



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0029111
(43) 공개일자 2021년03월15일

(51) 국제특허분류(Int. Cl.)
G06Q 40/02 (2012.01) G06N 20/00 (2019.01)
(52) CPC특허분류
G06Q 40/025 (2013.01)
G06N 20/00 (2019.01)
(21) 출원번호 10-2020-0112569
(22) 출원일자 2020년09월03일
심사청구일자 2020년09월03일
(30) 우선권주장
1020190110179 2019년09월05일 대한민국(KR)

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이성환
서울특별시 강남구 언주로30길 13, B동 2506호 (도곡동, 대림아크로빌)
강신한
서울특별시 강남구 언주로85길 32, 1239호 (역삼동)
(뒷면에 계속)
(74) 대리인
특허법인엠에이피에스

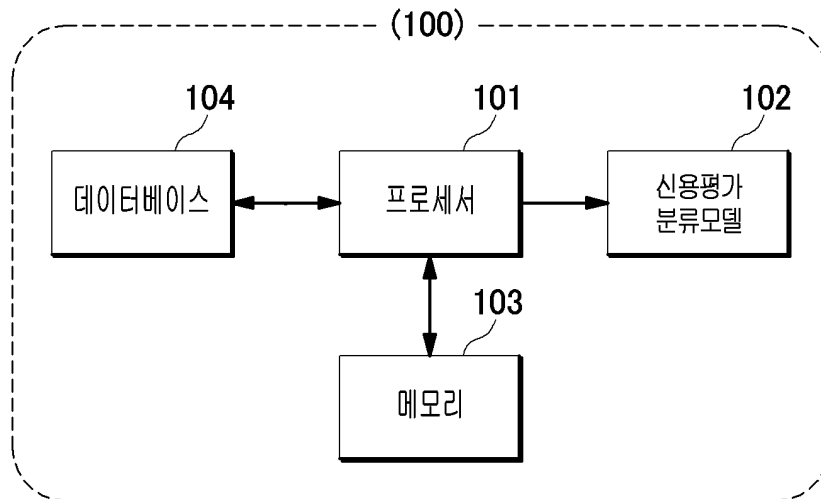
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치 및 방법

(57) 요약

본 발명의 일 측면에 따른 컴퓨터 장치에 의해 수행되는, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법은 (a) 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받는 단계; (b) 제1고객정보에 대하여 신용평가 분류 모델이 판단한 클래스를 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하는 단계; (c) 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하는 단계; (d) 제2고객정보의 입력에 대한 신용평가 분류 모델의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하는 단계; 및 (e) 제2고객정보가 목표 클래스로 판단되는지 여부를 비교하는 단계;를 포함한다.

대표도 - 도1



(72) 발명자

고한빈

경기도 광명시 사성로 91, 106동 1601호 (철산동,
두산위브아파트)

정홍규

서울특별시 성북구 안암로7길 44, 401호 (안암동4
가)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711102926
과제번호	2017-0-01779-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	혁신성장동력프로젝트(R&D)
연구과제명	의사결정 이유를 설명할 수 있는 인간 수준의 학습·추론 프레임워크 개발
기여율	1/2
과제수행기관명	한국과학기술원
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711108635
과제번호	2019-0-00079-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	인공지능대학원지원(고려대학교)
기여율	1/2
과제수행기관명	고려대학교산학협력단
연구기간	2020.01.01 ~ 2020.12.31

명세서

청구범위

청구항 1

컴퓨터 장치에 의해 수행되는, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법에 있어서,

- (a) 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받는 단계;
- (b) 상기 제1고객정보에 대하여 상기 신용평가 분류 모델이 판단한 클래스를 상기 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하는 단계;
- (c) 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하는 단계;
- (d) 상기 제2고객정보의 입력에 대한 상기 신용평가 분류 모델의 평가 결과로서 상기 목표 클래스가 도출되는 결과 확률을 산출하는 단계; 및
- (e) 상기 제2고객정보가 상기 목표 클래스로 판단되는지 여부를 비교하는 단계;를 포함하는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법.

청구항 2

제 1항에 있어서,

상기 (e) 단계에서

상기 제2고객정보가 상기 목표 클래스로 판단되지 않는 경우, 상기 특정 항목을 선택하는 과정부터 순차적으로 반복 수행하되, 반복될 때 마다 상기 특정 항목의 개수를 한 개씩 추가하는 것이고,

상기 제2고객정보가 상기 목표 클래스로 판단된 경우, 상기 제2 고객정보의 변경된 항목들을 검출하여 제공하는 단계를 더 포함하는 것인, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법.

