(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2015-0027938
(43) 공개일자 2015년03월13일

(51) 국제특허분류(Int. Cl.)
G06F 19/00 (2011.01) *G06F 17/10* (2006.01)
(21) 출원번호 10-2013-0106308
(22) 출원일자 2013년09월04일
심사청구일자 없음

(71) 출원인
삼성전자주식회사
경기도 수원시 영통구 삼성로 129 (매탄동)
(72) 발명자
손대순
서울 관악구 관악로 285, 103동 904호 (봉천동, 성현동아아파트)
안태진
서울 강남구 삼성로64길 5, 101동 602호 (대치동, 대치현대아파트)
(뒷면에 계속)
(74) 대리인
리앤목특허법인

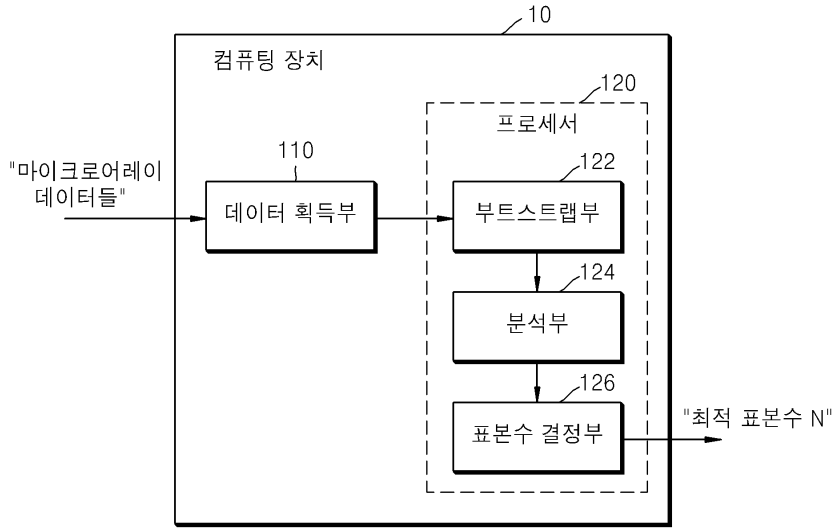
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **마이크로어레이 데이터를 분석하는 방법 및 장치**

(57) 요 약

피검체들에 대한 마이크로어레이 데이터들을 분석하는 방법 및 장치는, 상기 마이크로어레이 데이터들을 이용하여 부트스트랩(bootstrap) 데이터 세트들을 생성하고, 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 경험적 검정력(empirical power)을 산출함으로써 목표 검정력을 만족하는 최소의 표본수를 탐색한다.

대 표 도 - 도2a



(72) 발명자

이은진

서울 중랑구 동일로146길 33, 3층 (묵동, 신진빌라)

정종석

경기 화성시 병점1로 65, 106동 706호 (병점동, 늘벗마을신창1차아파트)

특허청구의 범위

청구항 1

피검체들에 대한 마이크로어레이 데이터들을 분석하는 방법에 있어서,

(a) 상기 마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응변수들을 갖는 표본수 N (N은 자연수)의 복수의 부트스트랩 데이터 세트들을 생성하는 단계;

(b) 상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 상기 표본수 N에 대한 경험적 검정력을 산출하는 단계; 및

(c) 상기 (a) 내지 (b) 단계들을 반복함으로써, 상기 산출된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수를 탐색하는 단계를 포함하고,

상기 (a) 내지 (c) 단계들은 컴퓨팅 장치에 의해 수행되는, 방법.

청구항 2

제 1 항에 있어서,

상기 이항반응변수들은

상기 부트스트랩 데이터 세트들에 포함된 상기 표본수 N에 대한 유전자 발현량 데이터를 이용하여 결정되는, 방법.

청구항 3

제 2 항에 있어서,

상기 (b) 단계는

상기 결정된 이항반응변수들에 대한, 상기 예측 모델들의 예측의 정확도를 나타내는 통계적 확률에 기초하여 상기 경험적 검정력을 산출하는, 방법.

청구항 4

제 1 항에 있어서,

상기 (a) 단계는

(a1) 상기 마이크로어레이 데이터들로부터 추출된 복수의 예비 데이터들(pilot data)에 대한 통계적 특성들 및 상기 예비 데이터들로부터 랜덤하게 추출된 유전자 발현 데이터들을 이용하여, 상기 표본수 N을 갖는 부트스트랩 유전자 발현 데이터들을 생성하는 단계;

(a2) 상기 부트스트랩 유전자 발현 데이터들을 이용하여, 상기 표본수 N에 대한 상기 이항반응변수들을 결정하는 단계; 및

(a3) 상기 (a1) 내지 (a2) 단계를 B회 (B는 자연수) 반복함으로써, 상기 결정된 이항반응변수들을 갖는 서로 다른 B개의 상기 부트스트랩 데이터 세트들을 생성하는 단계를 포함하고,

상기 생성된 부트스트랩 데이터 세트들은 상기 생성된 부트스트랩 유전자 발현 데이터들 및 상기 결정된 이항반응변수들을 포함하는, 방법.

청구항 5

제 4 항에 있어서,

상기 (a) 단계는

(a4) 단일변수(univariate) 로지스틱 회귀 모델을 이용하여, 상기 추출된 예비 데이터들에 포함된 유전자들 중 t개 (t는 자연수)의 유전자들을 결정하는 단계를 더 포함하고,

상기 (a2) 단계는

상기 부트스트랩 유전자 발현 데이터들에 대하여 상기 결정된 t개의 유전자들을 갖는 다중변수(multivariate) 로지스틱 회귀 모델을 적용시킴으로써, 상기 이항반응변수들을 결정하는, 방법.

청구항 6

제 5 항에 있어서,

상기 (a4) 단계는

상기 단일변수 로지스틱 회귀 모델에 의해 분석된, 상기 추출된 예비 데이터들에 포함된 유전자들 각각에 대응되는 효과 크기(effect size)에 기초하여, 상기 t개의 유전자들을 결정하는, 방법.

청구항 7

제 4 항에 있어서,

상기 (a2) 단계는

베르누이 시행(Bernoulli's trials)을 이용하여 상기 이항반응변수들을 결정하는, 방법.

청구항 8

제 1 항에 있어서,

상기 (b) 단계는

(b1) 상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 생성하는 단계; 및

(b2) 상기 생성된 예측 모델들에 대한 타당성(validity)을 결정하는 단계를 포함하는, 방법.

청구항 9

제 8 항에 있어서,

상기 (b1) 단계는

로지스틱 회귀 모델을 이용하여 상기 예측 모델들을 생성하는, 방법.

청구항 10

제 8 항에 있어서,

상기 (b2) 단계는

상기 생성된 부트스트랩 데이터 세트들 각각에 대하여 k-묶음 교차 검증(k-fold cross-validation)을 수행하는 단계; 및

상기 k-묶음 교차 검증의 수행 결과에 대하여 카이-제곱 검정(chi-square test)을 수행함으로써, 상기 표본수 N에 대한 경험적 검정력을 산출하는 단계를 포함하고,

상기 경험적 검정력은

상기 카이-제곱 검정의 수행 결과, 상기 생성된 부트스트랩 데이터 세트들 중, 상기 결정된 이항반응변수들에 대한 상기 예측 모델들의 예측의 정확도를 나타내는 통계적 확률이 소정의 유의 수준을 만족하는 부트스트랩 데이터 세트들의 분포 비율에 기초하여, 산출되는, 방법.

청구항 11

제 1 항에 있어서,

상기 (c) 단계는

이진 탐색 알고리즘(binary search algorithm)을 이용하여 상기 (a) 내지 (b) 단계들을 반복함으로써, 상기 최

적의 표본수를 탐색하는, 방법.

청구항 12

제 1 항 내지 제 11 항 중에 어느 한 항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

청구항 13

피검체들에 대한 마이크로어레이 데이터들을 분석하는 컴퓨팅 장치에 있어서,

피검체들에 대한 마이크로어레이 데이터들을 획득하는 데이터 획득부;

상기 획득된 마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응변수들을 갖는 표본수 N (N은 자연수)의 복수의 부트스트랩 데이터 세트들을 생성하는 부트스트랩부;

상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 상기 표본수 N에 대한 경험적 검정력을 산출하는 분석부; 및

상기 검정된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수가 결정될 때까지, 상기 부트스트랩 데이터 세트들의 생성 및 상기 경험적 검정력의 산출이 반복적으로 수행되도록 제어하는 표본수 결정부를 포함하는, 컴퓨팅 장치.

청구항 14

제 13 항에 있어서,

상기 이항반응변수들은

상기 부트스트랩 데이터 세트들에 포함된 상기 표본수 N에 대한 유전자 발현량 데이터를 이용하여 결정되는, 컴퓨팅 장치.

청구항 15

제 14 항에 있어서,

상기 분석부는

상기 결정된 이항반응변수들에 대한, 상기 예측 모델들의 예측의 정확도를 나타내는 통계적 확률에 기초하여 상기 경험적 검정력을 산출하는, 컴퓨팅 장치.

