



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0102583
(43) 공개일자 2024년07월03일

(51) 국제특허분류(Int. Cl.)
G06F 18/21 (2023.01) G06F 18/214 (2023.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 18/217 (2023.01)
G06F 18/214 (2023.01)
(21) 출원번호 10-2022-0184757
(22) 출원일자 2022년12월26일
심사청구일자 2022년12월26일

(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)
(뒷면에 계속)
(74) 대리인
특허법인엠에이피에스

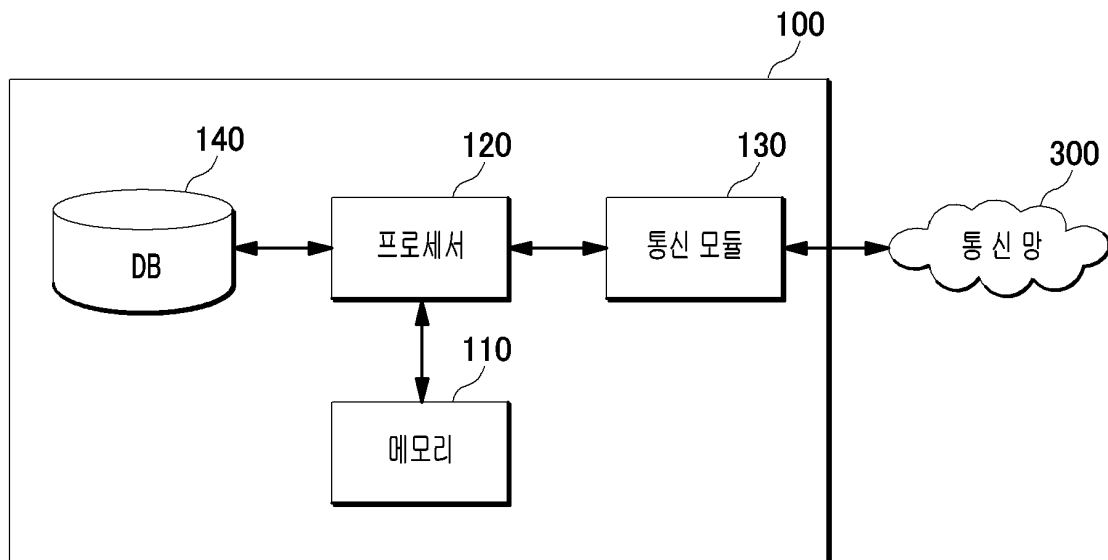
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 데이터 품질을 조정하는 시각적 분석 장치 및 이를 이용한 데이터 품질 조정 방법

(57) 요약

본 발명의 일 측면에 따른 데이터 품질을 조정하는 시각적 분석 장치는, 데이터 품질 조정 프로그램이 저장된 메모리; 상기 데이터 품질 조정 프로그램을 실행하는 프로세서를 포함하되, 상기 데이터 품질 조정 프로그램은, 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신하고, 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하고, 제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력하는 것이다.

대표도 - 도1



(52) CPC특허분류

G06N 20/00 (2021.08)

(72) 발명자

유상봉

서울특별시 광진구 군자로13길 18-5, 3층 201호 (군자동)

진예진

서울특별시 광진구 천호대로114길 44, B101 (능동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152617
과제번호	2021-0-00469-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	이종 융합 데이터 탐지 및 추적 기술 개발
기 여 율	1/2
과제수행기관명	(주)피씨엔
연구기간	2021.04.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711170541
과제번호	2021-0-02076-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	SW컴퓨팅산업원천기술개발
연구과제명	빅데이터 인과 분석을 위한 복잡계 기반 추론 인공지능(REX) 개발 및 실증
기 여 율	1/2
과제수행기관명	한국과학기술연구원(KIST)
연구기간	2022.04.01 ~ 2022.12.31

명세서

청구범위

청구항 1

데이터 품질을 조정하는 시각적 분석 장치에 있어서,

데이터 품질 조정 프로그램이 저장된 메모리;

상기 데이터 품질 조정 프로그램을 실행하는 프로세서를 포함하되,

상기 데이터 품질 조정 프로그램은, 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신하고,

제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하고,

제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력하는 것인, 시각적 분석 장치.

청구항 2

제 1 항에 있어서,

상기 기계 학습 모델의 유형은 분류(classification) 또는 예측(prediction)으로 구분되고,

상기 데이터 품질 조정의 대상이 되는 데이터의 범위는 입력 데이터의 특정 열(column)을 선택하는 형태로 설정되고,

대상 기계 학습 모델의 상세 정보는 예측을 위한 모델인 LR(Logistic Regression), LASSO(Least Absolute Shrinkage and Selection operator), RIDGE, EN(Elastic Net), LAR(Least Angle Regression), LLAR(Lasso Least Angle Regression), OMP(Orthogonal Matching Pursuit), BR(Bayesian Ridge), ARD(Automatic Relevance Determination), PAR(Passive Aggressive Regressor), RANSAC(Random Sample Consensus), TR(TheilSen Regressor), HUBER, KR(Kernel Ridge), SVM(Support Vector Machine), KNN(K-Nearest Neighbor), DT(Decision Trees), RF(Random Forest), ET(Extra Tree), ADA(ADABOOST), GBR(Gradient Boosting Regressor), MLP(Multi Layer Perceptron), XGBOOST(eXtreme Gradient Boosting), LIGHTGBM(Light Gradient Boosting Model) 또는 CATBOOST 중에서 선택되거나, 분류를 위한 모델인 LR(Logistic Regression), KNN(K-Nearest Neighbor), NB(Naive Bayes), DT(Decision Trees), SVM(Support Vector Machine), RBF SVM(Radial Basis Function Support Vector Machine), GPC(Gaussian Process Classifier), MLP(Multi Layer Perceptron), RIDGE, RF(Random Forest), QDA(Quadratic Discriminant Analysis), ADA(ADABOOST), GBC(Gradient Boosting Classifiers), LDA (Linear Discriminant Analysis), ET(Extra Tree), XGBOOST(eXtreme Gradient Boosting), LIGHTGBM(Light Gradient Boosting Model) 또는 CATBOOST 중에서 선택되고,

상기 대상 지표에 대한 정보는 예측을 위한 대상 지표인 MAE(Mean Average Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error), R2, RMSLE(Root Mean Square Logarithmic Error) 또는 MAPE(Mean Absolute Percentage Error) 중에서 선택되거나, 분류를 위한 대상 지표인 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), AUC(Area Under Curve), F1, Kappa, MCC(Matthews Correlation Coefficient) 중에서 선택되는 것인, 시각적 분석 장치.

