



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0149730
(43) 공개일자 2024년10월15일

(51) 국제특허분류(Int. Cl.)

G16B 40/20 (2019.01) G06N 3/04 (2023.01)
G06N 3/08 (2023.01) G16B 20/20 (2019.01)
G16B 30/00 (2019.01) G16B 50/40 (2019.01)
G16B 50/50 (2019.01) G16H 10/60 (2018.01)
G16H 50/30 (2018.01)

(52) CPC특허분류

G16B 40/20 (2019.02)
G06N 3/04 (2023.01)

(21) 출원번호 10-2023-0045687

(22) 출원일자 2023년04월06일

심사청구일자 2023년04월06일

(71) 출원인

울산과학기술원

울산광역시 울주군 언양읍 유니스트길 50

(72) 발명자

이정혜

울산광역시 울주군 언양읍 유니스트길 50 (반연리, 울산과학기술원)

이명훈

울산광역시 울주군 언양읍 유니스트길 50 (반연리, 울산과학기술원)

(뒷면에 계속)

(74) 대리인

특허법인 수

전체 청구항 수 : 총 26 항

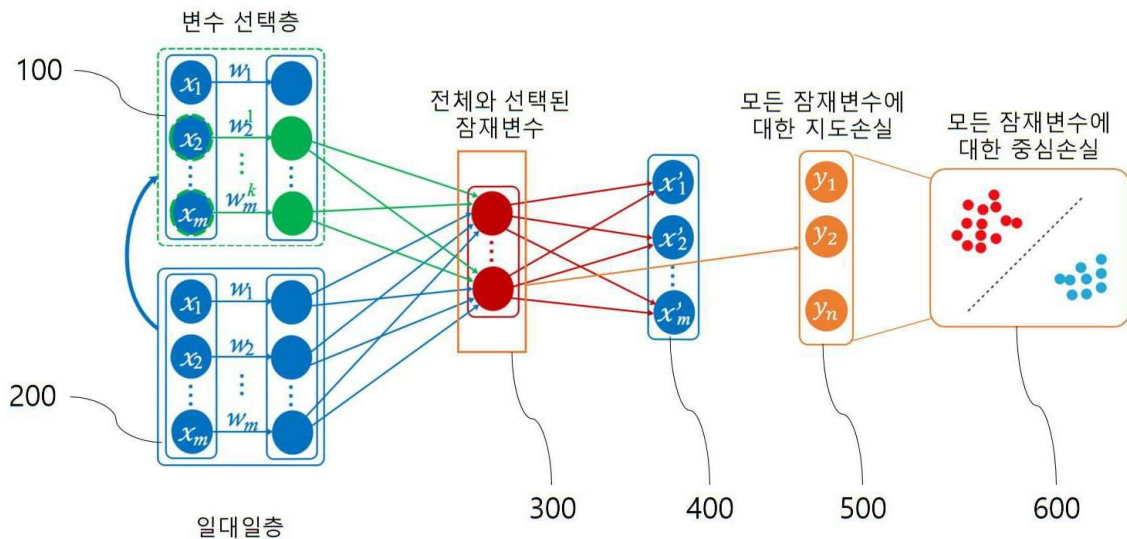
(54) 발명의 명칭 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법

(57) 요약

본 발명은 SNP 변이정보의 비선형적인 관계를 고려하고 질병 유무에 대한 정보를 반영하여 유전적 위험 점수의 실효성을 높이며, 해석이 가능하도록 유전적 위험 점수에 영향을 미치는 중요 바이오마커를 선택해낼 수 있는 방법까지 고안한 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법에 관한 것이다.

본 발명의 딥러닝을 활용한 유전적 위험 점수 산출 장치는 전처리가 수행된 SNP 일부에 대해 일대일층으로 구성된 제 1 입력층, SNP 전체에 대해 일대일층으로 구성된 제 2 입력층, 제 1 입력층 및 제 2 입력층을 인코딩하여 특정 질병의 잠재변수를 나타낸 은닉층, 및 제 1 입력층 과 제 2 입력층이 복원되도록 은닉층을 디코딩한 출력층으로 이루어진다.

대표도



(52) CPC특허분류

G06N 3/08 (2023.01)
G16B 20/20 (2019.02)
G16B 30/00 (2019.02)
G16B 50/40 (2019.02)
G16B 50/50 (2019.02)
G16H 10/60 (2021.08)
G16H 50/30 (2018.01)

김수현

울산광역시 울주군 언양읍 유니스트길 50 (반연리,
울산과학기술원)

(72) 발명자

임동철

울산광역시 울주군 언양읍 유니스트길 50 (반연리,
울산과학기술원)

손원호

울산광역시 울주군 언양읍 유니스트길 50 (반연리,
울산과학기술원)

이 발명을 지원한 국가연구개발사업

과제고유번호	1425157253
과제번호	P0016191
부처명	중소벤처기업부
과제관리(전문)기관명	한국산업기술진흥원
연구사업명	규제자유특구실증기반조성(R&D)
연구과제명	지능형 바이오 오믹스 분석기술 개발
기 여 율	50.00/100
과제수행기관명	울산과학기술원
연구기간	2022.01.01 ~ 2023.03.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711165204
과제번호	2020R1C1C1011063
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	딥러닝 기반의 개인정보보호가 보장되는 인공지능 연합학습 플랫폼 개발
기 여 율	25.00/100
과제수행기관명	울산과학기술원
연구기간	2022.03.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1485019478
과제번호	ARQ202101403003
부처명	환경부
과제관리(전문)기관명	한국환경산업기술원
연구사업명	환경보건 디지털 조사 기반 구축 기술개발 사업
연구과제명	GIS 기반 빅데이터 및 기계학습을 활용한 지역단위 환경보건 감시체계 평가 시스템
개발	
기 여 율	25.00/100
과제수행기관명	울산과학기술원
연구기간	2021.04.01 ~ 2024.12.31

명세서

청구범위

청구항 1

전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부에 대해 일대일층(one-to-one layer)으로 구성된 제 1 입력층;

SNP 전체에 대해 일대일층으로 구성된 제 2 입력층;

상기 제 1 입력층 및 상기 제 2 입력층을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 나타낸 은닉층; 및

상기 제 1 입력층 및 상기 제 2 입력층이 복원되도록 상기 은닉층을 디코딩한 출력층을 포함하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 2

제 1항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 변이정보, 나이, 성별이 주어졌을 때 질병에 걸릴 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 3

제 1항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 변이정보가 주어졌을 때 질병에 걸릴 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 4