청구항 16

제 13 항에 있어서,

상기 부트스트랩부는

상기 마이크로어레이 데이터들로부터 추출된 복수 개의 예비 데이터들(pilot data)의 통계적 특성들 및 상기 예비 데이터들로부터 랜덤하게 추출된 유전자 발현 데이터들을 이용하여, 상기 표본수 N을 갖는 부트스트랩 유전자 발현 데이터들을 생성하는 재샘플링부; 및

상기 부트스트랩 유전자 발현 데이터들을 이용하여, 상기 표본수 N에 대한 상기 이항반응변수들을 결정하는 변수 결정부를 포함하고,

상기 재샘플링부 및 상기 변수 결정부는

상기 부트스트랩 유전자 발현 데이터들의 생성 및 상기 이항반응변수들의 예측을 B회 (B는 자연수) 반복함으로써, 상기 결정된 이항반응변수들을 갖는 서로 다른 B개의 상기 부트스트랩 데이터 세트들을 생성하고,

상기 생성된 부트스트랩 데이터 세트들은 상기 생성된 부트스트랩 유전자 발현 데이터들 및 상기 결정된 이항반응변수들을 포함하는, 컴퓨팅 장치.

청구항 17

제 16 항에 있어서,

상기 부트스트랩부는

단일변수(univariate) 로지스틱 회귀 모델을 이용하여, 상기 추출된 예비 데이터들에 포함된 유전자들 중 t개(t는 자연수)의 유전자들을 결정하는 제 1 예측 모델 생성부를 더 포함하고,

상기 변수 예측부는

상기 부트스트랩 유전자 발현 데이터들에 대하여 상기 결정된 t개의 유전자들을 갖는 다중변수(multivariate) 로지스틱 회귀 모델을 적용시킴으로써, 상기 이항반응변수들을 결정하는, 컴퓨팅 장치.

청구항 18

제 17 항에 있어서,

상기 제 1 예측 모델 생성부는

상기 단일변수 로지스틱 회귀 모델에 의해 분석된, 상기 추출된 예비 데이터들에 포함된 유전자들 각각에 대응되는 효과 크기(effect size)에 기초하여, 상기 t개의 유전자들을 결정하는, 컴퓨팅 장치.

청구항 19

제 13 항에 있어서,

상기 분석부는

상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 생성하는 제 2 예측 모델 생성부; 및

상기 생성된 예측 모델들에 대한 타당성(validity)을 결정하는 검정부를 포함하는, 컴퓨팅 장치.

청구항 20

제 19 항에 있어서,

상기 검정부는

상기 생성된 부트스트랩 데이터 세트들 각각에 대하여 k-묶음 교차 검증(k-fold cross-validation)을 수행하는 교차 검증부; 및

상기 k-묶음 교차 검증의 수행 결과에 대하여 카이-제곱 검정(chi-square test)을 수행함으로써, 상기 표본수 N에 대한 경험적 검정력을 산출하는 검정력 산출부를 더 포함하고,

상기 경험적 검정력은

상기 카이-제곱 검정의 수행 결과, 상기 생성된 부트스트랩 데이터 세트들 중, 상기 결정된 이항반응변수들에 대한 상기 예측 모델들의 예측의 정확도를 나타내는 통계적 확률이 소정의 유의 수준을 만족하는 부트스트랩 데이터 세트들의 분포 비율에 기초하여, 산출되는, 컴퓨팅 장치.

명 세 서

기 술 분 야

[0001] 피검체들에 대한 마이크로어레이 데이터를 분석하는 방법 및 장치에 관한다.

배 경 기 술

[0002] 유전체(genome)란 한 생물이 가지는 모든 유전 정보를 말한다. 어느 한 개인의 유전체를 서열화(sequencing)하는 기술은 DNA 칩 및 차세대 서열화(Next Generation Sequencing) 기술, 차차세대 서열화(Next Next Generation Sequencing) 기술 등 여러 기술들이 개발되고 있다. 핵산 서열, 단백질 등과 같은 유전 정보들은 분석은 당뇨병, 암과 같은 질병을 발현시키는 유전자를 찾거나, 유전적 다양성과 개체의 발현 특성 간의 상관관계 등을 파악하기 위하여 폭넓게 활용된다. 특히, 개인으로부터 수집된 유전 정보들은 서로 다른 증상이나 질병의 진행과 관련된 개인의 유전적인 특징을 규명하는데 있어서 중요하다. 따라서, 개인의 핵산 서열, 단백질 등과

같은 유전 정보는 현재와 미래의 질병 관련 정보를 파악하여 질병을 예방하거나 질병의 초기 단계에서 최적의 치료 방법을 선택할 수 있도록 하는 핵심적인 데이터이다.

[0003] 임상 연구에서 마이크로어레이(microarray) 등 고차원(high-dimensional)자료를 사용하는 목적 중 하나는 생물학적 표지자(biomarker) 후보군을 도출하고, 이를 이용한 임상반응변수(clinical response variable)의 통계적 예측모델을 생성하는 것이다. 환자의 생존, 재발, 전이 여부를 비롯하여 약물의 반응성 등을 검사하는 데까지 마이크로어레이 등이 널리 사용되고 있다. 최근, Affymetrix Gene-Chip™은 마아크로어레이 최초로 임상적 진단 테스트 키트로 미국 식품의약국(U.S. Food and Drug Administration, U.S. FDA)의 승인을 받았으며, Illumina 사(社)도 임상적 서열분석(clinical sequencing) 장비를 포함하여 FDA의 허가를 앞두고 있다. 이는 향후 마이크로어레이 등의 high-throughput 기술을 이용한 임상연구가 더욱 활발하게 이루어질 것임을 반증한다.

발명의 내용

해결하려는 과제

[0004] 피검체들에 대한 마이크로어레이 데이터를 분석하는 방법 및 장치를 제공하는데 있다. 또한, 상기 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는 데 있다. 본 실시예가 해결하려는 기술적 과제는 상기된 바와 같은 기술적 과제들로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있다.

과제의 해결 수단

[0005] 일 측면에 따르면, 피검체들에 대한 마이크로어레이 데이터들을 분석하는 방법은, (a) 상기 마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응변수들을 갖는 표본수 N (N은 자연수)의 복수의 부트스트랩 데이터 세트들을 생성하는 단계; (b) 상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 상기 표본수 N에 대한 경험적 검정력을 산출하는 단계; 및 (c) 상기 (a) 내지 (b) 단계들을 반복함으로써, 상기 산출된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수를 탐색하는 단계를 포함하고, 상기 (a) 내지 (c) 단계들은 컴퓨팅 장치에 의해 수행된다.

[0006] 다른 일 측면에 따르면, 상기 마이크로어레이 데이터들의 분석 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공한다.

[0007] 또 다른 일 측면에 따르면, 피검체들에 대한 마이크로어레이 데이터들을 분석하는 컴퓨팅 장치는, 피검체들에 대한 마이크로어레이 데이터들을 획득하는 데이터 획득부; 상기 획득된 마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응변수들을 갖는 표본수 N (N은 자연수)의 복수의 부트스트랩 데이터 세트들을 생성하는 부트스트랩부; 상기 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 상기 표본수 N에 대한 경험적 검정력을 산출하는 분석부; 및 상기 검정된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수가 결정될 때까지, 상기 부트스트랩 데이터 세트들의 생성 및 상기 경험적 검정력의 산출이 반복적으로 수행되도록 제어하는 표본수 결정부를 포함한다.

발명의 효과

[0008] 상기된 바에 따르면, 질병, 약물 반응성 등에 대한 임상반응변수와 같은 예후를 예측하고자 할 때 요구되는 적절한 표본수(sample size)를 합리적인 근거를 통해 산출해 낼 수 있다.

도면의 간단한 설명

[0009] 도 1은 본 발명의 일 실시예에 따른 마이크로어레이 데이터 분석 시스템(100)을 도시한 도면이다.

도 2a는 본 발명의 일 실시예에 따른 마이크로어레이 데이터를 분석하는 컴퓨팅 장치(10)의 구성도이다.

도 2b은 본 발명의 일 실시예에 따른 프로세서(120)의 상세 구성도이다.

도 3은 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들(310), 예비 데이터들(320) 및 부트스트랩 데이터 세트들(330)의 관계를 도시한 도면이다.

도 4는 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들을 이용하여 최적의 표본수 N_{opt}을 탐색하는 과

정을 설명하기 위한 도면이다.

도 5는 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들을 이용하여 최적의 표본수 N_{opt}을 탐색하는 상세 과정을 설명하기 위한 도면이다.

도 6은 본 발명의 일 실시예에 따른 피검체들에 대한 마이크로어레이 데이터들을 분석하는 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0010] 이하에서는 도면을 참조하여 본 발명의 실시예들을 상세히 설명하도록 하겠다.

[0011] 도 1은 본 발명의 일 실시예에 따른 마이크로어레이 데이터 분석 시스템(100)을 도시한 도면이다. 도 1을 참고하면, 마이크로어레이 데이터 분석 시스템(100)은 컴퓨팅 장치(10)와 피검체 집단(1)의 유전 정보를 분석하기 위한 다수의 마이크로어레이들(2)을 포함한다. 여기서, 피검체 집단(1)은 암, 종양 등의 질병을 앓고 있는 환자들 또는 정상인들이 포함된 개인들의 집단이다. 또는 실험 대상의 동물들의 집단일 수 있다.

[0012] 도 1에는 비록 도시되지 않았지만, 마이크로어레이 데이터 분석 시스템(100)에서는 피검체 집단(1)으로부터 유전자 발현 패턴(gene expression pattern) 또는 유전자 발현 레벨(gene expression level) 등을 검출하기 위한, High Content Cell Imaging 장치, High Content Screening 장치 또는 High Throughput Screening 장치와 같은 이미지 분석 장치들이 추가로 포함될 수 있고, 마이크로어레이들(2) 대신에 중합효소 연쇄 반응(polymerase chain reaction, PCR) 장치 등도 사용될 수 있음을 당해 기술분야의 통상의 기술자라면 이해할 수 있다.