청구항 3

제 1 항에 있어서,

상기 제 2 인터페이스 블록에서,

상기 데이터 품질 문제의 종류는 결측 값(missing value), 이상값(outlier), 비 일관 값(inconsistent value), 데이터 분포 편향(data distribution bias) 및 특징 범위 차이(feature range difference) 또는 잡음이 있는 레이블(noisy label)을 포함하고,

데이터 품질 조정을 위해 사용된 해결 방법은 대체 방법과 변환 방법을 포함하고,

대체 방법은 제거, 최소값, 최대값, 평균값, 중앙값, 모드, LOCF(last observation carried forward) 및 EM(Expectation Maximization)을 포함하고,

변환 방법은 최대-최소(min-max), 표준화(standardization), robust, maxabs, sqrt, log 및 box-cox를 포함하는 것인, 시각적 분석 장치.

청구항 4

제 1 항에 있어서,

상기 제 3 인터페이스 블록에서,

2차원 그래프의 제 1 축을 통해 데이터 품질이 향상되는 단계를 나타내고, 2차원 그래프의 제 2 축을 통해 데이터 품질의 향상에 따른 기계 학습 모델의 성능 변화를 나타내는 것인, 시각적 분석 장치.

청구항 5

제 1 항에 있어서,

상기 제 3 인터페이스 블록에서,

데이터 품질 문제가 순차적으로 해소됨에 따른 데이터 변화를 각 단계별로 서로 구분되는 막대 그래프의 형태로 표시하고,

상기 막대 그래프에 대응 하여 배치되는 밀도 분포 그래프를 통해, 각 단계별 데이터의 분포와 정규 분포를 비교하여 표시하는 것인, 시각적 분석 장치.

청구항 6

제 1 항에 있어서,

제 4 인터페이스 블록을 통해, 사용자가 상기 데이터 품질 조정을 위해 사용된 해결 방법의 조합을 맞춤형으로 수정하는 것을 지원하는, 데이터 품질을 조정하는 시각적 분석 장치.

청구항 7

데이터 품질을 조정하는 시각적 분석 장치를 이용한 데이터 품질 조정 방법에 있어서,

(a) 입력 데이터를 수신하는 단계

(b) 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신하는 단계;

(c) 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하는 단계; 및

(d) 제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력하는 단계를 포함하는, 데이터 품질 조정 방법.

청구항 8

제 7 항에 있어서,

상기 기계 학습 모델의 유형은 분류(classification) 또는 예측(prediction)으로 구분되고,

상기 데이터 품질 조정의 대상이 되는 데이터의 범위는 입력 데이터의 특정 열(column)을 선택하는 형태로 설정

되고,

대상 기계 학습 모델의 상세 정보는 예측을 위한 모델인 LR(Logistic Regression), LASSO(Least Absolute Shrinkage and Selection operator), RIDGE, EN, LAR(Least Angle Regression), LLAR, OMP(Orthogonal Matching Pursuit), BR, ARD(Automatic Relevance Determination), PAR, RANSAC(Random Sample Consensus), TR, HUBER, KR, SVM(Support Vector Machine), KNN(K-Nearest Neighbor), DT(Decision Trees), RF, ET, ADA(Adaptive Boosting), GBR(Gradient Boosting Regressor), MLP(Multi Layer Perceptron), XGBOOST(eXtreme Gradient Boosting), LIGHTGBM(Light Gradient Boosting Model) 또는 CATBOOST 중에서 선택되거나, 분류를 위한 모델인 LR(Logistic Regression), KNN(K-Nearest Neighbor), NB(Naive Bayes), DT(Decision Trees), SVM(Support Vector Machine), RBFSVM(Radial Basis Function Support Vector Machine), GPC(Gaussian Process Classifier), MLP(Multi Layer Perceptron), RIDGE, RF, QDA(Quadratic Discriminant Analysis), ADA(Adaptive Boosting), GBC(Gradient Boosting Classifiers), LDA (Linear Discriminant Analysis), ET, XGBOOST(eXtreme Gradient Boosting), LIGHTGBM(Light Gradient Boosting Model) 또는 CATBOOST 중에서 선택되고,

상기 대상 지표에 대한 정보는 예측을 위한 대상 지표인 MAE(Mean Average Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error), R2, RMSLE(Root Mean Square Logarithmic Error) 또는 MAPE(Mean Absolute Percentage Error) 중에서 선택되거나, 분류를 위한 대상 지표인 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), AUC(Area Under Curve), F1, Kappa, MCC(Matthews Correlation Coefficient) 중에서 선택되는 것인, 데이터 품질 조정 방법.

청구항 9

제 7 항에 있어서,

상기 (c) 단계에서,

상기 데이터 품질 문제의 종류는 결측 값(missing value), 이상값(outlier), 비 일관 값(inconsistent value), 데이터 분포 편향(data distribution bias) 및 특징 범위 차이(feature range difference) 또는 잡음이 있는 레이블(noisy label)을 포함하고,

데이터 품질 조정을 위해 사용된 해결 방법은 대체 방법과 변환 방법을 포함하고,

대체 방법은 제거, 최소값, 최대값, 평균값, 중앙값, 모드, LOCF(last observation carried forward) 및 EM(expectation Maximization)을 포함하고,

변환 방법은 최대-최소(min-max), 표준화(standardization), robust, maxabs, sqrt, log 및 box-cox를 포함하는 것인, 데이터 품질 조정 방법.

청구항 10

제 7 항에 있어서,

상기 (d) 단계는,

2차원 그래프의 제 1 축을 통해 데이터 품질이 향상되는 단계를 나타내고, 2차원 그래프의 제 2 축을 통해 데이터 품질의 향상에 따른 기계 학습 모델의 성능 변화를 나타내는 것인, 데이터 품질 조정 방법.

청구항 11

제 7 항에 있어서,

상기 (d) 단계는,

데이터 품질 문제가 순차적으로 해소됨에 따른 데이터 변화를 각 단계별로 서로 구분되는 막대 그래프의 형태로 표시하고,

상기 막대 그래프에 대응 하여 배치되는 밀도 분포 그래프를 통해, 각 단계별 데이터의 분포와 정규 분포를 비교하여 표시하는 것인, 데이터 품질 조정 방법.

청구항 12

제 7 항에 있어서,

(e) 제 4 인터페이스 블록을 통해, 사용자가 상기 데이터 품질 조정을 위해 사용된 해결 방법의 조합을 맞춤형으로 수정하는 것을 지원하는 단계를 더 포함하는 것인, 데이터 품질 조정 방법.

청구항 13

제 7 항 내지 제 12 항 중 어느 한 항에 따른 데이터 품질 조정 방법을 수행하기 위한 컴퓨터 프로그램이 기록된 비일시적 컴퓨터 판독가능 기록매체.