제 1항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 나이, 성별이 주어졌을 때 변이정보에 대한 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 5

제 1항에 있어서,

상기 일대일층은, 신경망 구조(Neural network)에서 SNP 각각에 일대일 대응으로 가중치를 연결하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 6

제 1항에 있어서,

상기 잠재변수는, 유전적 위험 점수(Genomic Risk Score)에 질병 유무에 관한 정보를 반영하기 위해 완전 연결 층으로 연결되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 7

제 6항에 있어서,

상기 유전적 위험 점수는, 질병 유무에 대한 정보가 손실함수를 통해 학습된 잠재변수를 토대로 산출되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 8

제 7항에 있어서,

상기 유전적 위험 점수는, 상기 잠재변수의 값을 모두 더한 후 상기 잠재변수의 차원으로 나눈 값으로 산출되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 9

제 1항에 있어서,

상기 손실함수는, 손실함수를 최소화하도록 학습되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

청구항 10

제 9항에 있어서,

상기 손실함수는,

프랙탈 오토인코더(Fractal Autoencoders) 손실로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

$$\mathcal{L}_{FAE} = \min_{W_I, f, g} \|X - f(g(XW_I))\|_F^2 + \lambda_1 \|X - f(g(XW_I^{max_k}))\|_F^2 + \lambda_2 \|W_I\|_1$$

여기서,

$$X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$g = \text{encoder}, \quad f = \text{decoder}$

$\lambda_1, \lambda_2 : \text{weights}, \max_k : \text{pre-defined parameter for biomarker}$

청구항 11

제 9항에 있어서,

상기 손실함수는,

지도항목(Supervised term)손실로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

$$\mathcal{L}_{supervised} = \mathcal{L}_{BCElogit}(y, \psi(g(XW_I)))$$

여기서,

$$X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$$y \in \mathbb{R}^n$$

$g = \text{encoder}, \quad \psi : \text{fully connected layer}$

청구항 12

제 9항에 있어서,

상기 손실함수는,

센터 손실(center loss)로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험

점수 산출 장치.

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^n \|g(x_i W_I) - c_{y_i}\|_2^2$$

여기서,

$$g(x_i W_I) \in \mathbb{R}^d, \quad c_{y_i} \in \mathbb{R}^d, \quad g = \text{encoder},$$

n : # of samples, d : # of demensions

청구항 13

제 9항에 있어서,

상기 손실함수는,

퓨즈드 라소(fused lasso) 손실로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 장치.

$$\mathcal{L}_{fused\ lasso} = \sum_{j=2}^m \|w_j - w_{j-1}\|_1$$

여기서,

m : # of features, $w \in \mathbb{R}^m$

청구항 14

제 1 입력층에 전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부를 일대일층(one-to-one layer)으로 구성하는 제 1 입력층 전처리단계;

제 2 입력층에 SNP 전체를 일대일층으로 구성하는 제 2 입력층 전처리단계;

상기 제 1 입력층 및 상기 제 2 입력층을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 은닉층에 나타내는 인코딩단계; 및

상기 제 1 입력층 및 상기 제 2 입력층이 출력층에 복원되도록 상기 은닉층을 디코딩하는 디코딩단계;를 포함하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 15

제 14항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 변이정보, 나이, 성별이 주어졌을 때 질병에 걸릴 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 16

제 14항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 변이정보가 주어졌을 때 질병에 걸릴 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 17

제 14항에 있어서,

상기 전처리는, SNP 각각에 대한 변이정보가 상기 SNP 각각에 대해 나이, 성별이 주어졌을 때 변이정보에 대한 확률로 치환하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 18

제 14항에 있어서,

상기 일대일층은, 신경망 구조(Neural network)에서 SNP 각각에 일대일 대응으로 가중치를 연결하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 19

제 14항에 있어서,

상기 인코딩단계 후, 상기 잠재변수를 질병 유무에 관한 정보를 반영하기 위해 완전 연결층으로 연결하는 완전 연결층 변환단계를 더 포함하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 20

제 19항에 있어서,

상기 완전 연결층 변환단계 이후 분류단계를 더 포함하여, 상기 유전적 위험 점수 (Genomic Risk Score)를 질병 유무에 대한 정보가 손실함수를 통해 학습된 잠재변수를 토대로 산출하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 21

제 20항에 있어서,

상기 디코딩단계 및 상기 분류단계 이후, 유전적 위험 점수(Genomic Risk Score)가 상기 잠재변수를 모두 더한 후 상기 잠재변수의 차원으로 나눈 값으로 산출되도록 하는 잠재변수 추출 및 스코어링단계를 더 포함하는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 22

제 14항에 있어서,

상기 은닉층은, 손실함수를 최소화하도록 학습되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

청구항 23

제 22항에 있어서,

상기 손실함수는,

프랙털 오토인코더(Fractal Autoencoders) 손실로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

$$\mathcal{L}_{FAE} = \min_{W_I, f, g} \|X - f(g(XW_I))\|_F^2 + \lambda_1 \|X - f(g(XW_I^{max_k}))\|_F^2 + \lambda_2 \|W_I\|_1$$

여기서,

$X \in \mathbb{R}^{n \times m}$, n : # of samples, m : # of features

$W_I = \text{Diag}(w)$, $w \in \mathbb{R}^m$

g = encoder, f = decoder

λ_1, λ_2 : *wegihts*, max_k : pre-defined parameter for biomarker

청구항 24

제 22항에 있어서,

상기 손실함수는,

지도항목(Supervised term)손실로 다음과 같은 수학적식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전

적 위험 점수 산출 방법.

$$\mathcal{L}_{supervised} = \mathcal{L}_{BCElogit}(y, \psi(g(XW_I)))$$

여기서,

$$X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$$y \in \mathbb{R}^n$$

$g = \text{encoder}$, $\psi : \text{fully connected layer}$

청구항 25

제 22항에 있어서,

상기 손실함수는,

센터 손실(center loss)로 다음과 같은 수학식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^n \|g(x_i W_I) - c_{y_i}\|_2^2$$

여기서,

$$g(x_i W_I) \in \mathbb{R}^d, \quad c_{y_i} \in \mathbb{R}^d, \quad g = \text{encoder},$$

$$n : \# \text{ of samples}, \quad d : \# \text{ of demensions}$$

청구항 26

제 22항에 있어서,

상기 손실함수는,

퓨즈드 라소(fused lasso) 손실로 다음과 같은 수학식으로 정의되는 것을 특징으로 하는 딥러닝을 활용한 유전적 위험 점수 산출 방법.

$$\mathcal{L}_{fused\ lasso} = \sum_{j=2}^m \|w_j - w_{j-1}\|_1$$

여기서,

$$m : \# \text{ of features}, \quad w \in \mathbb{R}^m$$

발명의 설명

기술 분야

[0001] 본 발명은 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법에 관한 것으로, 상세하게는, 딥러닝을 활용해 SNP(Single Nucleotide Polymorphism; 단일염기다형) 변이정보의 비선형적인 관계를 고려하여 유전적 위험 점수의 정확성을 높일 수 있는 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법에 관한 것이다.