청구항 3

제 2항에 있어서,

상기 (b) 단계는

상기 목표 클래스가 도출되는 결과 확률에 기초하여 상기 제2고객정보에 대한 미분의 절대값이 큰 순서대로 상기 특정 항목을 선택하되,

상기 항목의 개수만큼 상기 미분의 절대값이 큰 순서대로 인덱싱을 수행하는 것인, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법.

청구항 4

제 3항에 있어서,

상기 (c) 단계는

상기 인덱싱된 항목에 대하여 상기 제1고객정보의 값을 상기 제2고객정보의 항목값으로 변경하되, 상기 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 상기 변경된 항목값을 결정하는 것인, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법.

청구항 5

기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치에 있어서,

기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램이 저장된 메모리;
및

상기 프로그램을 수행하는 프로세서;를 포함하며,

상기 프로세서는, 상기 프로그램의 실행에 따라,

복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받고,

상기 제1고객정보에 대하여 상기 신용평가 분류 모델이 판단한 클래스를 상기 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하고,

선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하고,

상기 제2고객정보의 입력에 대한 상기 신용평가 분류 모델의 평가 결과로서 상기 목표 클래스가 도출되는 결과 확률을 산출하고,

상기 제2고객정보가 상기 목표 클래스로 판단되는지 여부를 비교하는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치.

청구항 6

제 5항에 있어서,

상기 프로세서는

상기 제2고객정보가 상기 목표 클래스로 판단되지 않는 경우, 상기 특정 항목을 선택하는 과정부터 순차적으로 반복 수행하되, 반복될 때 마다 상기 특정 항목의 개수를 한 개씩 추가하는 것이고,

상기 제2고객정보가 상기 목표 클래스로 판단된 경우, 상기 제2 고객정보의 변경된 항목들을 검출하여 제공하는 것인, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치.

청구항 7

제 6항에 있어서,

상기 프로세서는

상기 목표 클래스가 도출되는 결과 확률에 기초하여 상기 제2고객정보에 대한 미분의 절대값이 큰 순서대로 상기 특정 항목을 선택하되,

상기 항목의 개수만큼 상기 미분의 절대값이 큰 순서대로 인덱싱을 수행하는 것인, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치.

청구항 8

제 7항에 있어서,

상기 프로세서는

상기 인덱싱된 항목에 대하여 상기 제1고객정보의 값을 상기 제2고객정보의 항목값으로 변경하되, 상기 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 상기 변경된 항목값을 결정하는 것인, 기계학

습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치.

발명의 설명

기술 분야

[0001] 본 발명은 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치 및 방법 에 대한 것이다.

배경 기술

[0002] 최근 기계학습 기술이 크게 발전하면서 금융 분야에서 대출과 관련되어 개인 및 기업의 신용평가를 위해 기계학습 모델을 적용하려고 하는 시도가 증가하였다. 금융 분야에서 기계학습 기반의 신용평가 분류 모델을 적용하는 것은 기존처럼 금융 분야의 전문가에 의해 신용평가가 이루어지는 것이 아닌 컴퓨터가 스스로 학습하고 데이터 특징을 분석하여 자동적으로 분류 가능하다는 것을 말한다. 또한 기계학습 기반 분류 모델은 항목간의 관계성을 알아낼 수도 있으며 복잡한 데이터로부터의 추론이 가능해졌으므로 높은 성능까지 도출 가능하다.

[0003] 그러나 동시에 학습해야 할 파라미터 수와 모델의 복잡성도 모두 비약적으로 증가함에 따라서 인간이 모델의 분류 결정 이유에 대해 이해할 수 없다는 치명적인 단점을 가지게 된다. 사람이 이해할 수 없다는 것은 신뢰성이 중요한 분야 또는 기업에서는 사용할 수 없다는 것을 의미한다. 특히 금융 분야 안에서 신용 평가 결정에 대한 문제는 기업의 이익과 관련된 직결되는 문제이므로 신용 평가 모델의 높은 성능과 관계없이 모델의 결정을 신뢰할 수 없다면 적용하기 힘들다.