발명의 설명

기술 분야

[0001] 본 발명은 기계 학습에 활용되는 데이터의 품질을 조정하는 시각적 분석 장치 및 이를 이용한 데이터 품질 조정 방법에 관한 것이다.

배경 기술

[0002] 최근 의료, 은행, 그리고 교육과 같은 분야에서 기계 학습 모델의 사용이 대중화됨에 따라 기계 학습 모델 성능 향상을 위한 연구가 증가하고 있다. 기계 학습 모델 성능 향상은 학습 모델의 구조를 수정하거나 데이터의 품질을 향상시키는 작업을 통해 향상된다. 그러나 기계 학습 모델의 구조를 수정하는 부분과 관련해서, 파라미터나 내부 구조 변경에 따른 성능의 변화는 공개되어 있지 않은 영역에 해당하여, 모델 성능을 예측하며 모델을 수정하는 작업은 비효율적이다. 또한 기계 학습 모델은 데이터 기반 학습을 수행함에도 불구하고 품질이 낮은 데이터의 양만 늘리는 것은 편향된 학습을 야기할 가능성이 있다. 따라서 데이터 품질을 향상시키는 작업은 기계 학습 모델 성능 향상에 중요하다.

[0003] 데이터의 품질을 향상시키는 방법은 데이터의 품질을 낮추는 원인에 따라 다르다. 데이터의 품질을 낮추는 원인은 결측 값, 이상 값, 비 일관 값, 데이터 분포 편향과 속성(attribute) 크기 불균형이 있으며 결측 값, 이상 값과 비일관 값은 대체(imputation), 데이터 분포 편향과 속성 크기 불균형은 변환(transformation)을 통해 데이터의 품질을 향상시킬 수 있다. 그러나 데이터 품질을 향상시키기 위해서는 분포나 통계적 특성과 같은 데이터의 상태에 따라 다른 데이터 품질 향상 방법이 적용된다. 따라서 데이터에 대한 충분한 지식을 가지고 있지 않은 사용자는 데이터 품질 향상을 위해 많은 시간과 노력을 소비하게 된다.

[0004] 품질 연구에서 시각적 분석 방법은 데이터 품질 향상 과정에 대한 의사 결정 지원을 위해 제안되어 왔다. 그러나 기존에 제안된 시스템은 데이터 자체에 대한 품질 향상에 초점을 두므로 데이터의 품질 향상이 기계 학습 모델의 성능에 미치는 영향에 대한 평가를 제공하지 않는다. 또한 기존에 제안된 시스템은 데이터 품질 향상 과정에 대한 모든 조합을 사용자가 직접 고민해야 하는 한계점이 남아있다.

[0005] 따라서 본 발명은 데이터 품질을 향상시키는 과정 중 사용자의 의사 결정을 지원하는 새로운 시각적 분석 시스템을 제안한다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 대한민국공개특허번호 제 10-2022-0129645호(발명의 명칭: 데이터 품질 자동 향상)

발명의 내용

해결하려는 과제

[0007] 본 발명은 전술한 문제점을 해결하기 위하여, 기계 학습에 사용되는 데이터의 품질을 손쉽게 조정할 수 있는 시각적 분석 장치 및 방법을 제공하는 것을 목적으로 한다.

[0008] 다만, 본 실시예가 이루고자 하는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0009] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 데이터 품질을 조정하는 시각적 분석 장치는, 데이터 품질 조정 프로그램이 저장된 메모리; 상기 데이터 품질 조정 프로그램을 실행하는 프로세서를 포함하되, 상기 데이터 품질 조정 프로그램은, 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신하고, 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하고, 제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력하는 것이다.

[0010] 또한, 본 발명의 다른 측면에 따른 데이터 품질을 조정하는 시각적 분석 장치를 이용한 데이터 품질 조정 방법에 있어서, (a) 입력 데이터를 수신하는 단계, (b) 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신하는 단계; (c) 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하는 단계; 및 (d) 제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력하는 단계를 포함한다.

발명의 효과

[0011] 전술한 과제의 해결 수단에 따라, 본 발명에서는 데이터 품질을 향상시키면서 의사 결정을 지원하는 새로운 시각적 분석 시스템을 제안한다. 제안하는 시스템은 데이터 기반 전략에서 가능한 품질 개선 방법을 식별하고 프로세스 기반 전략에서 최적의 품질 개선 프로세스를 제공한다. 본 발명은 DQI (Data Quality Improvement) 조합에 대체, 변환 및 시각적 대화형 레이블 지정을 포함하여 기계 학습 모델에 대한 포괄적인 DQI를 처리할 수 있도록 한다. 또한, 본 발명은 기계 학습 성능을 기반으로 품질 개선 프로세스를 재설계하고, 다양한 시각적 분석 기술을 사용하여 데이터 품질에 추가 개선이 필요한지 사용자에게 통찰력을 제공한다.

[0012] 특히, 시각적 인터페이스 블록을 통해, 기계 학습 모델의 성능과 데이터 상태에 따른 최적의 대체와 변환 방법 및 과정을 시각적으로 확인하기 쉽게 제시한다. 또한, 본 발명은 다양한 시각화 방법을 사용하여 데이터의 품질을 낮추는 원인의 위치나 분포, 데이터 품질 향상 단계마다 기계 학습 모델의 성능 변화와 같은 정보를 제공할 수 있다.

도면의 간단한 설명

[0013] 도 1은 본 발명의 일 실시예에 따른 시각적 분석 장치의 구성을 도시한 블록도이다.

도 2는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 실행되는 각 인터페이스 블록의 구성을 도시한 도면이다.

도 3은 본 발명의 일 실시예에 따른 시각적 분석 장치에서 사용되는 아이콘의 예를 도시한 도면이다.

도 4는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 사용되는 기계 학습 모델의 종류와 해결 방법의 종류를 도시한 도면이다.

도 5는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 실행되는 데이터 품질 조정 분석 방법을 도시한 도면이다.

도 6은 본 발명의 일 실시예에 따른 시각적 분석 장치가 수행하는 데이터 품질 조정 방법을 도시한 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0014] 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.

[0015] 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.

[0016] 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.

[0017] 이하에서 언급되는 시각적 분석 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, 데스크톱(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.

[0018] 도 1은 본 발명의 일 실시예에 따른 시각적 분석 장치의 구성을 도시한 블록도이다.