배경 기술

[0002] 현대의학은 단순한 질병 치료 목적에서 벗어나 건강하고 오래 살 수 있는 건강수명 연장에 초점을 두고있다.

[0003] 특히, Personalized Medicine(개인맞춤의학), Preventive Medicine(예방의학), Predictive Medicine(예측의학), Participatory Medicine(참여의학), 및 Precision Medicine(정밀의학)이라 불리는 5P 의학을 통해

진단, 치료 뿐만 아니라 질병을 예방하고 개인 건강의 최적화를 목표로 하고 있다.

- [0004] 이러한 초개인화 의료의 근간은 개인의 유전체를 분석하는 것이며 특히, 염기서열의 변이정보를 통해 유전 인자에 의해서 생물 내부적으로 결정되는 숨겨진 형질인 유전자형과, 그것이 겉으로 드러나는 것인 표현형 간의 연관 관계를 분석하는 것이다.
- [0005] 염기서열을 분석하는 방법은 RFLP (Restriction Fragment Length Polymorphism; 제한 단편 길이 다형), STR (Short Tandem Repeats), VNTR (Variable Number Tandem Repeat), SNP(Single Nucleotide Polymorphism; 단일 염기다형) 등이 알려져 있다.
- [0006] 이중 SNP는 A, T, G, C 4가지의 조합으로 구성된 유전자 염기서열 내 하나의 DNA 염기서열의 변이를 의미하는데, SNP 변이 정보의 조합을 유전형이라고 하며 이를 통해 증상 또는 질병(표현형)이 발현된다.
- [0007] SNP는 인간의 경우 약 1,000bp 마다 1회의 빈도로 발생하는 것으로 알려져 있는데, 개인마다 변이의 정보가 다르기 때문에 이를 분석하여 질병을 초기에 예방하거나 환자에게 반응이 뛰어나고 위험성을 적은 약물과 치료를 추천할 수 있으므로 이에 대한 연구가 지속되어 왔다.
- [0008] 그 일례로, 대한민국 공개특허공보 제10-2020-0132461호에서는 제2형 당뇨병과 관련하여 기 보고된 유전변이 175개 유전변이와 공복혈당과 연관성을 나타내는 56개 유전변이에 대한 유전적 위험 점수를 점수화하여 복합적으로 분석함으로써, 나이와 성별, 환경적인 영향 등 추가적인 검사 없이 각 개인의 유전정보만을 이용하여 제2형 당뇨의 유전적 고위험군을 선별할 수 있으며, 제2형 당뇨의 유전적 고위험군 선별에 따라 맞춤형 진료, 치료 및 예방에 활용할 수 있는 당뇨병 진단 또는 발병 위험 예측을 위한 키트 및 이를 이용한 진단 방법을 제안하였다.
- [0009] 그러나, 이 경우에도 22개 염색체 안의 SNP들이 매우 복잡하게 서로 연관된 비선형적인 관계를 유전적 위험 점수에 반영할 수 없어 실효성이 떨어진다는 단점이 있다.

선행기술문헌

특허문헌

- [0010] (특허문헌 0001) 대한민국 공개특허공보 제10-2020-0132461호(2020.10.14)

발명의 내용

해결하려는 과제

- [0011] 본 발명의 목적은, SNP(Single Nucleotide Polymorphism; 단일염기다형) 변이정보의 비선형적인 관계를 고려하고 질병의 유무에 대한 정보를 반영하여 유전적 위험 점수의 실효성을 높이는 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법을 제공하는 것이다.
- [0012] 본 발명은 해석이 가능하며 유전적 위험 점수에 영향을 미치는 바이오마커를 추출해낼 수 있는 방법까지 고안한 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법을 제공하는데 또 다른 목적이 있다.

과제의 해결 수단

- [0013] 본 발명에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치는 전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부에 대해 일대일층(one-to-one layer)으로 구성된 제 1 입력층, SNP 전체에 대해 일대일층으로 구성된 제 2 입력층, 제 1 입력층 및 제 2 입력층을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 나타낸 은닉층, 및 제 1 입력층과 제 2 입력층이 복원되도록 은닉층을 디코딩한 출력층을 포함할 수 있다.
- [0014] 여기서, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 변이정보, 나이, 성별이 주어졌을 때 질병에 걸릴 확률로 치환할 수 있다.
- [0015] 또한, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 변이정보가 주어졌을 때 질병에 걸릴 확률로 치환할 수 있다.

[0016] 여기서, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 나이, 성별이 주어졌을 때 변이정보에 대한 확률로 치환할 수 있다.

[0017] 또한, 일대일층은 신경망 구조(Neural network)에서 SNP 각각에 일대일 대응으로 가중치를 연결할 수 있다.

[0018] 또한, 잠재변수는 완전 연결층으로 연결된 분류기를 통해 특정 질병에 걸릴 확률로 산출되어 실제 질병 유무와 비교하는 손실함수를 최소화하도록 학습할 수 있다.

[0019] 여기서, 잠재변수는 질병 유무에 대한 정보가 반영되도록 학습되며 특정 질병에 걸릴 확률이 높을수록 유전적 위험 점수가 높게 학습되도록 할 수 있다.

[0020] 여기서, 유전적 위험 점수(Genomic Risk Score)는 잠재변수의 값을 모두 더한 후 잠재변수의 차원으로 나눈 값으로 산출될 수 있다.

[0021] 여기서, 은닉층은 손실함수를 최소화하도록 학습될 수 있다.