[0013] 즉, 도 1에 도시된 마이크로어레이 데이터 분석 시스템(100)은 본 실시예의 특징이 흐려지는 것을 방지하기 위하여 본 실시예에 관련된 구성요소들만이 도시되어 있으나, 도 1에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있다.

[0014] 개체의 DNA(DeoxyriboNucleic Acid)와 같은 핵산은 개체의 유전 정보를 포함하는 유전 물질, 즉 유전자에 해당된다. 이와 같은 핵산의 염기서열은 개체를 구성하는 세포, 조직 등에 대한 정보를 포함한다. 따라서, 개인의 완전한 핵산 서열의 정보에 대한 연구는 생명 현상을 이해하고, 신약의 개발, 질병의 진단 및 예방이나, 인간의 유전 연구 등과 같은 많은 분야에서 많이 수행되고 있다.

[0015] 이러한 생물학적 연구에는 환자의 임상시료(clinical specimen)가 사용된다. 생검(biopsy)이나 내시경, 수술 등의 방법으로 얻어진 임상시료는 병리학적 진단에 우선 사용한 후에 환자의 동의를 받고 나서야 연구용으로 사용할 수 있으므로 그 가치가 매우 높고, 희소한 경우가 대부분이다. 따라서, 연구 목적에 부합되는 적절한 표본수(sample size)를 추정하는 것이 중요하다. 너무 많은 표본수에 대해 실험하는 것은 표본들을 낭비하는 것이 될 수 있을 뿐만 아니라, 임상적 유용성이 없는 결과를 만들어 낼 수 있다. 반대로, 너무 적은 수의 표본들로 실험하는 것은 많은 무의미한 과학적 결론을 초래할 수 있다. 특히, 의학이나 임상분야는 사람을 대상으로 시험해야 하므로, 불필요하게 많은 사람들을 모집하여 검증되지 않은 방법으로 실험하는 것은 비윤리적인 실험에 해당될 수 있지만, 너무 적은 사람들로 과학적인 결론에 도달할 수 없는 경우 또한 윤리에 어긋날 수 있다. 따라서, 임상시험심사위원회(Institutional Review Board, IRB)에서 심사하는 연구계획서에서 통계적 근거를 갖는 표본수의 계산은 매우 중요하게 다루어지고 있는 심사항목이다.

[0016] 종래에는, 마이크로어레이 등을 이용하여 생물학적 표지자(biomarker) 등을 이용하여 질병, 약물 반응성에 대한 임상반응변수 등의 예후를 예측하고자 할 때, 적절한 표본수를 어떻게 산출할 것인지에 대해서는 거의 연구된 바가 없다. 이는 전체의 실험 과정이 완료되어야만 비로소, 예측 모델(prediction model)에 포함될 독립 변수(유전자)를 결정할 수 있기 때문이다.

[0017] 본 실시예에 따른 마이크로어레이 데이터 분석 시스템(100)에서는, 컴퓨팅 장치(10)를 이용하여 피검체 집단(1)에 대한 마이크로어레이 데이터들을 분석함으로써, 암, 종양 등의 질병 연구에 필요한 예측 모델의 생성에 필요한 표본수를 결정해 낼 수 있다. 이하에서는, 피검체 집단(1)에 대한 마이크로어레이 데이터들을 분석하여, 예측 모델의 생성에 필요한 표본수를 결정하기 위한, 컴퓨팅 장치(10)의 동작 및 기능에 대해 상세하게 설명하도록 하겠다.

[0018] 도 2a는 본 발명의 일 실시예에 따른 마이크로어레이 데이터를 분석하는 컴퓨팅 장치(10)의 구성도이다.

[0019] 도 2a를 참고하면, 컴퓨팅 장치(10)는 데이터 획득부(110) 및 프로세서(120)를 포함하고, 프로세서(120)는 부트

스트랩부(bootstrapping unit)(122), 분석부(124) 및 표본수 결정부(126)를 포함한다. 이와 같이, 부트스트랩부(bootstrapping unit)(122), 분석부(124) 및 표본수 결정부(126)를 포함하는 프로세서(120)는 적어도 하나의 프로세서로 구현될 수 있는 장치로서, 다수의 논리 게이트들의 어레이, 또는 범용적인 마이크로프로세서와 이 마이크로프로세서에서 실행될 수 있는 프로그램이 저장된 메모리의 조합의 형태로도 구현될 수도 있다. 또한, 프로세서(120)는 응용 프로그램의 모듈 형태로 구현될 수도 있다. 나아가서, 컴퓨팅 장치(10)는 본 실시예에서 설명할 동작들을 구현할 수 있는 다른 형태의 하드웨어로도 구현될 수 있음을 본 실시예가 속하는 기술분야의 통상의 기술자라면 이해할 수 있다.

[0020] 도 2b은 본 발명의 일 실시예에 따른 프로세서(120)의 상세 구성도이다.

[0021] 도 2b를 참고하면, 프로세서(120)는 앞서 설명한 바와 같이, 부트스트랩부(122), 분석부(124) 및 표본수 결정부(126)를 포함한다. 부트스트랩부(122)는 제 1 예측 모델 생성부(1221), 재샘플링부(resampling unit)(1223) 및 변수 결정부(1225)를 포함하고, 분석부(124)는 제 2 예측 모델 생성부(1241) 및 검정부(testing unit)(1243)를 포함한다. 여기서, 검정부(1243)는 교차 검증부(cross-validating unit)(1247) 및 검정력 산출부(power calculating unit)(1249)를 더 포함할 수 있다.

[0022] 이하에서는, 컴퓨팅 장치(10)의 동작 및 기능에 관하여 도 2a 및 2b를 연계하여 설명하도록 하겠다. 한편, 도 2a 및 2b에 도시된 컴퓨팅 장치(10)는 본 실시예의 특징이 흐려지는 것을 방지하기 위하여 본 실시예에 관련된 구성요소들만이 도시되어 있을 뿐이므로, 도 2a 및 2b에 도시된 구성요소들 외에 다른 범용적인 구성요소들이 더 포함될 수 있다.

[0023] 데이터 획득부(110)는 피검체 집단(1)에 대한 마이크로어레이 데이터들을 획득한다. 마이크로어레이 데이터들은 앞서 설명한 바와 같이, 유전자 발현 패턴 또는 유전자 발현 레벨 등의 검출 결과로서, High Content Cell Imaging 장치, High Content Screening 장치 또는 High Throughput Screening 장치와 같은 이미지 분석 장치들에 의해 피검체 집단(1)의 유전 정보가 분석된 결과에 해당된다.

[0024] 한편, 데이터 획득부(110)는, NCBI (National Center for Biotechnology Information)에서 운영하는 GEO (Gene Expression Omnibus)의 데이터베이스, 또는 NCI (National Cancer Institute) 및 NHGRI (National Human Genome Research Institute)에 의해 운영되는 TCGA (The Cancer Genome Atlas)의 데이터베이스 등으로부터 피검체 집단(1)의 마이크로어레이 데이터들을 획득할 수도 있다. 또는, 데이터 획득부(110)는, 직접 마이크로어레이를 이용하여 실험한 결과를 통해서도 마이크로어레이 데이터들을 획득할 수 있다. 즉, 연구의 대상과 목적, 실험방식(platform)이 동일하다면 어떠한 방식에 의해 제한되지 않는다.

[0025] 피검체 집단(1)에 대한 마이크로어레이 데이터들은 예를 들어, 다음의 표 1과 같이 피검체들 A, B, C 등에 대한 유전자 a, b, c 등의 유전자 발현량들의 데이터 형태로 획득될 수 있으나, 본 실시예는 이에 한정되지 않는다.

표 1

이항반응변수	0	1	1	· · ·
유전자 \ 피검체	Subject A	Subject B	Subject C	· · ·
Gene a	0.00001	0.01433	0.01232	· · ·
Gene b	0.00105	0.00133	0.00231	· · ·
Gene c	0.00035	0.00022	0.00004	· · ·
· · ·	· · ·	· · ·	· · ·	· · ·

<유전자 발현량>

[0026]

[0027] 부트스트랩부(122)의 제 1 예측 모델 생성부(1221)는 피검체 집단(1)에 대한 마이크로어레이 데이터들에 대하여 부트스트랩(bootstrapping)을 수행하기 전, 획득된 마이크로어레이 데이터들을 전처리(pre-processing)한다.

[0028] 우선, 부트스트랩부(122)는 피검체 집단(1) 전체에 대한 마이크로어레이 데이터들 중에서, 일부의(예를 들어, n명) 마이크로어레이 데이터들을 예비 데이터들(pilot data)로서 추출한다.

[0029] 예비 데이터들(pilot data)은, 총 n명의 피검체들의 총 g개의 유전자들에 대하여 x_{ij}를 i번째 피검체 및 j번째 유전자에 대한 유전자 발현량으로 정의함으로써, 수학식 1과 같이 표현될 수 있다.

수학식 1

[0030] $$M = \{(x_{i1}, \cdots, x_{ig}), \quad i = 1, \cdots, n\}$$

[0031] 예비 데이터들에는 n명의 피검체 각각에 대응되는, 특정 조건(specific condition) 또는 특정 반응(specific response)에 대한 이항반응변수들(binary response variables)에 대한 정보가 포함될 수 있다.

[0032] 여기서, 특정 조건 또는 특정 반응은, 종양(암)의 유무, 임파절 전이 양성/음성(lymph node metastasis +/-), 약물의 효과 유무 등을 예로 들 수 있으나, 이에 한정되지 않고, 주어진 데이터에 대하여 양성/음성 또는 유무를 구분시킬 수 있는 다양한 조건들 또는 다양한 반응들을 포함할 수 있다.

