(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0102166
(43) 공개일자 2022년07월20일

(51) 국제특허분류(Int. Cl.)
G16H 50/70 (2018.01) *G16H 70/00* (2018.01)
(52) CPC특허분류
G16H 50/70 (2018.01)
G16H 70/00 (2021.08)
(21) 출원번호 10-2021-0003233
(22) 출원일자 2021년01월11일
심사청구일자 2021년01월11일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
박유랑
서울특별시 송파구 올림픽로 435, 310동 103호 (신천동, 파크리오)
박지애
서울특별시 노원구 동일로245번길 162, 101동 1301호 (상계동, 은빛1단지아파트)
(74) 대리인
특허법인비엘티

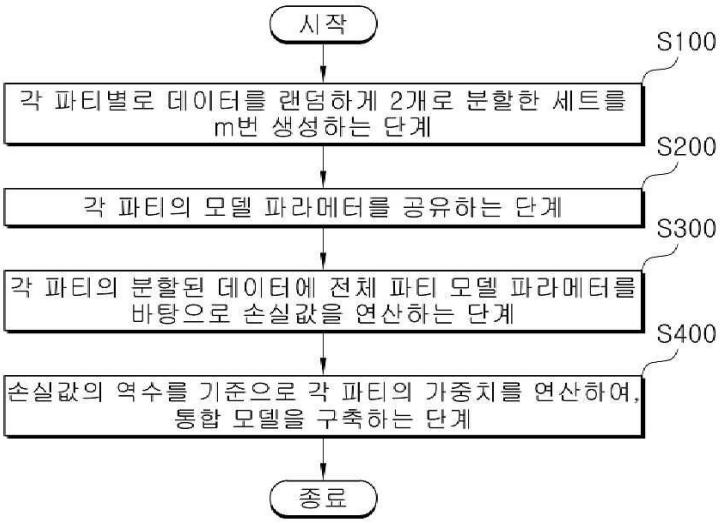
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법**

(57) 요 약

가중치 기반 통합 방법은, 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계, 각 파티의 모델 파라메터를 공유하는 단계, 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계 및 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계를 포함한다.

대 표 도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호	1465030035
과제번호	HI19C1015010020
부처명	보건복지부
과제관리(전문)기관명	한국보건산업진흥원
연구사업명	의료데이터 보호·활용 기술개발
연구과제명	다중 분할 임상 데이터 기반 분산형 컴퓨팅 기술 개발 및 검증
기 여 율	1/2
과제수행기관명	연세대학교 산학협력단
연구기간	2020.01.01 ~ 2020.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711096662
과제번호	2019M3E5D4064682
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	원천기술개발사업
연구과제명	공통자료모델 기반 임상 오믹스 정보 통합 개방형 플랫폼 구축 및 다기관 검증
기 여 율	1/2
과제수행기관명	연세대학교
연구기간	2020.03.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

서버에 의해 수행되는, 가중치 기반 통합 방법에 있어서,

각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계;

각 파티의 모델 파라메터를 공유하는 단계;

각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계; 및

손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계;를 포함하는 가중치 기반 통합 방법.

청구항 2

제1 항에 있어서,

각 파티별로 로지스틱 모델을 이용하여 파라메터 데이터를 생성하는 단계를 더 포함하는 가중치 기반 통합 방법.

청구항 3

제2 항에 있어서,

상기 로지스틱 모델은 $\ln\left(\frac{p}{1-p}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 에 따르는 것인 가중치 기반 통합 방법.

청구항 4

제1 항에 있어서,

상기 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계는,

$Z_{P_k,i}^{(1)}, Z_{P_k,i}^{(2)}$ $\leq i \leq m$ P_k

( )를 생성하는 단계로, 여기에서, 1 의 자연수이고, 는 k번째 파티를 지칭하는 변

$Z^{(1)}$ $\frac{n_k x}{x+1}$ $Z^{(2)}$ $\frac{n_k}{x+1}$

수이고, 은 의 크기를 가지는 첫번째 분할 부분이고, 는 의 크기를 가지는 두번째 분할 부분이며, x는 임의의 숫자인 가중치 기반 통합 방법.

청구항 5

제1 항에 있어서,

각 파티의 모델 파라메터를 공유하는 단계는,

각 파티별로 분할된 세트에서 추정한 파라메터 벡터값을 서로 다른 분할된 세트로 보내어 공유하는 단계인 가중치 기반 통합 방법.