[0022] 또한, 손실함수는 프랙털 오토인코더(Fractal Autoencoders) 손실로 다음과 같은 수학식으로 정의될 수 있다.

$$[0023] \mathcal{L}_{FAE} = \min_{W_I, f, g} \|X - f(g(XW_I))\|_F^2 + \lambda_1 \|X - f(g(XW_I^{max_k}))\|_F^2 + \lambda_2 \|W_I\|_1$$

[0024] 여기서,

$$[0025] X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$[0026] W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$$[0027] g = \text{encoder}, \quad f = \text{decoder}$$

$$[0028] \lambda_1, \lambda_2 : \text{weights}, \quad max_k : \text{pre-defined parameter for biomarker}$$

[0030] 여기서, 손실함수는 지도항목(Supervised term)손실로 다음과 같은 수학식으로 정의될 수 있다.

$$[0031] \mathcal{L}_{supervised} = \mathcal{L}_{BCElogit}(y, \psi(g(XW_I)))$$

[0032] 여기서,

$$[0033] X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$[0034] W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$$[0035] y \in \mathbb{R}^n$$

$$[0036] g = \text{encoder}, \quad \psi : \text{fully connected layer}$$

[0038] 또한, 손실함수는 센터 손실(center loss)로 다음과 같은 수학식으로 정의될 수 있다.

$$[0039] \mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^n \|g(x_i W_I) - c_{y_i}\|_2^2$$

[0040] 여기서,

$$[0041] g(x_i W_I) \in \mathbb{R}^d, \quad c_{y_i} \in \mathbb{R}^d, \quad g = \text{encoder},$$

$$[0042] n : \# \text{ of samples}, \quad d : \# \text{ of demensions}$$

한편, 손실함수는 퓨즈드 라소(fused lasso) 손실로 다음과 같은 수학적식으로 정의될 수 있다.

$$\mathcal{L}_{fused\ lasso} = \sum_{j=2}^m \|w_j - w_{j-1}\|_1$$

여기서,

$$m : \# \text{ of features}, \quad w \in \mathbb{R}^m$$

본 발명의 다른 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 방법은 제 1 입력층에 전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부를 일대일층(One-to-one layer)으로 구성하는 제 1 입력층 전처리단계, 제 2 입력층에 SNP 전체를 일대일층으로 구성하는 제 2 입력층 전처리단계, 제 1 입력층 및 제 2 입력층을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 은닉층에 나타내는 인코딩단계, 및 상기 제 1 입력층 및 제 2 입력층이 출력층에 복원되도록 은닉층을 디코딩하는 디코딩단계를 포함할 수 있다.

여기서, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 변이정보, 나이, 성별이 주어졌을 때 질병에 걸릴 확률로 치환할 수 있다.

또한, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 변이정보가 주어졌을 때 질병에 걸릴 확률로 치환할 수 있다.

여기서, 전처리는 SNP 각각에 대한 변이정보가 SNP 각각에 대해 나이, 성별이 주어졌을 때 변이정보에 대한 확률로 치환할 수 있다.

또한, 일대일층은 신경망 구조(Neural network)에서 SNP 각각에 일대일 대응으로 가중치를 연결할 수 있다.

한편, 인코딩단계 이후, 잠재변수를 질병 유무에 관한 정보를 반영하기 위해 완전 연결층으로 연결하는 완전 연결층 변환단계를 더 포함할 수 있다.

또한, 완전 연결층 변환단계 이후 분류단계를 더 포함하여, 유전적 위험 점수(Genomic Risk Score)를 분류기에 의해 특정 질병이 걸릴 확률로 산출하여 실제 질병 유무와 비교하는 손실함수를 최소화하도록 학습할 수 있다.

한편, 디코딩단계 및 분류단계 이후, 유전적 위험 점수(Genomic Risk Score)가 잠재변수의 값을 모두 더한 후 잠재변수의 차원으로 나눈 값으로 산출되도록 하는 잠재변수 추출 및 스코어링단계를 더 포함할 수 있다.

여기서, 은닉층은 손실함수를 최소화하도록 학습될 수 있다.

또한, 손실함수는 프랙털 오토인코더(Fractal Autoencoders) 손실로 다음과 같은 수학적식으로 정의될 수 있다.

$$\mathcal{L}_{FAE} = \min_{W_I, f, g} \|X - f(g(XW_I))\|_F^2 + \lambda_1 \|X - f(g(XW_I^{max_k}))\|_F^2 + \lambda_2 \|W_I\|_1$$

여기서,

$$X \in \mathbb{R}^{n \times m}, \quad n : \# \text{ of samples}, \quad m : \# \text{ of features}$$

$$W_I = \text{Diag}(w), \quad w \in \mathbb{R}^m$$

$$g = \text{encoder}, \quad f = \text{decoder}$$

$$\lambda_1, \lambda_2 : \text{weights}, \max_k : \text{pre-defined parameter for biomarker}$$

여기서, 손실함수는 지도항목(Supervised term)손실로 다음과 같은 수학적식으로 정의될 수 있다.

$$\mathcal{L}_{supervised} = \mathcal{L}_{BCElogit}(y, \psi(g(XW_I)))$$

여기서,

[0069] $X \in \mathbb{R}^{n \times m}$, n : # of samples, m : # of features

[0070] $W_I = \text{Diag}(w)$, $w \in \mathbb{R}^m$

[0071] $y \in \mathbb{R}^n$

[0072] $g = \text{encoder}$, ψ : *fully connected layer*

[0074] 또한, 손실함수는 센터 손실(center loss)로 다음과 같은 수학적식으로 정의될 수 있다.

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^n \|g(x_i W_I) - c_{y_i}\|_2^2$$

[0076] 여기서,

[0077] $g(x_i W_I) \in \mathbb{R}^d$, $c_{y_i} \in \mathbb{R}^d$, $g = \text{encoder}$,

[0078] n : # of samples, d : # of demensions

[0080] 한편, 손실함수는 퓨즈드 라소(fused lasso) 손실로 다음과 같은 수학적식으로 정의될 수 있다.

$$\mathcal{L}_{fused\ lasso} = \sum_{j=2}^m \|w_j - w_{j-1}\|_1$$

[0082] 여기서,

[0083] m : # of features, $w \in \mathbb{R}^m$

발명의 효과

[0085] 본 발명에 의한 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법은 SNP(Single Nucleotide Polymorphism: 단일염기다형) 변이정보의 비선형적인 관계를 고려하고 질병의 유무에 대한 정보를 반영하여 유전적 위험 점수의 실효성을 높이는 장점이 있다.

[0086] 또한, 본 발명에 의한 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법은 유전적 위험 점수를 산출함에 있어 일대일층에서 높은 가중치를 갖는 SNP를 바이오 마커에 활용할 수 있는 장점이 있다.

도면의 간단한 설명

[0087] 도 1은 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치를 나타낸 구성도이다.

도 2는 도 1의 제 1 입력층 및 제 2 입력층의 전처리를 나타낸 도면으로, 도 2(a)는 훈련셋 데이터 및 도 2(b)는 테스트셋 데이터를 나타낸 도표이다.