[0019] 시각적 분석 장치(100)는 메모리(110), 프로세서(120), 통신모듈(130), 및 데이터베이스(140)를 포함한다. 또한, 시각적 분석 장치(100)는 사용자를 위한 I/O 모듈을 더 포함할 수 있다. 시각적 분석 장치(100)는 시각화 결과물을 외부 컴퓨팅 장치에 제공하는 서버로서도 동작할 수 있으며, 예를 들면, SaaS (Software as a Service), PaaS (Platform as a Service) 또는 IaaS (Infrastructure as a Service)와 같은 클라우드 컴퓨팅 서비스 모델의 형태로도 동작할 수 있다.

[0020] 메모리(110)는 데이터 품질 조정 프로그램이 저장된다. 또한, 메모리(110)에는 시각적 분석 장치(100)의 구동을 위한 운영 체제나 데이터 품질 조정 프로그램의 실행 과정에서 발생하는 여러 종류가 데이터가 저장된다. 이때, 메모리(110)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다. 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.

[0021] 메모리(110)에 저장되는 데이터 품질 조정 프로그램은 제 1 인터페이스 블록을 통해 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 데이터 품질 측정을 위한 대상 지표에 대한 정보를 수신하고, 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력하고, 제 3 인터페이스 블록을 통해, 상기 해결 방법의 조합을 수행하는 단계마다 상기 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력한다. 또한, 데이터 품질 조정 프로그램은 제 4 인터페이스 블록을 통해 사용자가 데이터 품질 조정을 위해 사용된 해결 방법의 조합을 맞춤형으로 수정하는 것을 지원할 수 있다.

- [0022] 프로세서(120)는 메모리(110)에 저장된 데이터 품질 조정 프로그램을 실행하되, 데이터 품질 조정 프로그램의 실행에 따르는 전체 과정을 제어한다. 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, 마이 크로프로세서(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.
- [0023] 통신모듈(130)은 시각적 분석 장치(100)의 결과를 외부 컴퓨팅 장치에 전송하거나, 외부 컴퓨팅 장치로부터 데이터 품질 조정을 위한 입력 데이터나 데이터 품질 조정 프로그램의 실행 필요한 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.
- [0024] 데이터베이스(140)는 데이터 품질 조정 프로그램의 실행과정에서 생성되는 여러 데이터나, 다양한 기계 학습 모델, 각 기계 학습 모델의 구축에 사용된 학습 데이터가 저장될 수 있다.
- [0025] 도 2는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 진행되는 각 인터페이스 블록의 구성을 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 시각적 분석 장치에서 사용되는 아이콘의 예를 도시한 도면이며, 도 4는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 사용되는 기계 학습 모델의 종류와 해결 방법의 종류를 도시한 도면이다.
- [0026] 도시된 바와 같이, 데이터 품질 조정 프로그램은 제 1 인터페이스 블록(a), 제 2 인터페이스 블록(b), 제 3 인터페이스 블록(c)을 실행한다. 또한, 데이터 품질 프로그램은 데이터를 업로드하는 인터페이스 블록과 제 4 인터페이스 블록(d)을 추가로 실행할 수 있다.
- [0027] 데이터 품질 프로그램은 데이터를 업로드하는 인터페이스 블록을 통해 CSV 형식의 숫자 데이터를 수신할 수 있는데, 본 발명이 이에 제한되는 것은 아니며, 다른 형태의 데이터를 수신하는 것도 가능하다. 도면에서는 예시적으로 전복(abalone)의 크기 등에 대한 데이터가 입력되는 과정을 나타내고 있다.
- [0028] 데이터 품질 조정 프로그램은 제 1 인터페이스 블록(a) 을 통해 기계 학습 모델의 유형(Model type), 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위(Target column), 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보(Target model), 대상 기계 학습 모델의 성능 측정을 위한 대상 지표(Target metric)에 대한 정보를 수신한다. 이때, 기계 학습 모델의 유형은 분류(classification) 또는 예측(prediction)으로 구분된다. 그리고, 데이터 품질 조정의 대상이 되는 데이터의 범위는 입력 데이터의 특정 열(column)을 선택하는 형태로 설정될 수 있다.
- [0029] 또한, 예측을 위한 기계 학습 모델로는 도 4에 도시된 바와 같이, LR(Logistic Regression), LASSO(Least Absolute Shrinkage and Selection operator), RIDGE, EN(Elastic Net), LAR(Least Angle Regression), LLAR(Lasso Least Angle Regression), OMP(Orthogonal Matching Pursuit), BR(Bayesian Ridge), ARD(Automatic Relevance Determination), PAR(Passive Aggressive Regressor), RANSAC(Random Sample Consensus), TR(TheilSen Regressor), HUBER, KR(Kernel Ridge), SVM(Support Vector Machine), KNN(K-Nearest Neighbor), DT(Decision Trees), RF(Random Forest), ET(Extra Tree), ADA(ADABOOST), GBR(Gradient Boosting Regressor), MLP(Multi Layer Perceptron), XGBOOST(eXtreme Gradient Boosting), LIGHTGBM(Light Gradient Boosting Model), CATBOOST등 18개 대상 모델 중 하나 이상을 선택할 수 있다. 또한, 선택된 예측 기계 학습 모델의 성능을 측정하는 지표로는 MAE(Mean Average Error), MSE(Mean Squared Error), RMSE(Root Mean Squared Error), R2, RMSLE(Root Mean Square Logarithmic Error), MAPE(Mean Absolute Percentage Error)을 포함한 6가지 대상 지표를 선택할 수 있다.
- [0030] 또한, 분류를 위한 기계 학습 모델로는 도 4에 도시된 바와 같이, LR(Logistic Regression), KNN(K-Nearest Neighbor), NB(Naive Bayes), DT(Decision Trees), SVM(Support Vector Machine), RBFSVM(Radial Basis Function Support Vector Machine), GPC(Gaussian Process Classifier), MLP(Multi Layer Perceptron), RIDGE, RF(Random Forest), QDA(Quadratic Discriminant Analysis), ADA(ADABOOST), GBC(Gradient Boosting Classifiers), LDA (Linear Discriminant Analysis), ET(Extra Tree), XGBOOST(eXtreme Gradient Boosting),

LIGHTGBM(Light Gradient Boosting Model), CATBOOST등 18개 대상 모델 중 하나 이상을 선택할 수 있다. 또한, 선택된 분류 기계 학습 모델의 성능을 측정하는 지표로는 정확도(Accuracy), 재현율(Recall), 정밀도(Precision), AUC(Area Under Curve), F1, Kappa, MCC(Matthews Correlation Coefficient) 을 포함한 6가지 대상 지표(metric)를 선택할 수 있다. 위의 기계 학습 모델이나 각 측정 지표는 모두 종래 기술에 해당하는 것으로, 구체적인 내용에 대한 설명은 생략하기로 한다.