청구항 6

제5 항에 있어서,

각 파티의 모델 파라메터를 공유하는 단계는,

$Z_{P_k,i}^{(1)}$ $\hat{f}_{P_k,i}$ $\hat{f}_{P_k,1}, \hat{f}_{P_k,2}, \cdots, \hat{f}_{P_k,m}$

를 이용하여 를 추정하여 ( )를 도출하는 단계,

$\hat{f}_{P_k,i}$ $\hat{\beta}_{P_k,i}$ $(\hat{\beta}_{P_k,1}, \hat{\beta}_{P_k,2}, \cdots, \hat{\beta}_{P_k,m})$

를 이용하여 파라메터 벡터 를 추정하여 를 도출하는 단계, 및

$(\hat{\beta}_{P_k,1}, \hat{\beta}_{P_k,2}, \cdots, \hat{\beta}_{P_k,m})$ $\leq i \leq m$

각 파티별로 도출한 를 서로 공유하는 단계를 포함하고, 여기에서, 1 의 자연수

P_k $\hat{f}_{P_k,i}$

이고, 는 k번째 파티를 지칭하는 변수이고, 는 k번째 파티에 대한, i번째의 모델을 나타내는 파라메터인 가중치 기반 통합 방법.

청구항 7

제1 항에 있어서,

상기 손실값을 연산하는 단계는,

각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계,

제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 전달하는 단계, 및

제2 분할 세트의 손실값을 각 파티별로 연산하는 단계를 포함하는 가중치 기반 통합 방법.

청구항 8

제7 항에 있어서,

상기 각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계는, 파티 별로 피팅된 $(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \cdots, \hat{f}_{P_k,i}, \cdots, \hat{f}_{P_K,i})$

를 도출하는 단계이고,

$(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \cdots, \hat{f}_{P_k,i}, \cdots, \hat{f}_{P_K,i})$

제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 전달하는 단계는, 를 i

$Z_{P_k,i}^{(2)}(n = \frac{n_k}{x+1})$

에 대응하는 제2 분할 세트인 로 전달하는 단계이고,

$Loss_{P_k,i}(Z_{P_{k'},i}^{(2)})$

제2 분할 세트의 손실값을 각 파티별로 연산하는 단계는 로 표현되는 가중치 기반 통합 방법.

청구항 9

제8 항에 있어서,

손실값 연산 함수는 로지스틱 회귀 함수인 가중치 기반 통합 방법.

청구항 10

제9 항에 있어서,

$-\sum_{i=1}^{N}\{y_i ln p_i + (1-y_i) ln(1-p_i)\}$

손실값 연산 함수는 -ln $L(p)$ = 로 표현되고, 여기서

$p_i = 1/(1 + \exp(-\beta^T x_i))$

인 가중치 기반 통합 방법.

청구항 11

제1 항에 있어서,

$$W_{P_k,i} = \frac{E_{P_k,i}}{\sum_{k=1}^{K} E_{P_k,i}}, (i = 1,2,\dots,m)$$

상기 손실값의 역수를 기준으로 각 파티의 가중치를 연산은 를 통해 연산되고, 여기서 $E_{P_k,i}$ 는 $Loss_{P_k,i}$ 의 역수로 정의되며,

상기 통합 모델을 구축하는 단계는, $\hat{f}_{IM} = \hat{f}_{P_1} \times \hat{W}_{P_1} + \hat{f}_{P_2} \times \hat{W}_{P_2} + \cdots + \hat{f}_{P_K} \times \hat{W}_{P_K}$ 를 통해 연산되는 것인 가중치 기반 통합 방법.

청구항 12

제1 항에 있어서,

상기 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계는,

각 파티별로 이벤트 타임 데이터를 함께 생성하는 단계인 가중치 기반 통합 방법.

청구항 13

제12 항에 있어서,

상기 각 파티별로 이벤트 타임 데이터의 생성은 $\{t_{j1}, t_{j2}, \dots, t_{jn_{dj}}\}$ 로 표현되고, n_j 는 특정 사이트에서의 이벤트의 개수를 나타내는 가중치 기반 통합 방법.

청구항 14

제12 항에 있어서,

각 파티의 모델 파라메터를 공유하는 단계는,

각 파티별로 분할된 세트에서 추정한 파라메터 벡터값 및 이벤트 타임 데이터값을 서로 다른 분할된 세트로 보내어 공유하는 단계인 가중치 기반 통합 방법.

청구항 15

제12 항에 있어서,

상기 손실값을 연산하는 단계는, 콕스 모델(Cox model)의 손실값 함수를 이용하여 연산하는 것인 가중치 기반 통합 방법.

청구항 16

제12 항에 있어서,

상기 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계 이후에,

각 파티의 이벤트 타임 데이터에 대해서 서바이벌 함수를 연산하는 단계를 더 포함하는 가중치 기반 통합 방법.