도 3은 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 방법을 나타낸 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0088] 이하, 본 발명의 실시를 위한 구체적인 실시예를 첨부된 도면들을 참조하여 설명한다.

[0089] 본 발명을 설명함에 있어서 제 1, 제 2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 구성요소들은 용어들에 의해 한정되지 않을 수 있다. 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 된다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제 1 구성요소는 제 2 구성요소로 명명될 수 있고, 유사하게 제 2 구성요소도 제 1 구성요소로 명명될 수 있다.

[0090] 어떤 구성요소가 다른 구성요소에 연결되어 있거나 접속되어 있다고 언급되는 경우는, 그 다른 구성요소에 직

접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해될 수 있다.

- [0091] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함할 수 있다.
- [0092] 본 명세서에서, 포함하다 또는 구비하다 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것으로서, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해될 수 있다.
- [0093] 또한, 도면에서의 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.
- [0094] 이하, 첨부된 도면을 참조하여 본 발명에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법에 대해 상세히 설명한다.
- [0095] 도 1은 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치를 나타낸 구성도이며, 도 2는 도 1의 제 1 입력층 및 제 2 입력층의 전처리를 상세히 설명하기 위한 세부 도표이다.
- [0096] 이하, 도 1 내지 도 2를 참조하여 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치를 설명한다.
- [0097] 먼저, 도 1을 참조하면, 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치는 전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부에 대해 일대일층(one-to-one layer)으로 구성된 제 1 입력층(100), SNP 전체에 대해 일대일층으로 구성된 제 2 입력층(200), 제 1 입력층(100) 및 제 1 입력층(100)과 제 2 입력층(200)을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 나타낸 은닉층(300), 및 제 1 입력층(100)과 제 2 입력층(200)이 복원되도록 은닉층(300)을 디코딩한 출력층(400)으로 이루어져 학습을 수행한다.
- [0098] 여기서, SNP란 A, T, G, C 4가지의 조합으로 구성된 유전자 염기서열 내 하나의 DNA 염기서열의 변이를 의미하는 것으로 유전자 염기서열은 나선형 구조라는 특징에 의해 30억개의 유전자 염기서열 중 1,000개 마다 1개 이상의 변이가 나타나며, 종의 특징을 구분하는 주요 요소로 작용하고 있다.
- [0099] 본 발명에 의한 딥러닝을 활용한 유전적 위험 점수 산출 장치는 특정 질병에 대해 SNP를 분석한 결과를 학습함으로써 유전적 위험 점수를 산출하고 있으며, 유전적 위험 점수를 산출함에 있어 일대일층에서 높은 가중치를 갖는 SNP에 대해 바이오마커로 활용할 수 있는 장점이 있다.
- [0100] 한편, 제 1 입력층(100) 및 제 2 입력층(200)에 포함된 일대일층은 SNP 각각의 가중치를 토대로 다시 SNP 각각으로 일대일 대응되어 출력된다. 이때, 가중치는 임의의 값으로 초기화할 수 있으며 학습을 통해 특정 SNP가 향후 바이오마커로 활용이 가능한지의 여부를 조사하기 위해 사용될 수 있다.
- [0101] 은닉층(300)에서의 잠재변수는 유전적 위험 점수 산출을 위해 완전 연결층(500)으로 연결되어 1차원으로 변환(fully connected)되고 분류기(600)에 의해 특정 질병이 걸릴 확률로 산출된다.
- [0102] 여기서, 유전적 위험 점수(Genomic Risk Score)는 은닉층(300)의 값을 모두 더한 후 은닉층(300)의 차원으로 나눈 값으로 수학적 식 1과 같이 정의할 수 있다.

수학적 식 1

$$[0103] \text{Genomic Risk Score} = \frac{g(x_i W_1)}{\dim(g(x_i W_1))} = \frac{g(x_i W_1)}{d}$$

[0104] 여기서,

[0105] $g(\cdot)$: 인코더

[0106] $g(x_i W_1) \in \mathbb{R}^d$: 인코더 이후 잠재변수(Latent feature)

[0108] 즉, 수학식1은 제 1 입력층(100) 및 제 2 입력층(200)에 대해 인코더에 학습시킨 이후 산출된 잠재변수의 숫자를 모두 더하고 이것을 잠재변수의 차원의 수로 나누어 스코어를 계산한 것으로 볼 수 있다.

[0109] 이와 같이 산출된 스코어에 대해 은닉층(300)은, SNP를 제공한 제공자가 분류기에 의해 특정 질병이 걸릴 확률로 산출하여 실제 질병 유무와 비교하는 손실함수를 최소화하도록 학습된다.

[0110] 한편, 은닉층(300)은 손실함수를 최소화하도록 학습될 수 있는데, 손실함수는 총 4개로 구성되며, 적어도 어느 하나 이상을 고려하거나 총 4개의 손실함수를 모두 더하여 사용할 수도 있다.

[0111] 이러한 손실함수가 필요한 이유는 다음과 같은데, 즉 전체 SNP 보다는 일부 SNP에 대해 수행하는 프랙털 오토인코더(Fractal Autoencoders) 손실을 통해 유전자 정보를 요약하여 중요 변수의 선택을 돕고, 지도항목(Supervised term) 손실을 통해서 압축된 유전자 정보에 질병에 대한 정보를 반영하며, 센터 손실(center loss)을 통해 질병이 있는 잠재 변수와 없는 잠재 변수를 보다 명확히 구분하고, 퓨즈드 라소(fused lasso) 손실을 통해서 인접 유전정보를 같이 행동할 수 있게 학습할 필요가 있기 때문이다.

[0112] 여기서, 손실함수 중 프랙털 오토인코더 손실은 수학식 2와 같이 정의된다.

수학식 2

$$[0113] \mathcal{L}_{FAE} = \min_{W_1, f, g} \|X - f(g(XW_1))\|_F^2 + \lambda_1 \|X - f(g(XW_1^{max_k}))\|_F^2 + \lambda_2 \|W_1\|_1$$

[0114] 여기서,

[0115] $X \in \mathbb{R}^{n \times m}$, n : # of samples, m : # of features

[0116] $W_1 = \text{Diag}(w)$, $w \in \mathbb{R}^m$

[0117] g = encoder, f = decoder

[0118] λ_1, λ_2 : weights, max_k : pre-defined parameter for biomarker

[0120] 상기와 같이, 본 발명의 프랙털 오토인코더 손실 모델은 전체 SNP에 대해 복원 에러가 적도록 수행하는 기존의 오토인코더 손실 모델과, 선택된 SNP에 대해 복원 에러가 적도록 수행하는 손실 항목과, 오버피팅(overfitting)을 해결하기 위한 라소(Lasso) 손실 항목이 영향도에 따라 구분된 합으로 구성되는데, 이는 "입력 데이터의 정보를 보존하는 표현형 추출"을 목적으로 하는 것과 동시에 "데이터 변환"과 "변수 선택"을 목적으로 하며, 본 발명은 이를 통해 도출된 중요 변수를 바이오 마커 발굴에 활용할 수 있는 장점이 있다.