- [0031] 이와 같이, 사용자는 제 1 인터페이스 블록(a)을 통해 자신이 희망하는 기계 학습 모델의 유형, 데이터의 범위, 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표를 설정할 수 있다.
- [0032] 다음으로, 데이터 품질 조정 프로그램은 제 2 인터페이스 블록(b)을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 상기 대상 지표에서의 성능 측정 결과를 출력한다.
- [0033] 도면에서, 인덱스는 각 대상 기계 학습 모델에 해결 방법의 조합이 매칭된 것을 구분하기 위한 일련번호를 나타낸다. 그리고, 데이터 품질 문제의 종류(210)는 데이터 품질 조정을 위해 사용된 해결 방법의 조합에 의해 해결된 품질 문제의 상세 유형을 나타내는 것으로 예를 들면, 결측 값(missing value), 이상값(outlier), 비 일관 값(inconsistent value), 데이터 분포 편향(data distribution bias) 및 특징 범위 차이(feature range difference) 또는 잡음이 있는 레이블(noisy label) 등이 될 수 있다.
- [0034] 결측값은 공란인 값을 의미하고, 이상값은 대표값보다 작거나 큰 값을 의미하며, 비 일관 값은 대표값과 형식이 다른 값을 의미한다. 결측이나 비 일관 값은 특정 방법을 사용하지 않고도 식별할 수 있지만, 이상값의 식별에는 특정 기준이 필요하다. 예를 들어, $1.5 \times \text{IQR}$ (InterQuartile Range)을 이상값 기준으로 활용하고 그 값이 기준보다 크거나 작으면 이상값으로 판단할 수 있다. 또한, 데이터 분포 편향과 특징 범위 차이를 식별하기 위해 왜도(skewness), 첨도(kurtosis) 및 상자 그림(box plot)을 사용할 수 있다. 왜도 및 첨도는 데이터 분포의 비대칭 정도와 데이터 분포의 선명도를 나타내는 수치이다. 왜도 및 첨도가 0에 가까울수록 데이터 분포가 정규 분포와 유사해진다. 또한, 왜도와 첨도의 절대값이 각각 3 미만과 7 미만인 데이터는 정규 분포로 판단된다. 또한, 사용자가 노이즈 라벨을 판단할 수 있도록 시각적 대화형 라벨링(visual interactive labeling)을 지원할 수 있다. 그리고, 시각적 대화형 라벨링에 필요한 차원 축소 방법으로 TSNE(t-Distributed Stochastic Neighbor Embedding)와 PCA(Principle Component Analysis)사용하고, 산점도(scatter plot)를 시각화 기술로 사용할 수 있다.
- [0035] 이러한, 품질 문제의 종류는 도 3에 도시된 형태의 아이콘으로 표시될 수 있으며, 이러한 아이콘을 제 2 인터페이스 블록에 표시하여, 데이터 품질 문제의 종류를 보다 직관적으로 확인할 수 있도록 한다.
- [0036] 또한, 데이터 품질 조정을 위해 사용된 해결 방법의 조합(212)은 대체(imputation) 방법과 변환(transformation) 방법을 포함할 수 있고, 대체 방법은 제거, 최소값, 최대값, 평균값, 중앙값, 모드, LOCF(last observation carried forward) 및 EM(Expectation Maximization)을 포함하고, 변환 방법은 최대-최소(min-max), 표준화(standardization), robust, maxabs, sqrt, log 및 box-cox를 포함하는 것이다. 데이터 품질 조정을 위해 하나 이상의 해결 방법이 조합되어 사용될 수 있는데, 해결 방법의 조합은 데이터 품질을 악화시키는 데이터 품질 문제를 해소하기 위해 선택할 수 있는 다양한 해결 방법을 연결시킨 것을 의미한다.
- [0037] 이러한, 해결 방법의 상세한 종류는 도 3에 도시된 형태의 아이콘으로 표시될 수 있으며, 이러한 아이콘을 제 2 인터페이스 블록에 표시하여, 데이터 품질 조정을 위해 사용된 해결 방법의 종류를 보다 직관적으로 확인할 수 있도록 한다. 한편, 각각의 해결 방법은 모두 종래 기술에 해당하는 것으로, 구체적인 내용에 대한 설명은 생략하기로 한다.
- [0038] 한편, 제 2 인터페이스 블록은 각 기계 학습 모델별로 데이터 품질 문제의 종류와 품질 조정을 위해 사용된 해결 방법의 조합이 서로 매칭된 형태로 복수개가 나열되어 표시되는데, 각 해결 방법의 조합에 대상 지표에 의한 성능 측정 결과를 토대로 정렬된다. 즉, 도 2에서는 대상 지표로서 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)이 선택되어 있는데, 이 값이 모두 최대로 산출되는 해결 방법의 조합이 가장 상단에 표시되도록 정렬하여 표시한다. 이를 통해, 어떠한 해결 방법의 조합을 선택할때, 기계 학습 모델의 성능을 최대로 향상시킬 수 있을지를 쉽게 확인할 수 있다.
- [0039] 다음으로, 데이터 품질 조정 프로그램은 제 3 인터페이스 블록(c) 을 통해, 해결 방법의 조합을 수행하는 단계마다 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력한다. 특히, (c-1)에 도시된 바와 같이, 2차원 그래프의 X 축을 통해 데이터 품질이 향상되는 단계를 나타내고, 2차원 그래프의 Y 축을 통해 데이터 품질

의 향상에 따른 기계 학습 모델의 성능 변화를 나타낼 수 있다. 2차원 그래프는 연결된 산점도(connected scatter plot)로 도시될 수 있는데, X 축에서는 해결 방법의 조합이 단계적으로 실행될 때, 그에 따라 데이터 품질이 향상되는 정도를 단계적으로 표시하고, 데이터 품질의 향상에 따라 기계 학습 모델의 성능이 개선되는 정도를 Y 축에 표시한다. 도시된 바와 같이, 데이터의 품질이 향상됨에 따라, 기계 학습 모델에 대상 지표에서의 분류 성능이 향상됨을 확인할 수 있다.