[0004] 그래서 기계학습 기반 모델의 결정에 대해 이해할 수 있도록 하는 설명 가능한 인공지능에 관한 연구가 최근에 주목받고 있다. 주로 이러한 연구에서는 특정 클래스에 대해서 모델의 결정을 도출하는데 있어서 영향을 크게 미치는 항목을 찾는 방법을 사용하고 있다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 등록특허공보 제 10-2005734 호 (발명의 명칭: 머신 러닝 기반 유형자산 정량 평가 결과가 반영된 종합 금융 서비스 제공 시스템)

발명의 내용

해결하려는 과제

[0006] 본 발명은 전술한 문제점을 해결하기 위한 것으로, 목표하는 클래스로 판단되기 위한 효과적인 고객의 신용 정보를 찾아내고 해당 정보의 값을 어떻게 변경시켜야 신용평가 결과가 뒤바뀌는지 판단하는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치 및 방법을 제공하는 것을 기술적 과제로 한다.

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

과제의 해결 수단

[0008] 상술한 기술적 과제를 해결하기 위한 기술적 수단으로서, 본 발명의 일 측면에 따른 컴퓨터 장치에 의해 수행되는, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법은 (a) 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받는 단계; (b) 제1고객정보에 대하여 신용평가 분류 모델이 판단한 클래스를 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하는 단계; (c) 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하는 단계; (d) 제2고객정보의 입력에 대한 신용평가 분류 모델의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하는 단계; 및 (e) 제2고객정보가 목표 클래스로 판단되는지 여부를 비교하는 단계;를 포함한다.

[0009] (e) 단계에서 제2고객정보가 목표 클래스로 판단되지 않는 경우, 특정 항목을 선택하는 과정부터 순차적으로 반

복 수행하되, 반복될 때 마다 특정 항목의 개수를 한 개씩 추가하는 것이고, 제2고객정보가 목표 클래스로 판단된 경우, 제2 고객정보의 변경된 항목들을 검출하여 제공하는 단계를 더 포함한다.

[0010] (b) 단계는 목표 클래스가 도출되는 결과 확률에 기초하여 제2고객정보에 대한 미분의 절대값이 큰 순서대로 특정 항목을 선택하되, 항목의 개수만큼 미분의 절대값이 큰 순서대로 인덱싱을 수행한다.

[0011] (c) 단계는 인덱싱된 항목에 대하여 제1고객정보의 값을 제2고객정보의 항목값으로 변경하되, 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 변경된 항목값을 결정한다.

[0012] 본 발명의 다른 측면에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램이 저장된 메모리; 및 프로그램을 수행하는 프로세서;를 포함하며, 프로세서는, 프로그램의 실행에 따라, 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받고, 제1고객정보에 대하여 신용평가 분류 모델이 판단한 클래스를 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하고, 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하고, 제2고객정보의 입력에 대한 신용평가 분류 모델의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하고, 제2고객정보가 목표 클래스로 판단되는지 여부를 비교한다.

[0013] 프로세서는 제2고객정보가 목표 클래스로 판단되지 않는 경우, 특정 항목을 선택하는 과정부터 순차적으로 반복 수행하되, 반복될 때 마다 특정 항목의 개수를 한 개씩 추가하는 것이고, 제2고객정보가 목표 클래스로 판단된 경우, 제2 고객정보의 변경된 항목들을 검출하여 제공한다.

[0014] 프로세서는 목표 클래스가 도출되는 결과 확률에 기초하여 제2고객정보에 대한 미분의 절대값이 큰 순서대로 특정 항목을 선택하되, 항목의 개수만큼 미분의 절대값이 큰 순서대로 인덱싱을 수행한다.

[0015] 프로세서는 인덱싱된 항목에 대하여 제1고객정보의 값을 제2고객정보의 항목값으로 변경하되, 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 변경된 항목값을 결정한다.

발명의 효과

[0016] 전술한 본원의 과제 해결 수단 중 어느 하나에 의하면, 기계학습 기반 신용평가 분류 모델이 생각하기에 대출 가능/불가 클래스 사이의 분별력있는 항목을 찾아냄으로써 모델의 행동을 이해할 뿐만 아니라 모델이 집중적으로 고려하는 항목이 무엇인지 확인 가능하므로 추후 새로운 신용 평가 모델의 개발과정에서 중요히 고려해야 할 특징을 알 수 있다.