[0121] 또한, 손실함수 중 지도항목손실은 수학식 3과 같이 정의된다.

수학식 3

$$[0122] \mathcal{L}_{supervised} = \mathcal{L}_{BCElogit}(y, \psi(g(XW_1)))$$

[0123] 여기서,

[0124] $X \in \mathbb{R}^{n \times m}$, n : # of samples, m : # of features

[0125] $W_1 = \text{Diag}(w)$, $w \in \mathbb{R}^m$

[0126] $y \in \mathbb{R}^n$

[0127] $g = \text{encoder}, \psi : \text{fully connected layer}$

[0129] 지도항목손실은, 대표적인 분류 문제의 목적 함수인 이진교차엔트로피 손실(BCEwithLogit loss)을 활용하여 프랙털 오토인코더의 잠재 변수가 질병에 대한 정보를 담는 것을 목표로 하도록 한다.

[0130] 또한, 손실함수 중 센터 손실은 수학적 식 4와 같이 정의된다.

수학적 식 4

[0131]
$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^n \|g(x_i W_I) - c_{y_i}\|_2^2$$

[0132] 여기서,

[0133] $g(x_i W_I) \in \mathbb{R}^d, \quad c_{y_i} \in \mathbb{R}^d, \quad g = \text{encoder},$

[0134] $n : \# \text{ of samples}, d : \# \text{ of demensions}$

[0136] 일반적인 분류손실(classification loss)은 클래스간(inter-class) 만을 잘 구별하는 형태로 학습시켜 다른 클래스 간의 공간 분할이 뚜렷하지 않을 수 있으나, 본 발명에 의한 센터 손실은 같은 클래스에 속한 샘플 간 거리를 가깝게 하는 것을 목표로 하며, 이를 통해 질병 예측 성능을 강화할 수 있는 장점이 있다.

[0137] 마지막으로, 손실함수 중 정칙화 기법인 퓨즈드 라소 손실은 수학적 식 5와 같

수학적 식 5

[0138]
$$\mathcal{L}_{fused \text{ lasso}} = \sum_{j=2}^m \|w_j - w_{j-1}\|_1$$

[0139] 여기서,

[0140] $m : \# \text{ of features}, \quad w \in \mathbb{R}^m$

[0142] 이 정의된다.

[0143] 퓨즈드 라소 손실은 서로 거리가 가까운 유전자 위치끼리는 유전형이 섞이지 않고 함께 이동하여 연관불평형 블록(Linkage disequilibrium block)을 고려할 수 있도록 한다. 또한, 인접한 변수들을 동시에 선택이 가능하여 스무딩(smoothing) 효과도 있다.

[0144] 이를 상세히 설명하면, 한 염색체에 위치하는 2개의 SNPs간 거리가 아주 멀다면 감수분열 과정에서 교차(Crossover)가 빈번히 발생하는데, 이때 재조합은 단일염기수준이 아니라 큰 블록의 단위로 일어나게 되며 이렇게 함께 유전되는 단위를 일배체형(Haplotype)이라 한다.

[0145] 이러한 재조합이 일어남으로써 서로 다른 조합의 일배체형이 독립적으로 발생하게 되는데, 이처럼 독립적으로 가능한 조합이 모두 발생하는 경우 두 SNPs는 연관평형(linkage equilibrium) 상태에 있다고 한다. 그러나 SNPs 간의 거리가 매우 가깝다면 2개의 SNPs는 서로 연관되어 다음 세대에 같이 전달되게 되고, 이때, 일배체형의 조합이 모두 존재하지 않는다면 연관불평형(linkage disequilibrium) 상태에 있다고 한다.

[0146] 이상과 같은 본 발명에 의한 딥러닝을 활용한 유전적 위험 점수 산출 장치는 점수 산출에 영향을 미치는 바이오 마커를 추출해낼 수 있으며, 잠재 변수에 질병에 대한 정보를 내포시키도록 학습시켜 바이오 빅데이터 기반 유

전자 질병 나이(위험도)를 추정하는 유전적 위험 점수를 산출할 수 있는 장점이 있다.