[0040] 또한, (c-2)에 도시된 바와 같이, 데이터 품질 문제가 순차적으로 해소됨에 따른 데이터 변화를 각 단계별로 서로 구분되는 막대 그래프의 형태로 표시하고, 각 막대 그래프에 대응하여 배치되는 밀도 분포 그래프를 통해, 각 단계별 데이터의 분포와 정규 분포를 비교하여 표시할 수 있다. 도시된 바와 같이, 막대 그래프의 초기 상태(init)에서는 결측값(파랑색), 이상값(주황색), 비일관 값(초록색)의 개수가 최대로 나타나지만, 해결 방법의 조합을 순차적으로 수행함에 따라 결측값, 이상값, 비일관 값이 단계적으로 감소하는 것을 확인할 수 있도록 복수의 막대 그래프가 각각 표시된다. 그리고, 각 막대 그래프에 대응하는 밀도 분포 그래프를 통해, 데이터 품질이 향상될수록, 실제 데이터 분포(진회색)가 정규 분포(연회색)에 가까워짐을 확인할 수 있다. 데이터 품질이 향상될 수 록, 데이터 분포 편향(data distribution bias) 및 특징 범위 차이(feature range difference)도 개선되는 것도 확인할 수 있다.

[0041] 또한, 데이터 품질 조정 프로그램은 제 4 인터페이스 블록(d)을 통해 사용자가 데이터 품질 조정을 위해 사용된 해결 방법의 조합을 맞춤형으로 수정하는 것을 지원할 수 있다. 특히, 사용자가 데이터 도메인 지식을 보유하고 있는 경우에는, 이를 활용하여 데이터 품질 향상 조합을 맞춤화하도록 한다.

[0042] (d-1)은 데이터에 존재하는 품질 문제들을 배터리 차트(battery chart, 216)와 히트맵(218)의 형태로 출력한다. 배터리 차트는 데이터 품질 문제의 종류별로 데이터 품질 문제의 개수를 나타내며, 히트맵은 데이터 품질 문제들의 분포를 표시한다. 히트맵은 선택된 배터리 차트의 품질 문제의 종류와 같은 계열의 색상으로 인코딩되는데, 현재 도면에서는 결측 값의 분포를 표시하는 형태로 도시되어 있다. 사용자는 히트맵을 사용하여 데이터 품질 문제가 집중된 부분을 식별하고 맞춤형으로 데이터 품질 문제를 개선할 대상, 즉 데이터의 범위로서 특정 컬럼을 선택할 수 있게 된다.

[0043] (d-2)는 히스토그램 그래프와 산점도(scatter plot)를 제공한다. 앞서 설명한 (d-1)의 히트맵에서 선택한 컬럼에 대한 분포를 표시한다. 또한 히스토그램 아래의 브러시 바(brush bar) 그래프는 특정 컬럼 내 인스턴스들의 범위를 조정하는 용도로 사용된다. 사용자는 데이터 품질 조정 프로그램이 모든 컬럼에 대해 동일한 데이터 품질 향상 조합을 처리한 것을 수정하기 위한 목적으로 히스토그램을 사용할 수 있다. 산점도는 유사한 인스턴스끼리 밀집되는 결과를 보여표시한다. 산점도는 점에 해당하는 특정 인스턴스를 선택하여 대체 방법, 변환 방법 및 시각적 대화형 라벨링(visual interactive labeling)을 수행하는 기능을 제공한다. 사용자는 노이즈 라벨을 해결하거나 일부 인스턴스에 대해 수정할 때 산점도를 사용할 수 있다.

[0044] 데이터 품질 조정 프로그램은 제 5 인터페이스 블록(e)을 통해, 사용자가 입력한 질문과 관련된 시각화를 자동으로 생성할 수 있다. 이를 위해, NL4DV(A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries) 모델을 사용할 수 있는데, NL4DV 모델은 사용자가 입력한 자연어 쿼리에서 속성, 작업 및 vis 유형을 식별한다. 속성은 데이터에 존재하는 하나 이상의 열이다. 이 작업은 상관 관계, 필터 및 추세와 같은 자연어 쿼리를 사용하는 것을 목적으로 한다. vis 유형은 막대 차트, 파이 차트, 산점도 및 히스토그램을 포함하는 시각화 방법이다. 이를 통해, 사용자는 시스템이 제공하는 정보 외에 궁금한 점이 생겼을 때 NLP2Vis를 사용하여 관련된 시각화 정보를 획득할 수 있다.

[0045] 한편, 도 2에 도시된 각 인터 페이스 블록은 하나의 화면에서 출력되는 것으로 도시되어 있으나, 본 발명이 이에 제한되는 것은 아니며, 각각이 서로 구분되어 별도의 화면에서 표시되는 것도 가능하다.

[0046] 도 5는 본 발명의 일 실시예에 따른 시각적 분석 장치에서 실행되는 데이터 품질 조정 분석 방법을 도시한 도면이다.

[0047] 전체적인 구성은 도 2에 도시된 인터페이스 구성과 거의 동일하며, 기계 학습 모델의 유형이 예측 모델이라는 점과, 그에 따라 선택된 기계 학습 모델의 상세 정보, 해결 방법의 조합, 성능 측정을 위한 대상 지표의 상세한 내용이 달라졌음을 확인할 수 있다. 또한, 본 실시예에서는 자전거 대여 횟수를 예측하는 기계 학습 모델의 성능을 측정하는 것으로, 해결 방법의 조합이 각 단계별로 적용됨에 따라 데이터 품질 문제가 개선되고, 이에 따라 기계 학습 모델의 예측 결과의 에러가 점차적으로 감소하여 학습 모델의 성능이 향상되는 것을 확인할 수 있다.