[0017] 또한 다수의 고객 데이터에서 목표하는 클래스로 판별하는데 사용되는 공통적인 요소가 있다면 이후 모델링할 경우에도 해당 항목을 신용평가 분류 모델로 하여금 중요한 특징으로 고려되도록 제한하는 방식을 사용하여 기존보다 개선된 모델을 구축하는데 도움을 줄 수 있다.

[0018] 더불어, 신용 정보 중 해당 항목의 값을 변경시켜 목표하는 클래스로 판단되도록 하는 기술은 대출을 해주는 기업과 대출을 하는 고객 모두에게 이익을 줄 수 있다. 예를 들어 한 고객 정보에 대해서 신용평가 분류 모델이 대출 불가라고 한 경우에 대출을 해주는 기업입장에서는 고객에게 어떠한 항목의 값을 적절하게 변화시킬 경우 대출 가능하다고 설명가능하다. 즉, 신용이 높은 고객에게 대출을 많이 해줄수록 기업의 이익이 창출되는 기업 입장에서는 이는 좋은 효과로 작용될 수 있다. 또한 대출하려는 고객입장에서는 자신의 신용 정보 중 어떤 항목에 대한 값이 대출을 하는데 있어서 부정적으로 작용하는 지 알 수 있으므로 대출을 하려면 고려해야 할 항목에 대한 정보를 획득가능하다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치의 구성을 도시한 블록도이다.

도 2는 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치의 세부 모듈을 도시한 블록도이다.

도 3은 본 발명의 일 실시예에 따른 고객A의 신용정보와 이에 대한 신용평가 분류 모델에 따른 결과를 도시한 도면이다.

도 4는 본 발명의 일 실시예에 따른 신용평가 분류 모델에 따라 고객A의 신용정보 중 목표 클래스로 도출하기

위해, 신용정보 중 변경되어야 하는 항목을 도기한 도면이다.

도 5는 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 도시한 순서도이다.

발명을 실시하기 위한 구체적인 내용

- [0020] 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.
- [0021] 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.
- [0022] 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.
- [0023] 이하에서 언급되는 "기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치"는 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.
- [0024] 이하 첨부된 도면을 참고하여 본 발명의 일 실시예를 상세히 설명하기로 한다.
- [0025] 도 1은 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치의 구성을 도시한 블록도이다.
- [0026] 도1을 참조하면 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치(100)는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램이 저장된 메모리(103) 및 프로그램을 수행하는 프로세서(101)를 포함하며, 프로세서(101)는 프로그램의 실행에 따라, 복수의 항목으로 구성된 신용정보를 포함하는 제1고객정보와 신용평가 분류 모델(102)의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받고, 제1고객정보에 대하여 신용평가 분류 모델(102)이 판단한 클래스를 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하고, 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하고, 제2고객정보의 입력에 대한 신용평가 분류 모델(102)의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하고, 제2고객정보가 목표 클래스로 판단되는지 여부를 비교한다.
- [0027] 따라서, 본 발명의 기계학습 기반 신용평가 분류 모델(102)의 평가 이유를 설명하는 장치(100)는 어떠한 고객정보(제1고객정보)가 주어졌고, 해당 정보(예를 들어, 연봉, 과거 파산 경험, FICO 점수, 대출 신청 금액 등을 포함하는 복수의 신용정보의 항목)를 바탕으로 신용평가 분류 모델이 A라는 클래스로 판단하였을 때, 목표 클래스인 B로 판단되도록 원래의 고객정보(제1고객정보)에서 변경이 필요한 고객정보(제2고객정보)를 제공할 수

있다. 이를 통해 본 발명의 신용평가 분류 모델이 기존 클래스 A와 목표 클래스B사이의 분별력있는 특징적인 신용정보 항목이 무엇인지 해석을 제공할 수 있다.