[0033] 즉, 예비 데이터들에는, n명의 피검체들 각각에 대한 g개의 유전자들의 유전자 발현 데이터들과 함께, n명의 피검체들 각각에 대한 이항반응변수들이 포함될 수 있다.

[0034] n명의 피검체들 각각에 대응되는 이항반응변수 y_i 는 다음의 수학식 2와 같이 정의될 수 있다.

수학식 2

[0035] $$y_i = \begin{cases} 0, & if \ \ i^{th} \ \ subject \ \ is \ \ normal \\ 1, & if \ \ i^{th} \ \ subject \ \ has \ \ an \ \ event \end{cases}$$

[0036] 예를 들어, 약물반응성 실험인 경우, 이항반응변수가 0인 피검체는 대조군(control group)에 해당되고, 이항반응변수가 1인 피검체는 처리군(treatment group)에 해당되는 것으로 가정할 수 있다. 또한, 임파절 전이 실험인 경우, 이항반응변수가 0인 피검체는 임파절 전이 음성(lymph node metastasis -)에 해당되고, 이항반응변수가 1인 피검체는 임파절 전이 양성(lymph node metastasis +)에 해당되는 것을 가정할 수 있다. 나아가서, 종양에 대한 경우, 이항반응변수가 0인 피검체는 정상(normal)에 해당되고, 이항반응변수가 1인 피검체는 종양 군(tumor group)에 해당되는 것을 가정할 수 있다. 다만, 본 실시예는 이에 제한되지 않는다.

[0037] 제 1 예측 모델 생성부(1221)는 수학식 1과 같은 예비 데이터들을 이용하여, 예비 데이터들에 포함된 이항반응변수들 y_i 을 예측하는 예측 모델을 생성한다. 예측 모델의 생성은, 로지스틱 회귀 모델(logistic regression model)을 이용할 수 있으나, 본 실시예는 이에 제한되지 않고 로지스틱 회귀 모델 외에도 다른 종류의 모델 또는 알고리즘이 이용될 수 있다.

[0038] 먼저, 제 1 예측 모델 생성부(1221)는 예비 데이터들에 포함된 유전자들 각각을 단일 변수(single variable)로 하는, 단일변수 로지스틱 회귀 모델(univariate logistic regression model)을 이용하여 예측 모델을 생성한다.

[0039] 이 때, 제 1 예측 모델 생성부(1221)는 다음의 수학식 3과 같이 예비 데이터들을 표준화(normalize)하여 단일변수 로지스틱 회귀 모델을 이용할 수 있다.

수학식 3

$$X'_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}$$

$$where \quad \overline{X}_j = \sum_{i=1}^{n} X_{ij} / n, \quad S_j = \sqrt{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2 / (n-1)}$$

[0040]

[0041] 수학식 3을 참고하면, $\overline{X}_j$ 는 j번째 유전자의 유전자 발현량들의 평균, S_j는 j번째 유전자의 유전자 발현량들의 표준편차이다.

[0042] 제 1 예측 모델 생성부(1221)는 생성된 단일변수 로지스틱 회귀 모델을 이용하여, 예비 데이터들에 포함된 유전자들 각각에 대한 효과 크기(effect size, $\hat{\beta}_j$)와 함께 귀무가설 $H_0 : \beta_j = 0$ 의 검정 결과인 확률 P-값을 산출한다.

[0043] 제 1 예측 모델 생성부(1221)는 유전자들 각각에 대한 확률 P-값을 이용하여, 상위 t개의 유전자들을 결정한다. 이와 같이, 제 1 예측 모델 생성부(1221)에 의해 결정된 t개의 유전자들은 0 또는 1의 이항반응변수가 결정되는 데에 많은 영향을 미치는 유전자들에 해당될 수 있다. 한편, 제 1 예측 모델 생성부(1221)에서 총 g개의 유전자들 중 몇 개의 유전자들을 상위 t개의 유전자들로 결정할 것인지는 본 실시예의 사용 환경에 따라 다양하게 변경될 수 있다.

[0044] 제 1 예측 모델 생성부(1221)는 결정된 상위 t개의 유전자들을 독립변수들(independent variable)로 설정하는, 다중변수 로지스틱 회귀 모델(multivariate logistic regression model)을 생성한다. 이는 수학식 4를 참고하여 설명하도록 하겠다.

수학식 4

$$logit(p_i) = log(\frac{p_i}{1 - p_i}) = \hat{\beta}_0 + \hat{\beta}_1 X'_{i1} + \cdots + \hat{\beta}_t X'_{it}$$

[0045]

[0046] 수학식 4는 예비 데이터들을 이용하여 생성된, i번째 환자에 대한 다중변수 로지스틱 회귀 모델에 해당된다. 수학식 4를 참고하면, $\hat{\beta} = (\hat{\beta}_0, \ \hat{\beta}_1, \ \cdots, \ \hat{\beta}_t)$ 는 t개의 유전자들에 대한 다중변수 로지스틱 회귀 모델의 계수들을 의미하고, $X' = (X'_1, \ \cdots, \ X'_t)$ 는 t개의 유전자들에 대응되는 표준화된 유전자 발현량들을 의미한다.

[0047] 여기서, 예비 데이터들에 대한 다중변수 로지스틱 회귀 모델의 계수들 $\hat{\beta} = (\hat{\beta}_0, \ \hat{\beta}_1, \ \cdots, \ \hat{\beta}_t)$ 은 이하에서 설명될, 부트스트랩 데이터 세트들에 포함될 이항반응변수들을 결정하는데 이용된다.

[0048] 제 1 예측 모델 생성부(1221)에서 예비 데이터들에 대한 수학식 4와 같은 다중변수 로지스틱 회귀 모델의 생성이 완료된 경우, 부트스트랩부(122)는 예비 데이터들에 대한 부트스트랩(bootstrapping)을 수행한다.

[0049] 재샘플링부(1223)는 예비 데이터들의 통계적 특성들 및 예비 데이터들로부터 랜덤하게 추출된 유전자 발현량들

을 이용하여, 표본수 N (N은 자연수)을 갖는 부트스트랩 유전자 발현 데이터들(bootstrap gene expression data)을 생성한다.

[0050] 재샘플링부(1223)는 앞서 설명된 수학식 1과 같은 예비 데이터들 M에 대하여, 유전자 $j=(1,\ \cdots,\ g)$에 대응되는 표본평균 $\overline{X}_j$과 표준편차 S_j를 각각 산출한다.

[0051] 재샘플링부(1223)는 모든 $\epsilon_i, ..., \epsilon_N \sim iidN(0,1)$의 확률 변수에 대하여 다음의 수학식 5와 같이, 표본수 N(>n)을 갖는 부트스트랩 데이터 세트를 생성한다.

수학식 5

$$\widetilde{M}=\{(z_{i1},\ \cdots,\ z_{ig}),\quad i=1,\ \cdots,\ N\}$$

$$where\quad z_{ij}=\frac{(x_{i'j}-\overline{x_j})}{s_j},$$

$$i' \text{ is randomly chosen number from } (1,\ \cdots,\ n)$$

[0052]

[0053] 예비 데이터들 M이 주어진 경우, 생성된 부트스트랩 데이터 세트 $\widetilde{M}$의 조건부 공분산 구조는 M의 공분산 구조와 근사적으로 동일하고, 이는 수학식 6과 같이 설명할 수 있다.

수학식 6

$$\mathrm{cov}(\widetilde{M}\mid M)\rightarrow \mathrm{cov}(M),\quad as\quad n\rightarrow\infty$$

[0054]

[0055] 재샘플링부(1223)는 위와 같이 설명된 과정들을 통해, 예비 데이터들 M을 이용하여, 부트스트랩 데이터 세트 $\widetilde{M}$을 생성한다. 다만, 재샘플링부(1223)는 부트스트랩 데이터 세트 $\widetilde{M}$에 포함될 부트스트랩 유전자 발현 데이터들만 생성할 뿐이고, 부트스트랩 데이터 세트 $\widetilde{M}$에 포함될 이항반응변수들은 변수 결정부(1225)에서 결정된다.

[0056] 변수 결정부(1225)는 재샘플링부(1223)에서 생성된 부트스트랩 유전자 발현 데이터들을 이용하여, 부트스트랩 유전자 발현 데이터들에 대한 이항반응변수들을 결정한다.

[0057] 변수 결정부(1225)는 앞서 설명된, 예비 데이터들에 관한 수학식 4의 예측 모델을 이용하여, 부트스트랩 데이터 세트 $\widetilde{M}$에 포함된 N개의 표본들의 위험도(risk score)를 산출한다. 부트스트랩 데이터 세트 $\widetilde{M}$에 포함된 N개의 표본들의 위험도는 수학식 7과 같은 확률로 계산될 수 있다.

수학식 7

$$\hat{p}_i = P(y_i = 1 \mid z) = \frac{1}{1 + \exp\{-(\hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \cdots + \hat{\beta}_t z_{it})\}}$$

[0058]

[0059] 수학식 7을 참고하면, $\hat{\beta} = (\hat{\beta}_0,\ \hat{\beta}_1,\ \cdots,\ \hat{\beta}_t)$ 은 앞서 설명된, 예비 데이터들에 대한 다중변수 로지스틱 회귀 모델의 계수들이다.

[0060] 변수 결정부(1225)는 확률 $\hat{p}_i$ 를 갖는 수학식 8과 같은 베르누이 시행(Bernoulli's trials)을 이용하여, 부트스트랩 데이터 세트 $\widetilde{M}$ 에서 N개의 표본들 각각에 대응되는 이항반응변수들 $y_i' \quad (i = 1,\ \cdots,\ N)$ 을 산출한다.