청구항 17

제16 항에 있어서,

상기 서바이벌 함수를 연산하는 단계 이후에 각 파티에 대해서 서바이벌 함수값을 더해서, 중앙 타임 포인트에서의 중앙 서바이벌값을 추정하는 단계를 더 포함하는 가중치 기반 통합 방법.

청구항 18

제17 항에 있어서,

$\{t_1, t_2, \ldots, t_{N_d}\}$에서 각 파티별 $\Sigma_{j \in R(t)} \exp(x'_i \hat{\beta}_{IM})$ 를 더해서 중앙 서바이벌값을 추정하는 것인 가중치 기반 통합 방법.

청구항 19

제17 항에 있어서,

상기 중앙 서바이벌값을 추정하는 단계는, 이벤트 타임 데이터마다 복수 개의 통합 모델 파라메터를 바탕으로 중앙 서바이벌 값을 추정하고, 추정된 중앙 서바이벌 값의 추정치를 각 타임 포인트에서의 포인트 서바이벌 값을 추정하는 것인 가중치 기반 통합 방법.

청구항 20

컴퓨터인 하드웨어와 결합되어, 제1항 내지 제19항 중 어느 한 항의 방법을 실행하기 위해 매체에 저장된, 가중치 기반 통합 프로그램.

발명의 설명

기 술 분 야

[0001] 본 발명은 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법에 관한 것이다.

배 경 기 술

[0002] 연구 모집단의 대표성을 확보하는 것은 연구의 일반화 가능성을 높일 수 있기 때문에 바이오메디컬 연구에 있어서 중요하다. 이 점에서 다중 기관의 의료 데이터를 사용하는 것은 연구에 이점이 있다. 그러나 의료 데이터의 비밀유지의무 및 기밀 특성으로 인해 개인 정보 문제가 발생하기 때문에 의료 데이터를 물리적으로 결합하기는 곤란하다. 따라서 여러 기관의 의료 데이터를 연구에 활용하기 위해서는 기관 간 실체적 데이터 공유없이 모델을 구축 할 수 있는 방법의 개발이 요구된다.

선행기술문헌

특허문헌

[0003] (특허문헌 0001) 등록특허공보 제 10-1799823 호, 2017.11.15

발명의 내용

해결하려는 과제

[0004] 본 발명이 해결하고자 하는 과제는 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법을 제공하는 것이다.

[0005] 본 발명이 해결하고자 하는 과제들은 이상에서 언급된 과제로 제한되지 않으며, 언급되지 않은 또 다른 과제들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0006] 상술한 과제를 해결하기 위한 본 발명의 일 면에 따른 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법은, 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계, 각 파티의 모델 파라메터를 공유하는 단계, 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계, 및 손실값의 역수를 기준으로 각 파티의 가중치를 연

산하여, 통합 모델을 구축하는 단계;를 포함한다.

[0007] 각 파티별로 로지스틱 모델을 이용하여 파라메터 데이터를 생성하는 단계를 더 포함할 수 있다.

[0008] 상기 로지스틱 모델은 $\ln\left(\frac{p}{1-p}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 에 따르는 것일 수 있다.

[0009] 상기 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계는, $Z_{P_k,i}^{(1)}, Z_{P_k,i}^{(2)}$ $\leq i \leq m$ P_k ( )를 생성하는 단계로, 여기에서, 1 의 자연수이고, 는 k번째 파티를 지칭하는 변수이고, $Z^{(1)}$ $\frac{n_k x}{x+1}$ 은 의 크기를 가지는 첫번째 분할 부분이고, $Z^{(2)}$ $\frac{n_k}{x+1}$ 는 의 크기를 가지는 두번째 분할 부분이며, x는 임의의 숫자일 수 있다.

[0010] 각 파티의 모델 파라메터를 공유하는 단계는, 각 파티별로 분할된 세트에서 추정한 파라메터 벡터값을 서로 다른 분할된 세트로 보내어 공유하는 단계일 수 있다.