- [0028] 다시 말해서, 본 발명은 신용평가 분류 모델의 전체적인 해석보다는 주어진 고객 정보(복수의 신용 정보 항목)를 입력으로 하여 산출된 모델의 결과(목표 클래스)에 대해 설명을 제공하고자 한다. 이때 설명의 목적은 목표 클래스로 판별하는데 있어서 중요한 신용 정보의 항목을 시각화 하며 대상 항목의 값을 변경시켜 목표하는 클래스로 판단되도록 하는 방법을 제안하는 것이다.
- [0029] 즉, 본 발명은 정확하게 특정 고객에 대해 어떠한 항목을 변경하여야 목표 클래스로 판단되는지 설명함으로써 모델의 행동을 이해하고 해석하는데 있어서 효과적인 기술이다.
- [0030] 도1에 도시된 바와 같이 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치(100)는 메모리(103), 프로세서(101), 데이터베이스(104) 및 신용평가 분류 모델(102)을 포함할 수 있다.
- [0031] 메모리(103)에는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램이 저장된다. 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램은 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받고, 제1고객정보에 대하여 신용평가 분류 모델이 판단한 클래스를 목표 클래스로 변경하는데 영향을 주는 특정 항목을 선택하고, 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하고, 제2고객정보의 입력에 대한 신용평가 분류 모델의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하고, 제2고객정보가 목표 클래스로 판단되는지 여부를 비교한다.
- [0032] 신용평가 분류 모델(102)은 고객의 신용정보를 입력으로 인공신경망을 통해 학습되어 대출 여부를 예측하는 것이다.
- [0033] 이러한 메모리(103)에는 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치(100)의 구동을 위한 운영 체제나 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램의 실행 과정에서 발생하는 여러 종류가 데이터가 저장된다.
- [0034] 이때, 메모리(103)는 전원이 공급되지 않아도 저장된 정보를 계속 유지하는 비휘발성 저장장치 및 저장된 정보를 유지하기 위하여 전력이 필요한 휘발성 저장장치를 통칭하는 것이다.
- [0035] 또한, 메모리(103)는 프로세서(101)가 처리하는 데이터를 일시적 또는 영구적으로 저장하는 기능을 수행할 수 있다. 여기서, 메모리(103)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.
- [0036] 프로세서(101)는 메모리(103)에 저장된 프로그램을 실행하되, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 수행하기 위한 프로그램의 실행에 따르는 전체 과정을 제어한다. 프로세서(101)가 수행하는 각각의 동작에 대해서는 추후 보다 상세히 살펴보기로 한다.
- [0037] 이러한 프로세서(101)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.
- [0038] 데이터베이스(104)는 프로세서(101)의 제어에 따라, 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치(100)에 필요한 데이터를 저장 또는 제공한다. 이러한 데이터베이스(104)는 메모리(103)와는 별도의 구성 요소로서 포함되거나, 또는 메모리(103)의 일부 영역에 구축될 수도 있다.
- [0039] 도 2는 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치의 세부 모듈을 도시한 블록도이다.
- [0040] 한편, 프로세서(101)는 메모리(103)에 저장된 프로그램의 실행에 따라 다양한 기능을 수행할 수 있는데, 각 기능에 따라 프로세서(101)에 포함되는 세부 모듈들을 입력부(110), 항목 선택부(120), 항목값 변경부(130), 분류 모델 결과부(140), 비교부(150) 및 설명부(160)로 나타낼 수 있다.