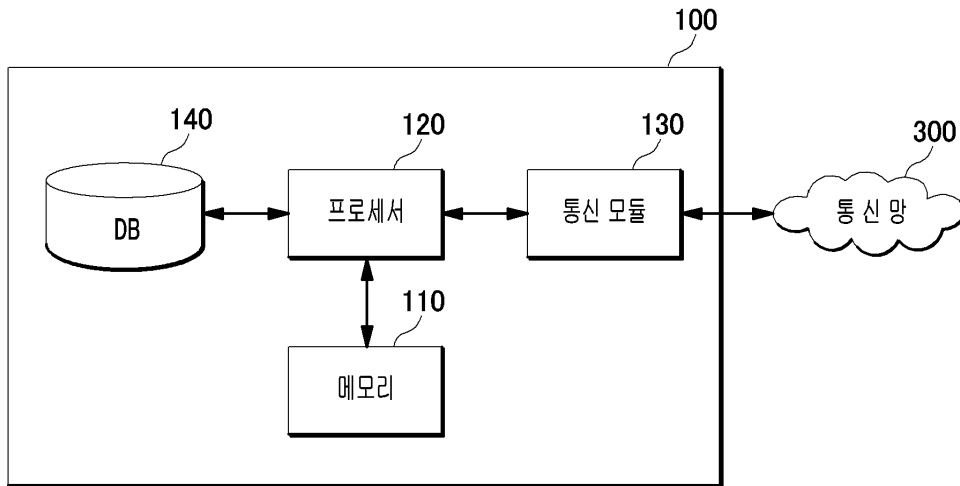
- [0048] 이와 같이, 본 발명에 따른 시각적 분석 장치(100)는 데이터 품질 향상을 위해 적절한 해결 방법의 조합을 구축하고, 조합에 따라 품질이 향상된 데이터 세트를 생성하고, 품질이 향상된 데이터를 바탕으로 대상 기계 학습 모델의 성능 측정 결과를 출력하여, 사용자가 적절한 해결 방법의 조합을 선택할 수 있게한다. 또한, 해결 방법의 조합의 일부를 행동 상세(action detail)을 통해 사용자가 커스터마이징할 수 있도록 지원하는데, 행동 상세는 사용자가 도메인 지식을 활용하여 세부적으로 데이터를 선택하고 수정할 수 있도록 한다. 또한 이 단계에서 사용자에게 피드백을 주어 추가 작업에 대한 의사결정을 돕는다.
- [0049] 도 6은 본 발명의 일 실시예에 따른 시각적 분석 장치가 수행하는 데이터 품질 조정 방법을 도시한 순서도이다.
- [0050] 먼저, 시각적 분석 장치(100)는 데이터 품질 조정 프로그램의 실행에 따라 디스플레이에 표시된 입력 인터페이스를 통해, 데이터 품질 조정의 대상이 되는 입력 데이터를 수신한다(S610).
- [0051] 다음으로, 데이터 품질 조정 프로그램이 실행하는 제 1 인터페이스 블록을 통해, 기계 학습 모델의 유형, 입력 데이터 중 데이터 품질 조정의 대상이 되는 데이터의 범위, 품질 측정 대상이 되는 대상 기계 학습 모델의 상세 정보, 대상 기계 학습 모델의 성능 측정을 위한 대상 지표에 대한 정보를 수신한다(S620). 그리고, 이와 같이 수신된 정보에 따라, 성능 측정의 대상이 되는 기계 학습 모델과 성능 지표를 구체적으로 설정한다.
- [0052] 다음으로, 데이터 품질 조정 프로그램이 실행하는 제 2 인터페이스 블록을 통해, 각 대상 기계 학습 모델별로 데이터 품질 문제의 종류, 데이터 품질 조정을 위해 사용된 해결 방법의 조합 및 품질이 조정된 데이터에 기초한 대상 기계 학습 모델에 대한 대상 지표에서의 성능 측정 결과를 출력한다(S630).
- [0053] 이를 위해, 데이터 시각적 분석 장치(100)는 각각의 해결 방법의 조합을 대상 데이터에 적용하여 데이터를 수정하고, 수정된 데이터를 이용하여 각 기계 학습 모델을 새롭게 구축하고, 갱신된 기계 학습 모델에 대하여 앞서 설정된 대상 지표를 이용하여 성능 측정을 수행한다. 본 발명에서 사용하는 기계 학습 모델과 대상 지표는 모두 종래에 알려진 기술에 해당하므로, 해당 기술을 이용하여 각 기계 학습 모델을 새롭게 구축하고 성능을 평가하는 절차의 수행이 가능하다.
- [0054] 다음으로, 데이터 품질 조정 프로그램이 제 3 인터페이스 블록을 통해 해결 방법의 조합을 수행하는 단계마다 기계 학습 모델의 성능과 데이터 품질의 변화하는 상태를 출력한다(S640).
- [0055] 앞서, 도 2를 통해 설명한 2차원 그래프나, 막대 그래프 및 밀도 분포 그래프를 통해, 데이터 품질이 향상 되는 단계별로 기계 학습 모델의 성능이 변화하는 것을 확인할 수 있다.
- [0056] 추가적으로, 데이터 품질 조정 프로그램이 제 4 인터페이스 블록을 사용자가 데이터 품질 조정을 위해 사용된 해결 방법의 조합을 맞춤형으로 수정할 수 있도록 하고, 수정된 설정에 따라 그 결과를 출력한다(S650).
- [0057] 본 발명의 일 실시예에 따른 방법은 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 판독가능 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.
- [0058] 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.
- [0059] 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.
- [0060] 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

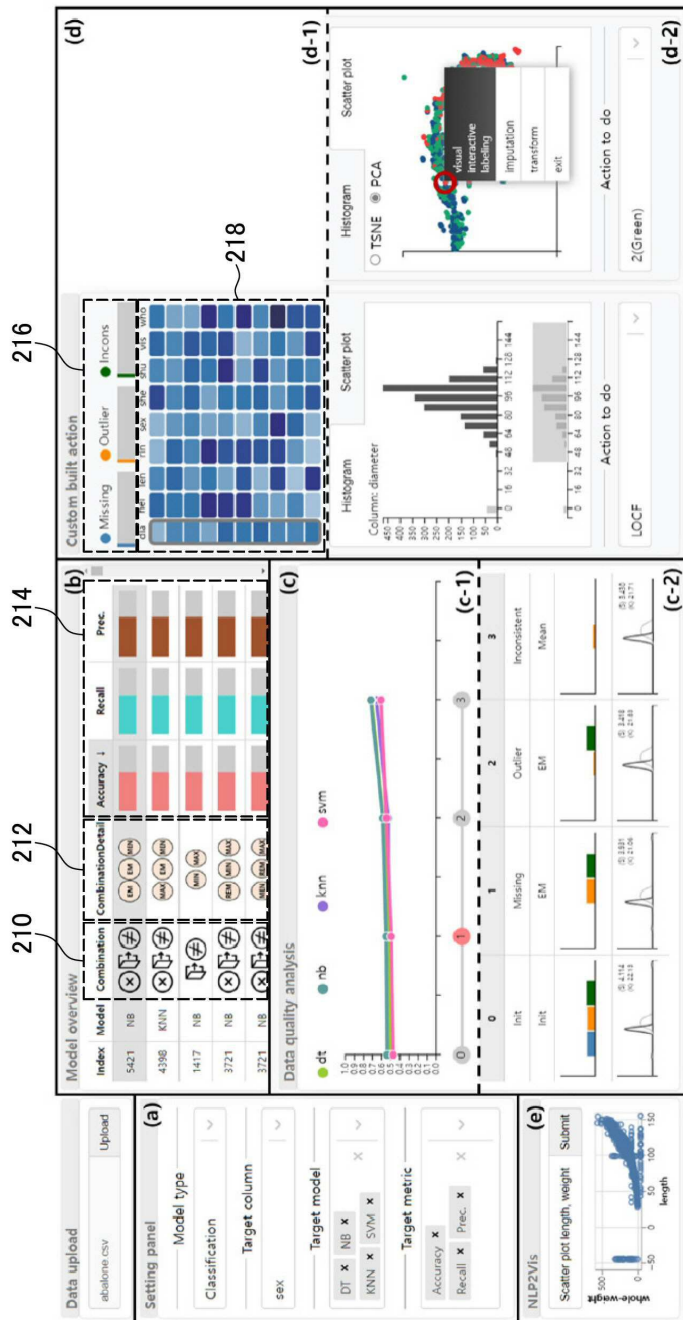
- [0061]
- 100: 시각적 분석 장치
 - 110: 메모리
 - 120: 통신 모듈
 - 130: 프로세서
 - 140: 데이터베이스