수학식 8

$$y_i' \sim \quad Bernoulli(\hat{p}_i)$$

[0061]

[0062] 결국, 부트스트랩부(122)는 재샘플링부(1223)에서 생성된 부트스트랩 유전자 발현 데이터들 $\widetilde{M} = \{(z_{i1},\ \cdots,\ z_{ig}),\quad i = 1,\ \cdots,\ N\}$ 및 변수 결정부(1225)에서 결정된 이항반응변수들 $y_i' \quad (i = 1,\ \cdots,\ N)$ 을 포함하는, b번째의 부트스트랩 데이터 세트 $\widetilde{M}_b$을 다음의 수학식 9와 같이 생성한다.

수학식 9

$$\widetilde{M}_b = \{y_i',\ (z_{i1},\ \cdots,\ z_{ig})\}$$
$$where \quad i = (1,\ \cdots,\ N), \quad b = (1,\ \cdots,\ B)$$

[0063]

[0064] 앞서 설명된 바는, 부트스트랩부(122)에서 1회의 부트스트랩 과정을 수행하는 경우에 대해서 설명하였으나, 수학식 9에서 설명된 바와 같이, 부트스트랩부(122)는 앞서 설명된 부트스트랩 과정을 B회 반복함으로써, 서로 다른 B개의 부트스트랩 데이터 세트들 $\widetilde{M}_b \quad (b = 1,\ \cdots,\ B)$ 을 생성한다.

[0065] 도 3은 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들(310), 예비 데이터들(320) 및 부트스트랩 데이터 세트들(330)의 관계를 도시한 도면이다.

[0066] 도 3을 참고하면, 예를 들어 환자 168명에 대한 마이크로어레이 데이터들(310)이 존재하는 경우를 가정할 수 있고, 이는 데이터 획득부(110)에 의해 획득된다.

[0067] 부트스트랩부(122)는 이 중에서 일부만을 임의로 추출하여, 예비 데이터들(320)을 생성한다. 예비 데이터(320)는 환자 30명(n=30)에 대한 데이터, 또는 환자 50명(n=50)에 대한 데이터 등 다양한 인원수(n)에 대한 데이터일

수 있고, 어느 하나에 의해 제한되지 않는다.

[0068] 환자 30명(n=30)의 예비 데이터(321)를 이용하는 경우를 예로 들면, 부트스트랩부(122)는 예비 데이터(321)를 이용하여 N개(N=100)의 표본수를 갖는 부트스트랩 데이터 세트들(330)을 B개(B=1000) 생성할 수 있다.

[0069] 도 3은 마이크로어레이 데이터들(310), 예비 데이터들(320) 및 부트스트랩 데이터 세트들(330)의 관계를 개략적으로 설명하기 위한 것일 뿐이므로, 도 3에서 설명된 수치들은 임의의 수치들에 해당된다.

[0070] 다시 도 2b를 참고하면, 분석부(124)는 부트스트랩부(122)에서 생성된 B개의 부트스트랩 데이터 세트들 $\tilde{M}_b \ (b=1, \cdots, B)$ 각각에 대응되는 예측 모델들을 이용하여, 경험적 검정력(empirical power)을 산출한다.

[0071] 보다 상세하게 설명하면, 우선, 분석부(124)의 제 2 예측 모델 생성부(1241)는 앞서 예비 데이터들과 유사한 방식으로, 단일변수 로지스틱 회귀 모델 및 다중변수 로지스틱 회귀 모델을 이용하여, b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 대응되는 예측 모델을 생성한다. 여기서, 예측 모델은 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 포함된 표본수 N의 유전자 발현량들을 이용하여 이항반응변수들 y_i' 을 예측하기 위한 모델일 수 있다.

[0072] 검정부(1243)는 제 2 예측 모델 생성부(1241)에서 생성된, b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 대응되는 예측 모델을 검정한다.

[0073] 검정부(1243)의 교차 검증부(1247)는 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 대하여 k-묶음 교차 검증(k-fold cross-validation)을 수행한다.

[0074] k-묶음 교차 검증은 부트스트랩 데이터 세트를 대략 동일한 수의 데이터들로 구성된 k-묶음(fold)으로 랜덤하게 분류한 후, k-1개의 묶음을 학습 집합(training set)으로, 나머지 1개의 묶음을 검증 집합(testing set)으로 사용하는 검증 방법이다. 그리고, k-묶음 교차 검증은 위와 같은 묶음 과정을 k회 반복하여 학습 집합에서 구성된 예측 모델을 검증 집합에 대해 평가하는 방법이다. 여기서, k-1개의 묶음의 학습 집합에 대한 예측 모델은 제 2 예측 모델 생성부(1241)에 의해 생성될 수 있다.

[0075] 교차 검증부(1247)는 학습 집합에 대해 생성된 다중변수 로지스틱 회귀 모델의 예측 모델에 기초하여 검증 집합에서의 예측 확률 $\hat{p}_i$을 산출한다.

[0076] 검정부(1243)의 검정력 산출부(1249)는 카이-제곱 검정(chi-squre test, χ^2 test)을 이용하여, b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 대한 k-묶음 교차 검증의 결과를 검정한다.

[0077] 보다 상세하게 설명하면, 검정력 산출부(1249)는 N개의 표본들 각각에 대하여 이항반응변수의 예측 값 $\hat{y}_i$을 수학식 10을 이용하여 산출한다. 다만, 수학식 10에서 0.5의 수치는 변경될 수 있다.

수학식 10

[0078]
$$\hat{y}_i = \begin{cases} 0, & if \quad \hat{p}_i < 0.5 \\ 1, & if \quad \hat{p}_i \geq 0.5 \end{cases}$$

[0079] 검정력 산출부(1249)는 위와 같이, 산출된 N개의 이항반응변수들의 분포를 아래의 표 2와 같은 카이-제곱 검정을 이용하여 검정한다.

표 2

		예측값 ($\hat{y}_i$)		합계
		0	1	
실제값 (y_i)	0	n00	n01	n0+
	1	n10	n11	n1+
합계		n+0	n+1	n++ (=N)

[0080]

[0081] 여기서, 실제값 y'_i 는 변수 결정부(1225)에 의해 결정된 이항반응변수의 값이고, 예측값 $\hat{y}_i$ 는 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$ 에 대응되는 예측 모델에 대한 교차 검증의 결과로 예측된 이항반응변수의 값이다.

[0082] 그리고 나서, 검정력 산출부(1249)는 수학식 11과 같은 카이-제곱(χ^2) 검정통계량을 산출한다.

수학식 11

$$\chi^2 = \frac{(n_{00} \cdot n_{11} - n_{01} \cdot n_{10})^2 \cdot n_{++}}{n_{0+} \cdot n_{1+} \cdot n_{+0} \cdot n_{+1}} \sim \chi^2(1)$$

[0083]

[0084] 검정력 산출부(1249)는 수학식 11에 의해 산출된 검정통계량을 이용하여, b번째 부트스트랩 데이터 세트 $\tilde{M}_b$ 에 대한 예측 모델의 타당성(validity)을 나타내는 P-값 p'_b 를 산출한다.

[0085] 그리고 나서, 검정력 산출부(1249)는 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$에 대해 산출된 P-값 p'_b를 소정의 유의 수준 α (예를 들어, 5%)와 비교함으로써, b번째 부트스트랩 데이터 세트 $\tilde{M}_b$가 타당한지 또는 타당하지 않은지 여부를 결정한다. 예를 들어, 검정력 산출부(1249)는, 산출된 P-값 p'_b이 소정의 유의 수준 α 이하인 경우 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$는 타당하고, 산출된 P-값 p'_b이 소정의 유의 수준 α 초과인 경우 b번째 부트스트랩 데이터 세트 $\tilde{M}_b$는 타당하지 않은 것으로 결정할 수 있다. 다만, 이에 제한되지 않는다.

[0086] 검정력 산출부(1249)는 부트스트랩부(122)에서 생성된 B개의 부트스트랩 데이터 세트 $\tilde{M}_b$ 전체 중에서 소정의

유의 수준 α 에 의해 타당한 것으로 결정된 부트스트랩 데이터 세트들의 분포 비율을 산출함으로써, 표본수 N에 대한 경험적 검정력(empirical power)을 산출한다. 여기서, 소정의 유의 수준 α 는 다양하게 변경될 수 있는 값이다.

[0087] 검정력 산출부(1249)는 수학식 13을 이용하여 표본수 N에 대한 경험적 검정력 $(1-\hat{\beta}_N)$ 을 산출할 수 있다.

수학식 13

$$(1-\hat{\beta}_N)=\frac{1}{B}\sum_{b=1}^{B} I(p_b' < \alpha)$$

[0088]

[0089] 표본수 결정부(126)는 목표 검정력을 만족하는지 여부에 따라, 최적의 표본수 N_{opt} 에 해당되는지 여부를 결정한다.

[0090] 예를 들어, 목표 검정력 $(1-\beta)$ 을 0.80으로 가정한 경우, 검정력 산출부(1249)에서 산출된 표본수 N에 대한 경험적 검정력 $(1-\hat{\beta}_N)$ 이 0.80을 초과한다면, 표본수 N은 특정 반응 또는 특정 조건에 대한 실험 또는 연구를 수행하기 위하여 모집될 최적의 표본수 N_{opt} 에 해당되는 것으로 결정한다. 그러나, 검정력 산출부(1249)에서 산출된 표본수 N에 대한 경험적 검정력 $(1-\hat{\beta}_N)$ 이 0.80 이하인 경우, 표본수 N은 최적의 표본수 N_{opt} 에 해당되지 않는 것으로 결정한다.

