
n-1
softmax
output layer
dense
학습 모델
(3) 예측 단계
현재 데이터
계산
(input data)
학습 모델
전문가 행동 확률
(output data)
상위 확률
3가지 추천

도면7


350
352
submit
Create scatter plot pm10, pm25
Provenance flow
Action: 결측 값 최대 값(40)으로 대체
Action: 결측 값 최소 값(23)으로 대체
Action: 결측 값 평균 값(39)으로 대체
Show extreme value distribution
Show missing value distribution
pm25
pm10
attributes
tasks
vis type
pm10pm25
correlation
scatterplot

도면8


800
missing
extreme
습도 온도 pm25 pm10
810
humi
pm10
pm25
temp
820
830
normal
current
840
normal
current