- [0041] 입력부(110)는 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 사용자로부터 입력받을 수 있다. 예를 들어, 사용자는 입력부(110)를 통해 제1고객정보(개인 또는 기업 고객의 원래 신용정보)와 목표 클래스(대출 가능 또는 대출 불가)를 입력할 수 있다. 이때, 신용정보는 연봉, 과거 파산 경험, FICO 점수, 대출 신청 금액 등의 복수의 항목으로 구성될 수 있다.
- [0042] 항목 선택부(120)는 제1고객정보에 대하여 신용평가 분류 모델(102)이 판단한 클래스를 목표 클래스로 변경하는데 영향을 주는 특정 항목을 선택할 수 있다.
- [0043] 항목값 변경부(130)는 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성할 수 있다.
- [0044] 구체적으로 항목 선택부(120)는 목표 클래스가 도출되는 결과 확률에 기초하여 제2고객정보에 대한 미분의 절대값이 큰 순서대로 특정 항목을 선택하되, 항목의 개수만큼 미분의 절대값이 큰 순서대로 인덱싱을 수행할 수 있다. 여기서 미분의 절대값이 가장 큰 항목의 의미는 해당 항목값의 변경이 다른 항목에 비해 목표 클래스의 확률의 변화를 가장 크게 주는 것을 의미한다. 즉, 항목 선택부(120)는 제1고객정보가 목표 클래스로 판단되는 데 있어서, 영향을 주는 효과적인 신용정보의 항목을 하나씩 증가시키면서 선택하는 역할을 한다.
- [0045] 또한 항목값 변경부(130)는 인덱싱된 항목에 대하여 제1고객정보의 값을 제2고객정보의 항목값으로 변경하되, 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 변경된 항목값을 결정할 수 있다. 여기서 손실 함수는 하기의 [수학식1]과 같다.
- [0046] [수학식1]
- [0047]
$$\left\| \sum_{k=1}^K f_k(\text{변경된 } A \text{ 고객 정보}) - \frac{1}{N} \sum_{i=1}^N f_k(\text{학습 데이터}_{i,t}) \right\|_2 + \lambda \| \text{변경된 } A \text{ 고객 정보} - \text{원래의 } A \text{ 고객 정보} \|_2$$
- [0048] f_k 는 클래스 k 에 대한 로직 점수(소프트 맥스 함수 이전 값), K 는 클래스 수, N 은 하이퍼 파라미터로 트레이닝 데이터의 수, 학습데이터 i , t 는 목표 클래스로 t 로 판별되는 학습데이터, λ 는 하이퍼 파라미터를 의미한다. 본 발명에서는 N 은 100, λ 는 0.3으로 설정한다. 손실 함수의 앞 변의 값을 최소화한다는 것은 제2고객정보(변경된 A 의 고객 정보)의 로직 점수가 목표 클래스로 판단되는 학습 데이터들의 로직 점수와 비슷해지도록 하는 것이다. 뒤 변의 값을 최소화하는 것은 제2고객정보(변경된 A 의 고객 정보)가 제1고객정보(원래의 고객 정보)와 비슷해지도록 하는 것이다.
- [0049] 분류 모델 결과부(140)는 제2고객정보의 입력에 대한 신용평가 분류 모델(102)의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출할 수 있다. 즉, 분류 모델 결과부(140)는 목표 클래스(예를 들어, 대출 가능)를 도출하기 위해서, 제2고객정보(변경된 A 고객 정보)를 입력으로 한 신용평가 분류 모델(102)의 결과 확률값을 산출할 수 있다.
- [0050] 비교부(150)는 제2고객정보가 목표 클래스로 판단되는지 여부를 비교할 수 있다. 비교부(150)는 제2고객정보가 목표 클래스로 판단되지 않는 경우, 특정 항목을 선택하는 과정에서 순차적으로 반복 수행하되, 반복될 때 마다 특정 항목의 개수를 한 개씩 추가할 수 있다. 또한, 비교부(150)는 제2고객정보가 목표 클래스로 판단된 경우, 제2 고객정보의 변경된 항목들을 검출하여 설명부(160)를 통해 제공할 수 있다.
- [0051] 설명부(160)는 제2고객정보가 목표 클래스로 판단되는 데, 제2고객정보의 변경된 항목을 시각화 하여 제공할 수 있다. 또한, 설명부(160)는 기계학습 기반 신용평가 분류 모델의 평가 이유를 표시하기 위한 장치이며, 후술하는 도 3및 도4에 도시된 바와 같이 고객의 신용정보 및 대출 가능 여부를 표시하는 것도 가능하다. 이로 인해, 사용자는 한 고객 정보로부터 원래의 클래스(제1고객정보에 따른 결과)와 목표 클래스(제2고객정보에 따른 결과)사이의 분별력있는 특징(변경된 신용 정보의 항목)이 무엇인지 확인할 수 있다.
- [0052] 도 3은 본 발명의 일 실시예에 따른 고객A의 신용정보와 이에 대한 신용평가 분류 모델에 따른 결과를 도시한 도면이다. 도 4는 본 발명의 일 실시예에 따른 분류 모델에 따라 고객A의 신용정보 중 목표 클래스로 도출하기 위해, 신용정보 중 변경되어야 하는 항목을 도시한 도면이다.
- [0053] 예시적으로 도 3에 도시된 바와 같이, 제1고객정보(10)에 대한 신용평가 분류 모델(102)의 분류 결과는 대출 불가 상태(대출 가능: 40%, 대출 불가: 60%)로 도출될 수 있다. 이때 제1고객정보는 연봉: 3,000만원, 부채: 2,000만원, 대출 금액: 7,000만원, 신용 거래 횟수 5회를 포함하는 원래의 고객A의 신용정보이다.