[0091] 다만, 프로세서(120)는, 표본수 결정부(126)에서 목표 검정력을 초과하는 최적의 표본수 N_{opt}가 결정될 때까지, 이진 탐색 알고리즘(binary search algorithm)을 이용하여 표본수 N을 적절한 값으로 반복적으로 변경하면서, 앞서 설명된 부트스트랩 데이터 세트들의 생성 및 경험적 검정력의 산출 등의 과정들을 반복적으로 수행한다. 다만, 프로세서(120)는, 이진 탐색 알고리즘 외에도 이와 유사한 다른 알고리즘을 이용하여 최적의 표본수 N_{opt} 를 탐색할 수 있다.

[0092] 즉, 프로세서(120)의 부트스트랩부(122), 분석부(124) 및 표본수 결정부(126)는 표본수 N의 값을 점차 변화시키면서 경험적 검정력 $(1-\hat{\beta}_N)$ 을 반복적으로 산출함으로써, 목표 검정력을 만족하는 최소의 표본수 N을 탐색한다. 그리고, 표본수 결정부(126)는 목표 검정력을 만족하는 최소의 표본수 N을 최적의 표본수 N_{opt}로 결정한다.

[0093] 표본수 결정부(126)에 의해 최종적으로 결정된 최적의 표본수 N_{opt} 의 의미는, 앞서 설명하였던, 특정 반응 또는 특정 조건에 대한 실험 또는 연구를 수행하기 위하여 모집될 최적의 표본수를 의미한다.

[0094] 도 4는 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들을 이용하여 최적의 표본수 N_{opt}을 탐색하는 과정을 설명하기 위한 도면이다. 도 4를 참고하면, 본 실시예에 따른 마이크로어레이 데이터 분석 방법은 도 2a 및 2b에 도시된 컴퓨팅 장치(10)에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하 생략된 내용이라 하더라도 앞서 설명한 도면들에 관한 내용은 본 실시예에 따른 표본수 탐색 과정에도 적용될 수 있다.

[0095] 도 4를 참고하면, 부트스트랩부(122)는 우선, 데이터 획득부(110)에서 획득된 마이크로어레이 데이터들로부터 일부 데이터들을 추출하여, 예비 데이터(pilot data)(410)를 생성한다. 이는 앞서 도 3에서 설명하였던, 환자 30명(n=30)에 대한 예비 데이터(321)에 해당될 수 있다.

[0096] 부트스트랩부(122)는 예비 데이터(410)를 이용하여, 표본수 N을 갖는 B개의 부트스트랩 데이터 세트들(421, 422, 423)을 생성한다. 도 4에서는 B=1000인 것으로 가정하였다.

[0097] 분석부(124)는 B개의 부트스트랩 데이터 세트들(421, 422, 423)에 대한 경험적 검정력을 산출하기 위하여, B개의 부트스트랩 데이터 세트들(421, 422, 423)에 대한 검정(431, 432, 433)을 각각 수행한다.

[0098] 부트스트랩 데이터 세트 1(421)을 예로 들어 설명하면, 제 2 예측 모델 생성부(1241)는 로지스틱 회귀 모델을 이용하여 부트스트랩 데이터 세트 1(421)에 대한 예측 모델 1을 생성한다(4311). 그리고, 교차 검증부(1247)는 생성된 예측 모델 1에 대하여 k-묶음 교차 검증을 수행한다(4312). 나아가서, 검정력 산출부(1249)는 k-묶음 교차 검증의 결과를 이용하여 카이-제곱 검정을 수행함으로써, 소정의 유의 수준 α을 기준으로 부트스트랩 데이터 세트 1(421)에 대한 타당성을 결정한다(4313).

[0099] 분석부(124)는 이와 같은 과정들을 나머지 B-1개의 부트스트랩 데이터 세트들(422, 423) 각각에 대해서도 반복하여 수행한다. 그 결과, 나머지 B-1개의 부트스트랩 데이터 세트들(422, 423) 각각에 대하여도 타당성들이 결정된다(4323, 4333).

[0100] 분석부(124)는 B개의 부트스트랩 데이터 세트들(421, 422, 423)에 대해 결정된 타당성들을 이용하여 표본수 N에 대한 경험적 검정력(450)을 산출한다.

[0101] 표본수 결정부(126)는 목표 검정력을 만족하는지 여부에 따라, 최적의 표본수 N_{opt}에 해당되는지 여부를 결정한다.

[0102] 만약, 산출된 경험적 검정력(450)이 목표 검정력을 만족하지 못하는 경우에는, 표본수 N을 다른 값으로 설정하여 부트 스트랩 데이터 세트들(421, 422, 423)을 생성하거나, 또는 다른 종류의 예비 데이터(410)를 이용할 수 있다. 즉, 표본수 결정부(126)에서 경험적 검정력(450)이 목표 검정력을 만족하는 것으로 판단될 때까지, 프로세서(120)는 앞서 설명된 과정들의 설정들을 변경시키면서, 최적의 표본수 N_{opt}을 탐색한다.

[0103] 도 5는 본 발명의 일 실시예에 따른, 마이크로어레이 데이터들을 이용하여 최적의 표본수 N_{opt}을 탐색하는 상세 과정을 설명하기 위한 도면이다. 도 5를 참고하면, 본 실시예에 따른 마이크로어레이 데이터 분석 방법은 도 2a 및 2b에 도시된 컴퓨팅 장치(10)에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하 생략된 내용이라 하더라도 앞서 설명한 도면들에 관한 내용은 본 실시예에 따른 표본수 탐색 과정에도 적용될 수 있다.

[0104] 510 단계에서, 부트스트랩부(122)는 데이터 획득부(110)에서 획득된 마이크로어레이 데이터들로부터 일부 데이터들을 추출하여, 예비 데이터(pilot data)를 생성한다.

[0105] 520 단계에서, 제 1 예측 모델 생성부(1221)는 로지스틱 회귀 모델을 이용하여 예비 데이터에 대한 예측 모델을 생성한다.

[0106] 보다 상세하게 설명하면, 521 단계에서, 제 1 예측 모델 생성부(1221)는 예비 데이터에 포함된 유전자들 각각을 단일 변수로 하는, 단일변수 로지스틱 회귀 모델(univariate logistic regression model)을 생성한다. 그리고, 522 단계에서, 제 1 예측 모델 생성부(1221)는 유전자들 각각에 대한 확률 P-값을 이용하여 상위 t개의 유전자들을 결정하고, 결정된 상위 t개의 유전자들을 독립변수들로 하는, 다중변수 로지스틱 회귀 모델(multivariate logistic regression model)을 생성한다.

[0107] 530 단계에서, 제 1 예측 모델 생성부(1221)는 t개의 유전자들에 대한 다중변수 로지스틱 회귀 모델의 계수들 $\hat{\beta} = (\hat{\beta}_0,\ \hat{\beta}_1,\ \cdots,\ \hat{\beta}_t)$ 을 결정한다.

[0108] 540 단계에서, 재샘플링부(1223)는 예비 데이터들을 이용하여, 표본수 N을 갖는 부트스트랩 유전자 발현 데이터들을 생성한다. 그리고, 변수 결정부(1225)는 재샘플링부(1223)에서 생성된 부트스트랩 유전자 발현 데이터들을 이용하여, 부트스트랩 유전자 발현 데이터들에 대한 이항반응변수들을 결정한다.

[0109] 즉, 부트스트랩부(122)는 재샘플링부(1223)에서 생성된 부트스트랩 유전자 발현 데이터들 및 변수 결정부(1225)에서 결정된 이항반응변수들을 포함하는, b번째의 부트스트랩 데이터 세트를 생성한다.

[0110] 550 단계에서, 교차 검증부(1247)는 b번째 부트스트랩 데이터 세트에 포함된 유전자 발현 데이터들을 k-묶음으로 분류하여, k-묶음 교차 검증(k-fold cross-validation)을 수행한다.

[0111] 560 단계에서, 제 2 예측 모델 생성부(1241)는 k-1개의 묶음의 학습 집합(training set)에 대하여 예측 모델을 생성한다.

[0112] 보다 상세하게 설명하면, 561 단계에서, 제 2 예측 모델 생성부(1241)는 학습 집합에 포함된 유전자들 각각을 단일 변수로 하는, 단일변수 로지스틱 회귀 모델을 생성한다. 그리고, 562 단계에서, 제 2 예측 모델 생성부(1241)는 유전자들 각각에 대한 확률 값을 이용하여 상위 t개의 유전자들을 결정하고, 결정된 상위 t개의 유전자들을 독립변수들로 하는, 다중변수 로지스틱 회귀 모델을 생성한다.

[0113] 570 단계에서, 교차 검증부(1247)는 562 단계에서 생성된 다중변수 로지스틱 회귀 모델을 이용하여 나머지 1개의 검증 집합(testing set)을 예측한다.

[0114] 580 단계에서, 검정력 산출부(1249)는 카이-제곱 검정(chi-squre test, χ^2 test)을 이용하여, b번째 부트스트랩 데이터 세트에 대한 k-묶음 교차 검증의 결과를 검정한다.

[0115] 590 단계에서, 검정력 산출부(1249)는 검정 결과, b번째 부트스트랩 데이터 세트에 대한 예측 모델의 타당성을 나타내는 P-값을 산출한다.

[0116] 595 단계에서, 표본수 결정부(126)는 B개의 부트스트랩 데이터 세트들에 대해 산출된 경험적 검정력을 이용하여, 목표 검정력을 만족하는지 여부에 따라 최적의 표본수 N_{opt} 인지를 결정한다.