- [0148] 도 2는 도 1의 제 1 입력층(100) 및 제 2 입력층(200)의 전처리를 나타낸 도면으로, 도 2(a)는 훈련셋(Training set) 데이터를 나타낸 도표이고 도 2(b)는 테스트셋(Test set) 데이터를 나타낸 도표이다.
- [0149] 도 2에서 볼 수 있는 바와 같이, 원본 데이터는 하나의 SNP 당 변이 개수인 SNP의 변이정보로 0, 1, 2 만으로 구성된 매우 고차원의 데이터이다. 이러한 명목형 데이터는 수치적 정보를 담고 있지 않기 때문에 딥러닝을 활용하는데 많은 어려움을 준다.
- [0150] 또한, 이러한 명목형 데이터에 대해 one-hot encoding (0과 1로 구성하는 인코딩 방법)을 사용할 경우 더 큰 차원으로 확장되는 단점이 있다.
- [0151] 본 발명에서는 이것을 극복하기 위해 특정 조건에 따른 확률값(조건부 확률)을 계산하여 각각의 변수(SNP)에 맵핑을 함으로써, 명목형 변수를 수치형 변수로 바꿀 수 있고 이를 통해 딥러닝을 원활히 수행할 수 있는 장점이 있다.
- [0152] 예를 들어, SNP를 제공하는 제공자에 대한 나이, 성별, 변이정보를 토대로 확률값을 산출할 수 있는데, 이러한 훈련셋(Training set)을 기준으로 확률값을 산출하고 테스트셋(Test set)에는 계산없이 훈련셋(Training set)의 조건에 맞게 매핑할 수 있다.
- [0153] 즉, 제 1 입력층(100) 및 제 2 입력층(200)을 생성하기 위한 전처리는, SNP 각각에 대해 변이정보, 나이, 성별이 주어졌을 때 질병에 걸릴 확률로 치환하거나, 변이정보가 주어졌을 때 질병에 걸릴 확률로 치환하거나, 나이, 성별이 주어졌을 때 변이정보에 대한 확률로 치환할 수도 있다.
- [0154] 도 2(a)에서는 훈련셋 데이터에 대해 성별은 남자, 나이는 40대, 및 변이정보가 0인 SNP에 대해 질병에 걸릴 확률을 변이정보 대신 치환한 경우이다. 즉, N번째 염색체의 변이정보가 0인 40대 남자가 질병에 걸릴 확률은 총 3명 중 2명으로 0.666이므로, N번째 염색체 변이정보 0 대신 0.666을 치환한 경우를 나타낸다.
- [0155] 또한, 도 2(b)에서는 테스트셋 데이터에 대해 훈련셋에서 치환한 확률값을 그대로 적용한 경우를 나타낸다. 이를 통해 훈련셋 데이터를 토대로 기존의 명목형으로 구성된 데이터의 한계를 극복할 수 있으며, 본 발명의 딥러닝을 활용한 유전적 위험 점수 산출 장치의 학습 수행을 통해 잠재변수도 산출하고 잠재변수에 기여한 중요 SNP를 선택하여 이를 바이오 마커에 활용할 수 있다.
- [0157] 도 3은 본 발명의 일 실시예에 따른 딥러닝을 활용한 유전적 위험 점수 산출 방법을 나타낸 순서도이다.
- [0158] 도 3에서 알 수 있는 바와 같이, 딥러닝을 활용한 유전적 위험 점수 산출 방법은 제 1 입력층(100)에 전처리가 수행된 SNP(Single Nucleotide Polymorphism; 단일염기다형) 일부를 일대일층(one-to-one layer)으로 구성하는 제 1 입력층 전처리단계(S100), 제 2 입력층(200)에 SNP 전체를 일대일층으로 구성하는 제 2 입력층 전처리단계(S200), 제 1 입력층(100) 및 제 2 입력층(200)을 인코딩하여 특정 질병의 잠재변수(Latent feature)를 은닉층(300)에 나타내는 인코딩단계(S300), 및 상기 제 1 입력층(100) 및 제 2 입력층(200)이 출력층(400)에 복원되도록 은닉층(300)을 디코딩하는 디코딩단계(S400)로 이루어져 학습을 수행한다.
- [0159] 여기서, SNP란 A, T, G, C 4가지의 조합으로 구성된 유전자 염기서열 내 하나의 DNA 염기서열의 변이를 의미하는 것으로 유전자 염기서열은 나선형 구조라는 특징에 의해 30억개의 유전자 염기서열 중 1,000개 마다 1개 이상의 변이가 나타나며, 종의 특징을 구분하는 주요 요소로 작용하고 있다.
- [0160] 한편, 제 1 입력층(100) 및 제 2 입력층(200)에 포함된 일대일층은 SNP 각각의 가중치를 토대로 다시 SNP 각각으로 일대일 대응되어 출력된다. 이때, 가중치는 임의의 값으로 초기화할 수 있으며 학습을 통해 특정 SNP가 향후 바이오마커로 활용이 가능한지의 여부를 조사하기 위해 사용될 수 있다.
- [0162] 본 발명에 의한 딥러닝을 활용한 유전적 위험 점수 산출 방법은 인코딩단계(S300) 이후 잠재변수에 대해 질병 유무에 관한 정보를 반영하기 위해 완전 연결층(500)으로 연결하는 완전 연결층 변환단계(S500)를 더 포함할 수 있다.
- [0163] 또한, 완전 연결층 변환단계(S500) 이후 분류단계(S600)를 더 포함하여 유전적 위험 점수(Genomic Risk Score)

를 분류기에 의해 특정 질병이 걸릴 확률로 산출하고 실제 질병 유무와 비교하는 손실함수를 최소화하도록 학습할 수 있다.

[0164] 한편, 본 발명의 딥러닝을 활용한 유전적 위험 점수 산출 방법은 디코딩단계(S400) 및 분류단계(S600) 이후 유전적 위험 점수(Genomic Risk Score)가 잠재변수의 값을 모두 더한 후 잠재변수의 차원으로 나눈 값으로 산출되도록 하는 잠재변수 추출 및 스코어링단계(S700)를 더 포함할 수 있다.

[0165] 여기서, 유전적 위험 점수는 상기의 수학식 1과 같이 정의할 수 있으며, 도 1에서 이미 설명한 바가 있으므로 이에 대한 설명은 생략한다.

[0166] 한편, 은닉층(300)은 손실함수를 최소화하도록 학습될 수 있는데, 손실함수는 총 4개로 구성되며, 적어도 어느 하나 이상을 고려하여 사용할 수 있다. 예를 들어 총 4개의 손실함수를 모두 더하여 사용할 수도 있다.

[0167] 이러한 손실함수가 필요한 이유는 다음과 같은데, 즉 전체 SNP 보다는 일부 SNP에 대해 수행하는 프랙털 오토인코더(Fractal Autoencoders) 손실을 통해 유전자 정보를 요약하여 중요 변수 선택하고, 지도항목(Supervised term) 손실을 통해서 압축된 유전자 정보에 질병에 대한 정보를 반영하며, 센터 손실(center loss)을 통해 질병이 있는 잠재 변수와 없는 잠재 변수를 보다 명확히 구분하고, 퓨즈드 라소(fused lasso) 손실을 통해서 인접 유전정보를 같이 행동할 수 있게 학습할 필요가 있기 때문이다.

[0168] 각 손실함수와, 제 1 입력층(100) 및 제 2 입력층(200)을 생성하기 위한 전처리에 대해서는 도 1 및 도 2에서 이미 설명한 바가 있으므로 이에 대한 설명은 생략한다.

[0170] 이상과 같은 본 발명에 따른 딥러닝을 활용한 유전적 위험 점수 산출 장치 및 방법은 SNP(Single Nucleotide Polymorphism; 단일염기다형) 변이정보의 비선형적인 관계를 고려하여 유전적 위험 점수의 실효성을 높이는 장점이 있으며, 유전적 위험 점수를 산출함에 있어 일대일층에서 높은 가중치를 갖는 SNP을 바이오 마커에 활용할 수 있는 장점이 있다.