- [0054] 예시적으로, 도4를 참조하면 프로세서(101)는 상기 제1고객정보(10)와 목표 클래스가 대출 가능으로 입력되면 목표 클래스(대출 가능)로 변경하는 데 영향을 주는 특정 항목으로, 신용 정보 중 연봉 및 부채를 선택하고, 각 항목의 값을 변경(연봉: 6,000만원, 부채: 1,000만원)한 제2고객정보(20)으로 목표 클래스(대출 가능)가 도출되는 결과(대출 가능: 90%)를 산출하고, 제2고객정보(20)가 목표 클래스로 판단된 경우, 제2고객정보(20)의 변경된 항목(연봉, 부채)을 제공할 수 있다.
- [0055] 즉, 사용자는 제2고객정보(20)의 변경된 항목을 통해 신용평가 분류 모델의 평가 이유를 파악할 수 있다. 고객 A의 경우, 기존에는 연봉이 3,000만원이고 부채 금액이 2,000만원(제1고객정보)이어서 목표 클래스(대출 가능)로 판단되는 데 방해되었다. 그러나, 고객 A의 신용 정보 중 연봉을 6,000만원으로 올리고, 부채 금액을 1,000만원 줄일 경우, 신용평가 분류 모델(102)이 대출 가능이라고 판단할 결과가 90%로 바뀐다는 것을 의미한다.
- [0056] 이에 따라, 대출해주는 기업측에서는 고객에게 대출이 가능한 조건을 제시할 수 있어, 대출 실적을 증가시킬 수 있는 동기부여가 되고, 고객측에서도 본인의 신용 정보 중 어떤 부분이 대출하는데 부족한지 정확하게 알 수 있다는 효과가 있다.
- [0057] 이하에서는 상술한 도1 내지 도4에 도시된 구성 중 동일한 기능을 수행하는 구성의 경우 설명을 생략하기로 한다.
- [0058] 도 5는 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법을 도시한 순서도이다.
- [0059] 도 5를 참조하면, 본 발명의 일 실시예에 따른 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 방법은 복수의 항목으로 구성된 신용 정보를 포함하는 제1고객정보와 신용평가 분류 모델의 평가 결과로서 도출하고자 하는 목표 클래스를 입력받는 단계(S110), 제1고객정보에 대하여 신용평가 분류 모델이 판단한 클래스를 목표 클래스로 변경하는 데 영향을 주는 특정 항목을 선택하는 단계(S120), 선택된 특정 항목의 값을 변경하고, 변경된 항목값을 포함하는 제2 고객정보를 생성하는 단계(S130), 제2고객정보의 입력에 대한 신용평가 분류 모델의 평가 결과로서 목표 클래스가 도출되는 결과 확률을 산출하는 단계(S140) 및 제2고객정보가 목표 클래스로 판단되는지 여부를 비교하는 단계(S150)를 포함한다.
- [0060] S150단계에서 제2고객정보가 목표 클래스로 판단되지 않는 경우, 특정 항목을 선택하는 과정부터 순차적으로 반복 수행하되, 반복될 때 마다 특정 항목의 개수를 한 개씩 추가하는 것이고, 제2고객정보가 목표 클래스로 판단된 경우, 제2 고객정보의 변경된 항목들을 검출하여 제공하는 단계를 더 포함할 수 있다.
- [0061] S120단계는 목표 클래스가 도출되는 결과 확률에 기초하여 제2고객정보에 대한 미분의 절대값이 큰 순서대로 특정 항목을 선택하되, 항목의 개수만큼 미분의 절대값이 큰 순서대로 인덱싱을 수행할 수 있다.
- [0062] S130단계는 인덱싱된 항목에 대하여 제1고객정보의 값을 제2고객정보의 항목값으로 변경하되, 제1및 제2고객정보에 기초하여 기 설정된 손실 함수의 값을 최소화하여 변경된 항목값을 결정할 수 있다.
- [0063] 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.
- [0064] 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.
- [0065] 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.
- [0066] 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

로 해석되어야 한다.