[0117] 만약, 산출된 경험적 검정력이 목표 검정력을 만족하지 못하는 경우에는, 표본수 N을 다른 값으로 설정하여 부트 스트랩 데이터 세트들을 생성하거나, 또는 다른 종류의 예비 데이터를 이용할 수 있다. 즉, 표본수 결정부(126)에서 경험적 검정력이 목표 검정력을 만족하는 것으로 판단될 때까지, 프로세서(120)는 앞서 설명된 과정들의 설정들을 변경시키면서, 최적의 표본수 N_{opt} 을 탐색한다.

[0118] 도 6은 본 발명의 일 실시예에 따른 피검체들에 대한 마이크로어레이 데이터들을 분석하는 방법의 흐름도이다. 도 6을 참고하면, 본 실시예에 따른 마이크로어레이 데이터 분석 방법은 도 2a 및 2b에 도시된 컴퓨팅 장치(10)에서 시계열적으로 처리되는 단계들로 구성된다. 따라서, 이하 생략된 내용이라 하더라도 앞서 설명한 도면들에 관한 내용은 본 실시예에 따른 마이크로어레이 데이터 분석 방법에도 적용될 수 있다.

[0119] 601 단계에서, 부트스트랩부(122)는 마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응변수들을 갖는 표본수 N의 복수의 부트스트랩 데이터 세트들을 생성한다.

[0120] 602 단계에서, 분석부(124)는 생성된 부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 표본수 N에 대한 경험적 검정력을 산출한다.

[0121] 603 단계에서, 표본수 결정부(126)는 601 단계 내지 602 단계들이 반복하여 수행되도록 제어함으로써, 산출된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수를 탐색한다.

[0122] 한편, 상술한 본 발명의 실시예들은 컴퓨터에서 실행될 수 있는 프로그램으로 작성 가능하고, 컴퓨터로 읽을 수 있는 기록매체를 이용하여 상기 프로그램을 동작시키는 범용 디지털 컴퓨터에서 구현될 수 있다. 또한, 상술한 본 발명의 실시예에서 사용된 데이터의 구조는 컴퓨터로 읽을 수 있는 기록매체에 여러 수단을 통하여 기록될

수 있다. 상기 컴퓨터로 읽을 수 있는 기록매체는 마그네틱 저장매체(예를 들면, 롬, 플로피 디스크, 하드 디스크 등), 광학적 판독 매체(예를 들면, 시디롬, 디브이디 등)와 같은 저장매체를 포함한다.

[0123] 이제까지 본 발명에 대하여 그 바람직한 실시예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

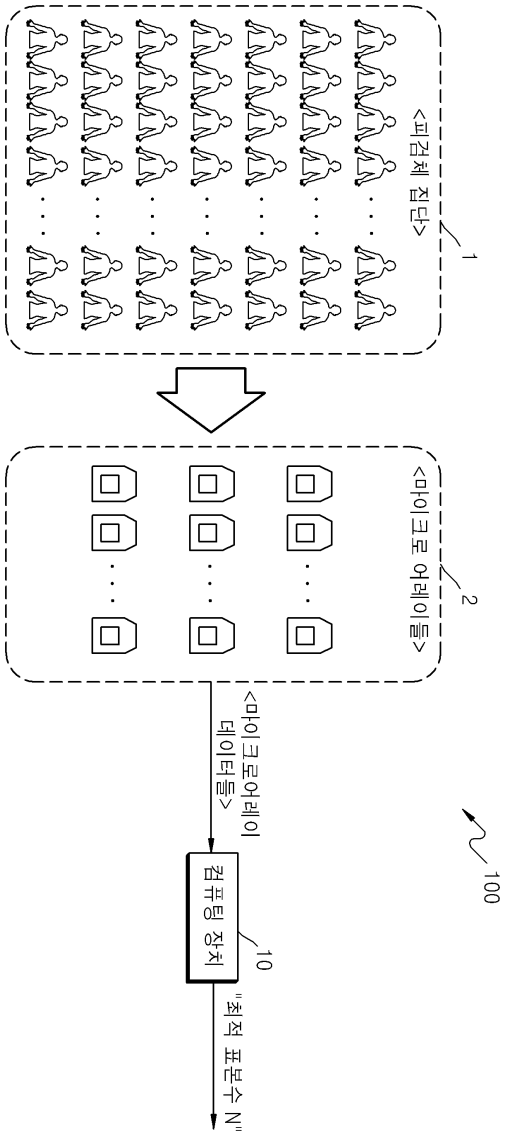
부호의 설명

[0124]

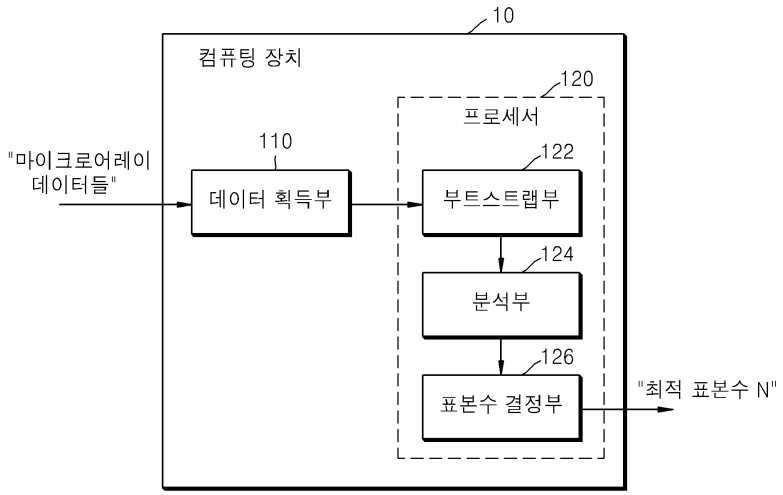
100:	마이크로어레이 데이터 분석 시스템	1:	피검체 집단
2:	마이크로어레이들	10:	컴퓨팅 장치
110:	데이터 획득부	120:	프로세서
122:	부트스트랩부	124:	분석부
126:	표본수 결정부		

도면

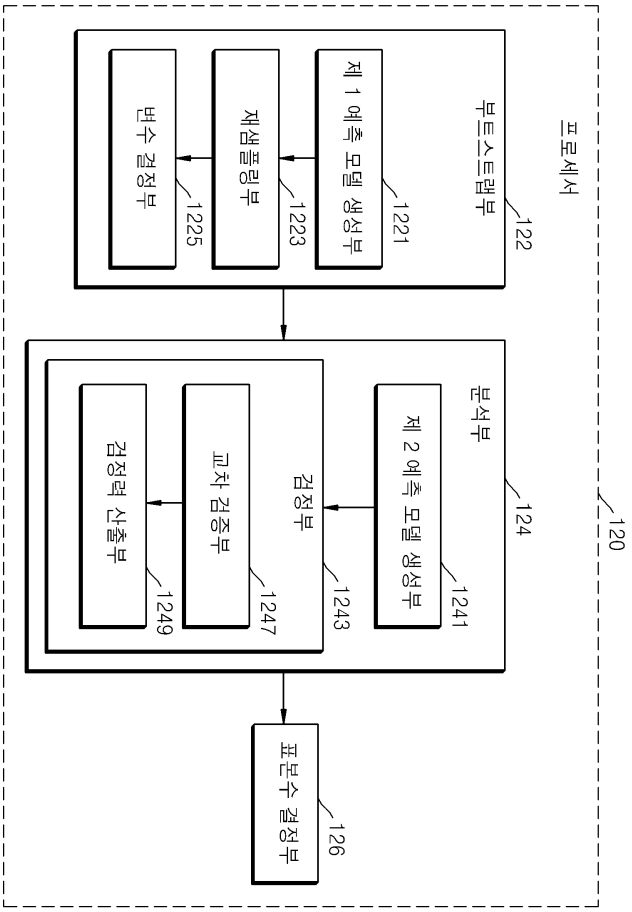
도면1



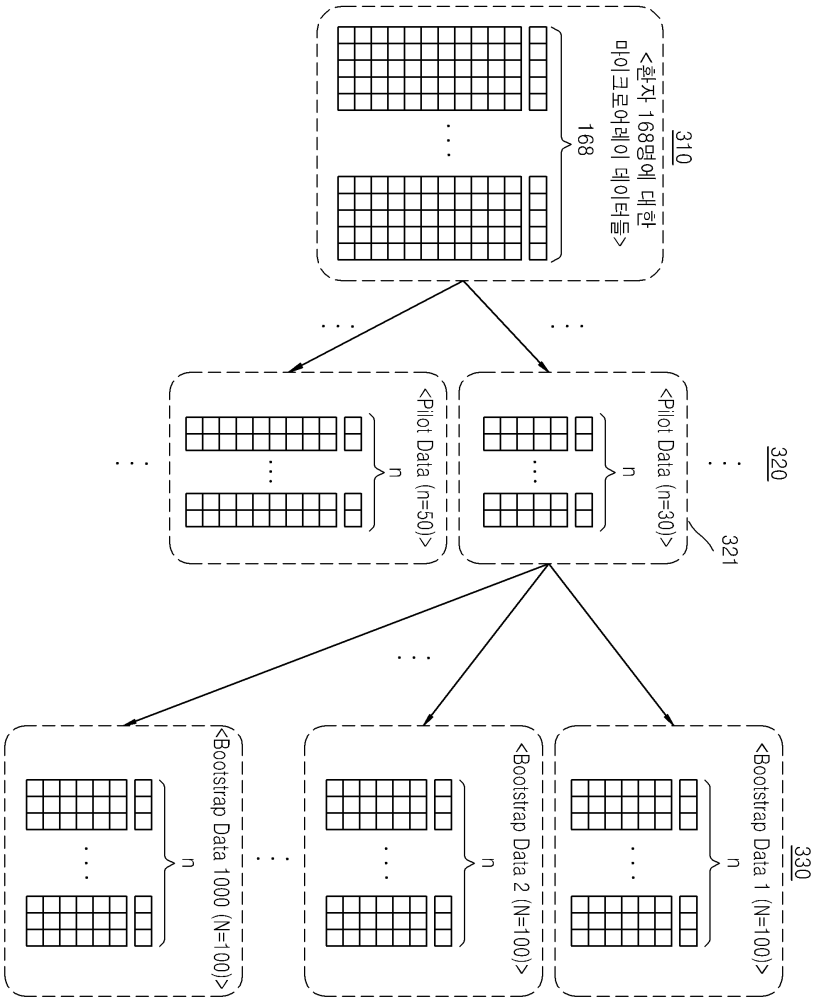
도면2a


10
컴퓨팅 장치
120
프로세서
110
"마이크로어레이 데이터들"
데이터 획득부
122
부트스트랩부
124
분석부
126
표본수 결정부
"최적 표본수 N"