[0172] 본 발명의 기술 분야에서 통상의 지식을 가진 자는 여기에 개시된 실시예들과 관련하여 설명된 다양한 예시적인 논리 블록들, 모듈들, 프로세서들, 수단들, 회로들 및 알고리즘 단계들이 전자 하드웨어, (편의를 위해, 여기에서 소프트웨어로 지칭되는) 다양한 형태들의 프로그램 또는 설계 코드 또는 이들 모두의 결합에 의해 구현될 수 있다는 것을 이해할 것이다. 하드웨어 및 소프트웨어의 이러한 상호 호환성을 명확하게 설명하기 위해, 다양한 예시적인 컴포넌트들, 블록들, 모듈들, 회로들 및 단계들이 이들의 기능과 관련하여 위에서 일반적으로 설명되었다. 이러한 기능이 하드웨어 또는 소프트웨어로서 구현되는지 여부는 특정한 애플리케이션 및 전체 시스템에 대하여 부과되는 설계 제약들에 따라 좌우된다. 본 발명의 기술 분야에서 통상의 지식을 가진 자는 각각의 특정한 애플리케이션에 대하여 다양한 방식으로 설명된 기능을 구현할 수 있으나, 이러한 구현 결정들은 본 발명의 범위를 벗어나는 것으로 해석되어서는 안 될 것이다.

[0173] 여기서 제시된 다양한 실시예들은 방법, 장치, 또는 표준 프로그래밍 및/또는 엔지니어링 기술을 사용한 제조물품(article)으로 구현될 수 있다. 용어 제조물품은 임의의 컴퓨터-판독가능 저장장치로부터 액세스 가능한 컴퓨터 프로그램, 캐리어, 또는 매체(media)를 포함한다. 예를 들어, 컴퓨터-판독가능 저장매체는 자기 저장 장치(예를 들면, 하드 디스크, 플로피 디스크, 자기 스트립, 등), 광학 디스크(예를 들면, CD, DVD, 등), 스마트 카드, 및 플래쉬 메모리 장치(예를 들면, EEPROM, 카드, 스틱, 키 드라이브, 등)를 포함하지만, 이들로 제한되는 것은 아니다. 또한, 여기서 제시되는 다양한 저장 매체는 정보를 저장하기 위한 하나 이상의 장치 및/또는 다른 기계-판독가능한 매체를 포함한다.

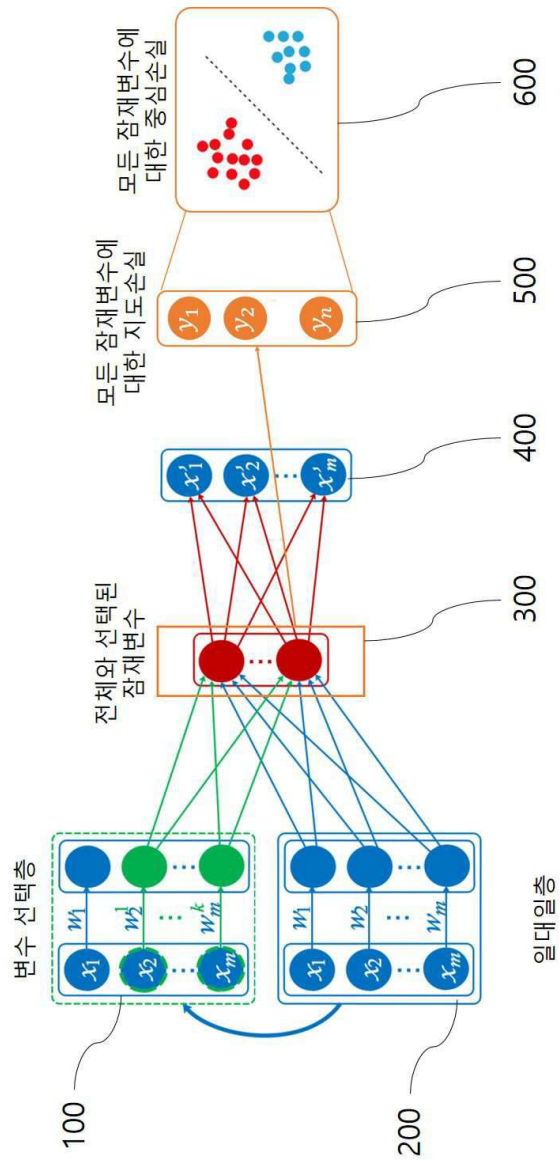
[0174] 제시된 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조는 예시적인 접근들의 일례임을 이해하도록 한다. 설계 우선순위들에 기반하여, 본 발명의 범위 내에서 프로세스들에 있는 단계들의 특정한 순서 또는 계층 구조가 재배열될 수 있다는 것을 이해하도록 한다. 첨부된 방법 청구항들은 샘플 순서로 다양한 단계들의 엘리먼트들을 제공하지만 제시된 특정한 순서 또는 계층 구조에 한정되는 것을 의미하지는 않는다.

[0175] 제시된 실시예들에 대한 설명은 임의의 본 발명의 기술 분야에서 통상의 지식을 가진 자가 본 발명을 이용하거나 또는 실시할 수 있도록 제공된다. 이러한 실시예들에 대한 다양한 변형들은 본 발명의 기술 분야에서 통상의 지식을 가진 자에게 명백할 것이며, 여기에 정의된 일반적인 원리들은 본 발명의 범위를 벗어남이 없이 다른 실시예들에 적용될 수 있다. 그리하여, 본 발명은 여기에 제시된 실시예들로 한정되는 것이 아니라, 여기에 제시

된 원리들 및 신규한 특징들과 일관되는 최광의의 범위에서 해석되어야 할 것이다.

도면

도면1



도면2

Sample ID	질병유무	성별	나이	N번째 염색체
NIH_01	True	M	41	0
NIH_02	True	F	53	1
NIH_07	False	F	61	0
NIH_08	True	M	43	0
NIH_09	False	M	44	0
NIH_10	False	M	69	1

Sample ID	질병유무	성별	나이	N번째 염색체
NIH_01	True	M	41	0.6666
NIH_02	True	F	53	1
NIH_07	False	F	61	0
NIH_08	True	M	43	0.6666
NIH_09	False	M	44	0.6666
NIH_10	False	M	69	1

(a)

Sample ID	질병유무	성별	나이	N번째 염색체 변이
NIH_11	True	M	47	0
NIH_12	False	M	42	0
NIH_13	True	F	52	1
NIH_14	True	M	64	2
NIH_15	False	F	55	0

Sample ID	질병유무	성별	나이	N번째 염색체 변이
NIH_11	True	M	47	0.6666
NIH_12	False	M	42	0.6666
NIH_13	True	F	52	1
NIH_14	True	M	64	2
NIH_15	False	F	55	0

(b)

도면3