부호의 설명

[0067] 100: 기계학습 기반 신용평가 분류 모델의 평가 이유를 설명하는 장치

101: 프로세서

102: 신용평가 분류 모델

103: 메모리

104: 데이터베이스

110: 입력부

120: 항목 선택부

130: 항목값 변경부

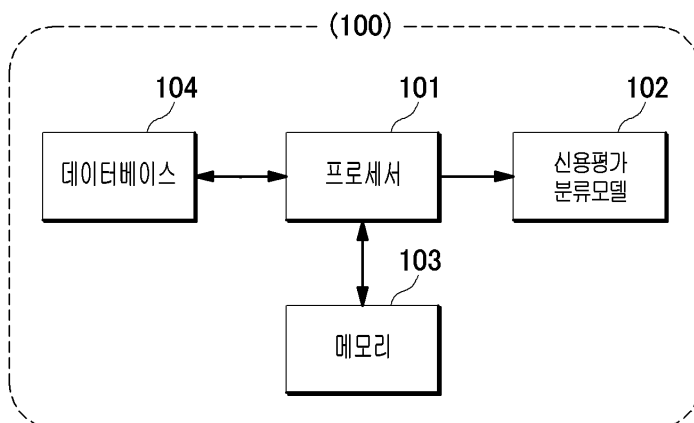
140: 분류 모델 결과부

150: 비교부

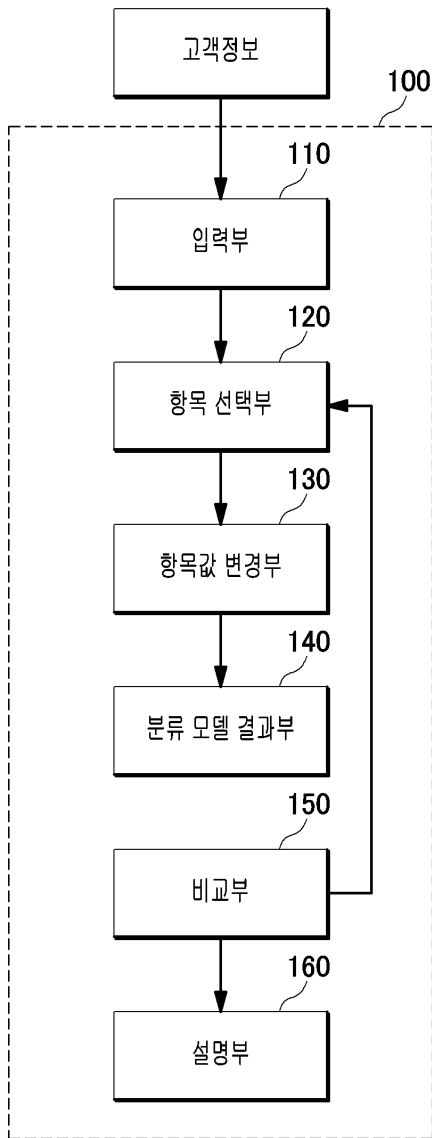
160: 설명부

도면

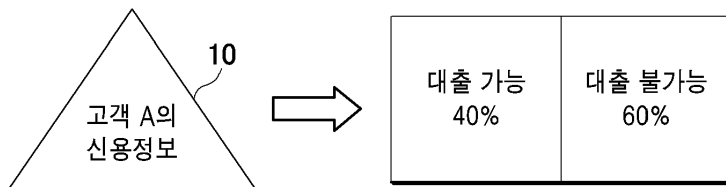
도면1



도면2



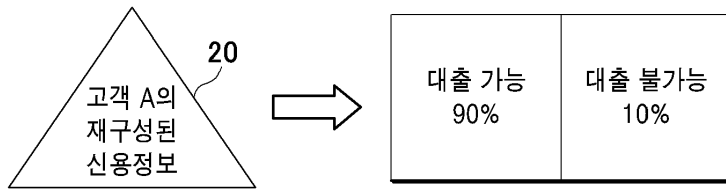
도면3



고객 A의 신용정보

항목	연봉 (기준 : 만원)	부채 (기준 : 만원)	대출 금액 (기준 : 만원)	신용 거래 횟수
값	3,000	2,000	7,000	5

도면4



설명 모델을 통한 재구성된 고객 A의 신용정보

항목	연봉 (기준 : 만원)	부채 (기준 : 만원)	대출 금액 (기준 : 만원)	신용 거래 횟수
값	6,000	1,000	7,000	5

도면5