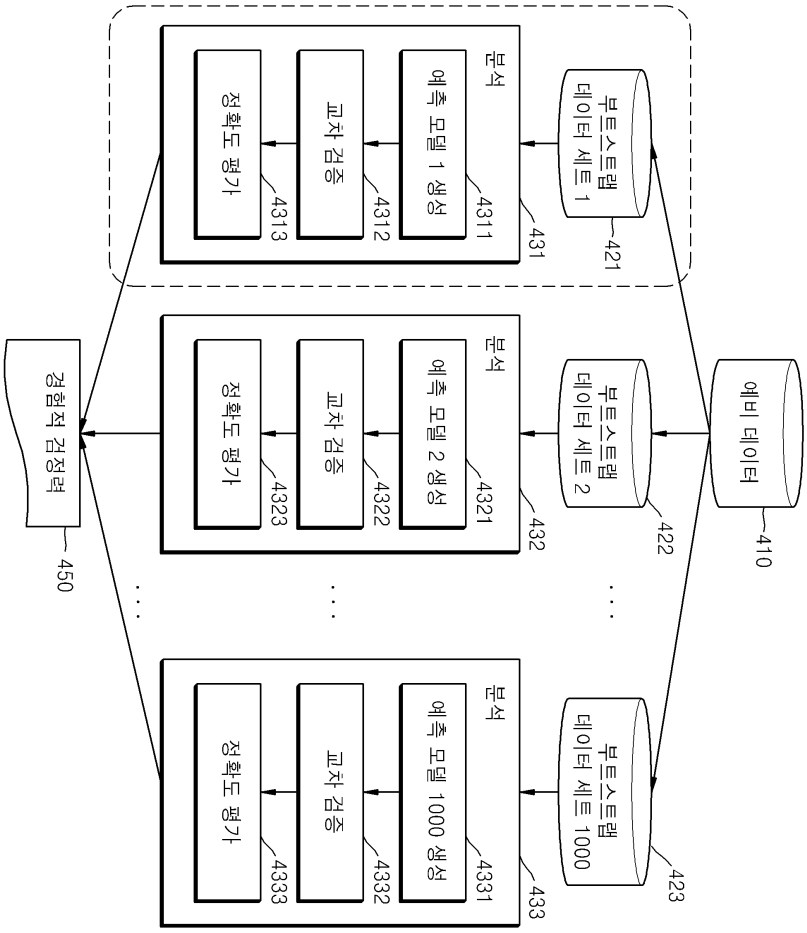
도면2b


프로세서
부트스트랩부
122
제 1 예측 모델 생성부
1221
재샘플링부
1223
변수 결정부
1225
분석부
124
제 2 예측 모델 생성부
1241
검정부
1243
교차 검증부
1247
검정력 산출부
1249
표본수 결정부
126
120

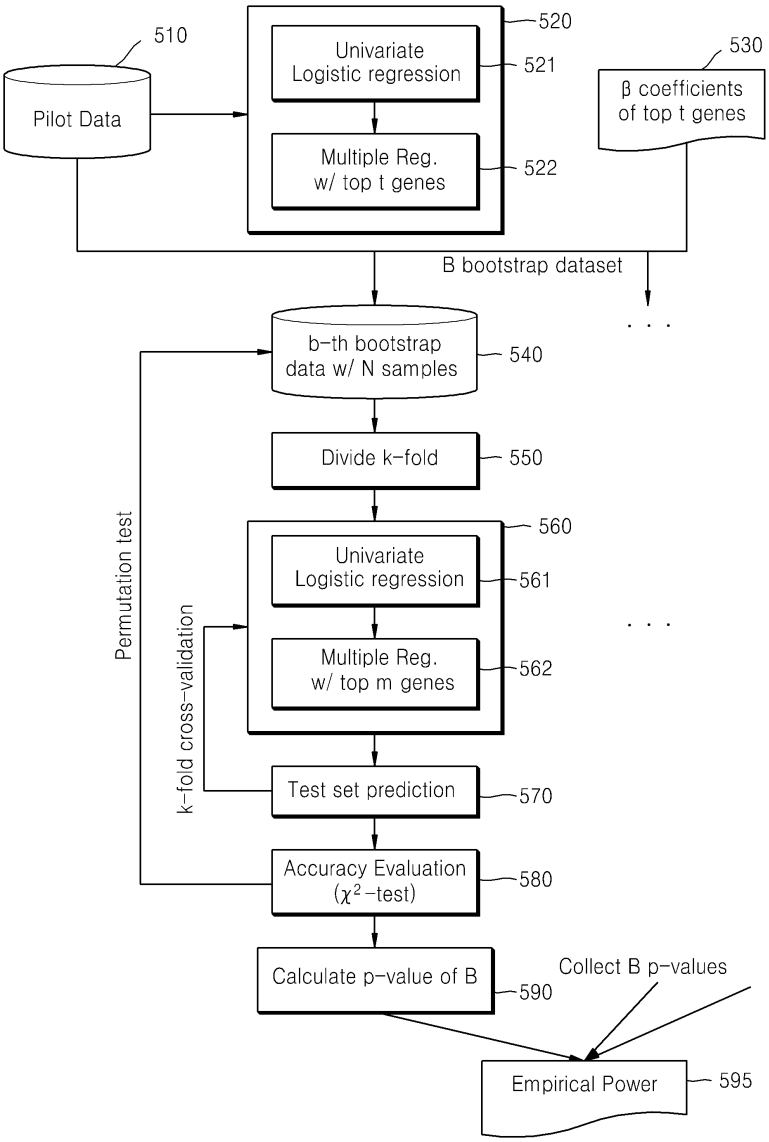
도면3


<환자 168명에 대한 마이크로어레이 데이터들>
168
310
<Pilot Data (n=50)>
<Pilot Data (n=30)>
n
320
321
<Bootstrap Data 1000 (N=100)>
<Bootstrap Data 2 (N=100)>
<Bootstrap Data 1 (N=100)>
330

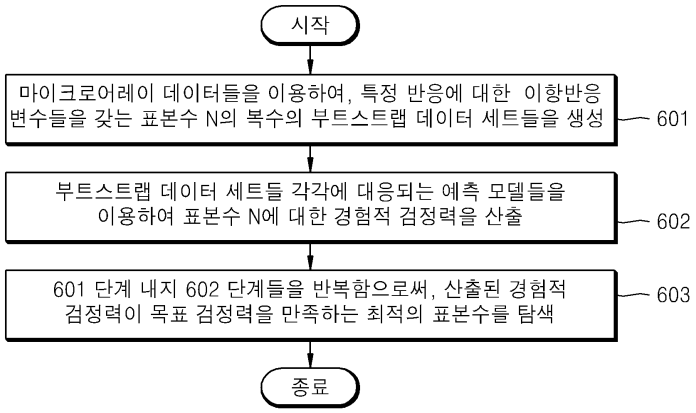
도면4


분석
예측 모델 1 생성
4311
교차 검증
4312
정확도 평가
4313
431
부트스트랩 데이터 세트 1
421
분석
예측 모델 2 생성
4321
교차 검증
4322
정확도 평가
4323
432
부트스트랩 데이터 세트 2
422
분석
예측 모델 1000 생성
4331
교차 검증
4332
정확도 평가
4333
433
부트스트랩 데이터 세트 1000
423
예비 데이터
410
경험적 검정력
450

도면5


510
Pilot Data
520
Univariate
Logistic regression
521
Multiple Reg.
w/ top t genes
522
530
β coefficients
of top t genes
B bootstrap dataset
b-th bootstrap
data w/ N samples
540
Divide k-fold
550
560
Univariate
Logistic regression
561
Multiple Reg.
w/ top m genes
562
Permutation test
k-fold cross-validation
Test set prediction
570
Accuracy Evaluation
(χ^2-test)
580
Calculate p-value of B
590
Collect B p-values
Empirical Power
595

도면6


시작
마이크로어레이 데이터들을 이용하여, 특정 반응에 대한 이항반응 변수들을 갖는 표본수 N의 복수의 부트스트랩 데이터 세트들을 생성
601
부트스트랩 데이터 세트들 각각에 대응되는 예측 모델들을 이용하여 표본수 N에 대한 경험적 검정력을 산출
602
601 단계 내지 602 단계들을 반복함으로써, 산출된 경험적 검정력이 목표 검정력을 만족하는 최적의 표본수를 탐색
603
종료