도면

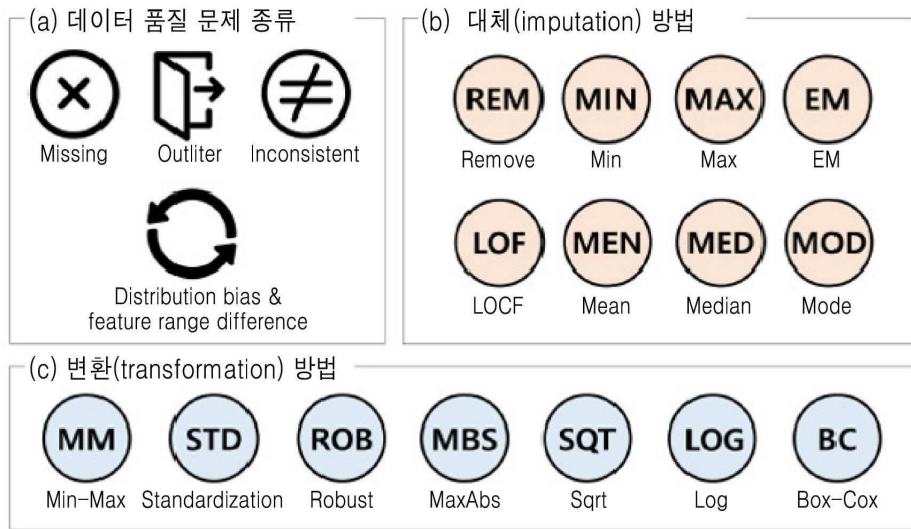
도면1



도면2



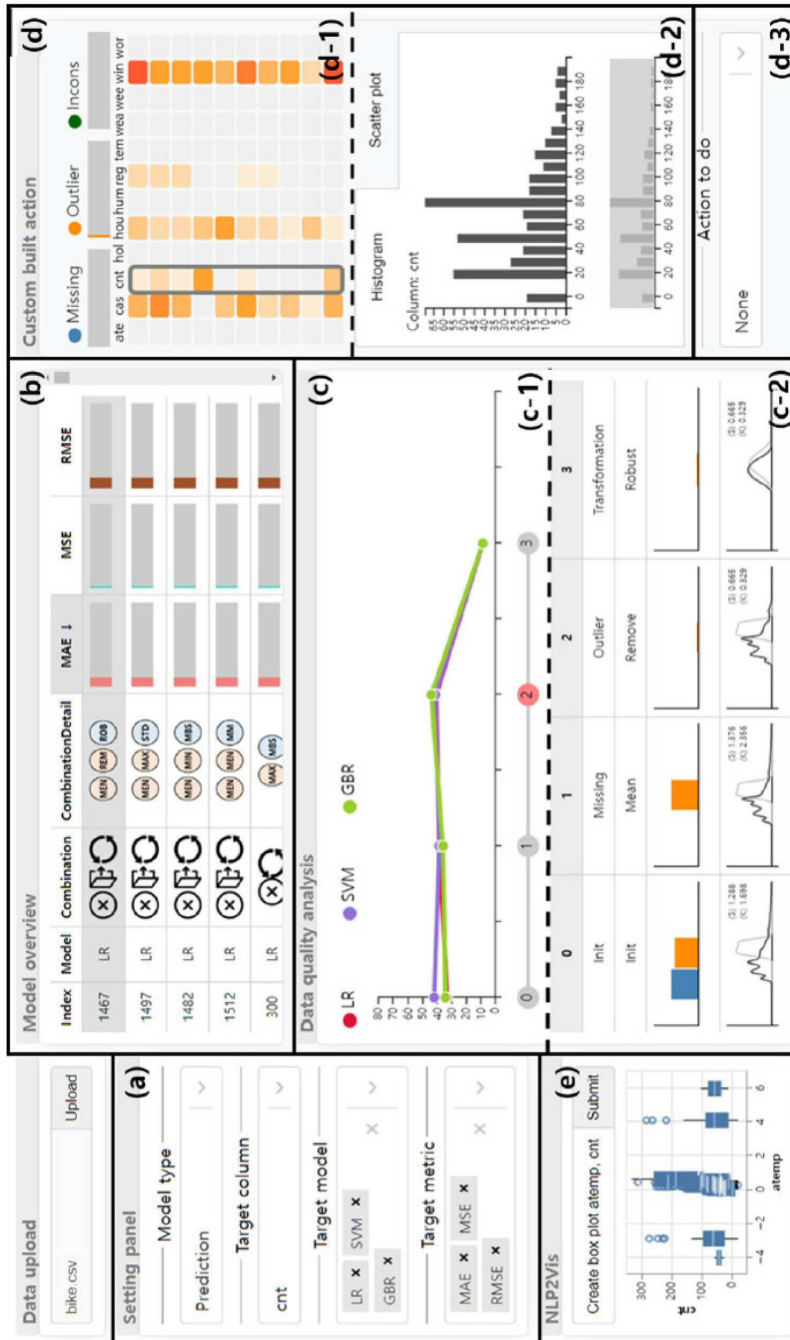
도면3



도면4

대체(Imputation)		Remove, Min, Max, Mean, Median, Mode, LOCF, EM
변환(Transformation)		Min-Max, Standardization, Robust, MaxAbs, Sqrt, Log, Box-Cox
차원축소(Dimension reduction)		TSNE, PCA
예측 (Prediction)	모델(model)	LR, LASSO, RIDGE, EN, LAR, LLAR, OMP, BR, ARD, PAR, RANSAC, TR, HUBER, KR, SVM, KNN, DT, RF, ET, ADA, GBR, MLP, XGBOOST, LIGHTGBM, CATBOOST
	지표(metric)	MAE, MSE, RMSE, R2, RMSLE, MAPE
분류 (Classification)	모델(model)	LR, KNN, NB, DT, SVM, RBFSVM, GPC, MLP, RIDGE, RF, QDA, ADA, GBC, LDA, ET, XGBOOST, LIGHTGBM, CATBOOST
	지표(metric)	Accuracy, AUC, Recall, Precision, F1, Kappa, MCC

도면5



도면6