[0011] 각 파티의 모델 파라메터를 공유하는 단계는, $Z_{P_k,i}^{(1)}$ 를 이용하여 $\hat{f}_{P_k,i}$ 를 추정하여 ($\hat{f}_{P_k,1}, \hat{f}_{P_k,2}, \cdots, \hat{f}_{P_k,m}$)를 도출하는 단계, $\hat{f}_{P_k,i}$ 를 이용하여 파라메터 벡터 $\hat{\beta}_{P_k,i}$ 를 추정하여 $(\hat{\beta}_{P_k,1}, \hat{\beta}_{P_k,2}, \cdots, \hat{\beta}_{P_k,m})$ 를 도출하는 단계, 및 각 파티별로 도출한 $(\hat{\beta}_{P_k,1}, \hat{\beta}_{P_k,2}, \cdots, \hat{\beta}_{P_k,m})$ $\leq i \leq m$ 를 서로 공유하는 단계를 포함하고, 여기에서, 1 의 자연수이고, P_k 는 k번째 파티를 지칭하는 변수이고, $\hat{f}_{P_k,i}$ 는 k번째 파티에 대한, i번째의 모델을 나타내는 파라메터 일 수 있다.

[0012] 상기 손실값을 연산하는 단계는, 각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계, 제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 전달하는 단계, 및 제2 분할 세트의 손실값을 각 파티별로 연산하는 단계를 포함할 수 있다.

[0013] 상기 각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계는, 파티 별로 피팅된 $(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \cdots, \hat{f}_{P_k,i}, \cdots, \hat{f}_{P_K,i})$ 를 도출하는 단계이고, 제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 $(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \cdots, \hat{f}_{P_k,i}, \cdots, \hat{f}_{P_K,i})$ $Z_{P_k,i}^{(2)}\left(n = \frac{n_k}{x+1}\right)$ 전달하는 단계는, 를 i에 대응하는 제2 분할 세트인 로 전달하는 단계이고, 제2 분할 세트의 손실값을 각 파티별로 연산하는 단계는 $Loss_{P_k,i}(Z_{P_{k'},i}^{(2)})$ 로 표현될 수 있다.

[0014] 손실값 연산 함수는 로지스틱 회귀 함수일 수 있다.

[0015] 손실값 연산 함수는 $-\ln L(p) = -\sum_{i=1}^{N}\{y_i \ln p_i + (1-y_i)\ln(1-p_i)\}$ 로 표현되고, 여기서 $p_i = 1/(1+\exp(-\beta^T x_i))$ 일 수 있다.

[0016] 상기 손실값의 역수를 기준으로 각 파티의 가중치를 연산은 $W_{P_k,i} = \frac{E_{P_k,i}}{\sum_{k=1}^{K} E_{P_k,i}}, (i = 1,2,\ldots,m)$ 를 통해 연산되

$E_{P_k,i}$ $Loss_{P_k,i}$

고, 여기서 는 의 역수로 정의되며, 상기 통합 모델을 구축하는 단계는,

$$\hat{f}_{IM} = \hat{f}_{P_1} \times \hat{W}_{P_1} + \hat{f}_{P_2} \times \hat{W}_{P_2} + \cdots + \hat{f}_{P_K} \times \hat{W}_{P_K}$$

를 통해 연산되는 것일 수 있다.

[0017] 상기 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계는, 각 파티별로 이벤트 타임 데이터를 함께 생성하는 단계일 수 있다.

$\{t_{j1}, t_{j2}, \ldots, t_{jn_{dj}}\}$ n_j

[0018] 상기 각 파티별로 이벤트 타임 데이터의 생성은 로 표현되고, 는 특정 사이트에서의 이벤트의 개수를 나타낼 수 있다.

[0019] 각 파티의 모델 파라메터를 공유하는 단계는, 각 파티별로 분할된 세트에서 추정한 파라메터 벡터값 및 이벤트 타임 데이터값을 서로 다른 분할된 세트로 보내어 공유하는 단계일 수 있다.

[0020] 상기 손실값을 연산하는 단계는, 콕스 모델(Cox model)의 손실값 함수를 이용하여 연산하는 것일 수 있다.

[0021] 상기 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계 이후에, 각 파티의 이벤트 타임 데이터에 대해서 서바이벌 함수를 연산하는 단계를 더 포함할 수 있다.

[0022] 상기 서바이벌 함수를 연산하는 단계 이후에 각 파티에 대해서 서바이벌 함수값을 더해서, 중앙 타임 포인트에서의 중앙 서바이벌값을 추정하는 단계를 더 포함할 수 있다.

[0023] $\{t_1, t_2, \ldots, t_{N_d}\}$에서 각 파티별 $\sum_{j \in R(t)} \exp(x'_i \hat{\beta}_{IM})$ 를 더해서 중앙 서바이벌값을 추정하는 것일 수 있다.

[0024] 상기 중앙 서바이벌값을 추정하는 단계는, 이벤트 타임 데이터마다 복수 개의 통합 모델 파라메터를 바탕으로 중앙 서바이벌 값을 추정하고, 추정된 중앙 서바이벌 값의 추정치를 각 타임 포인트에서의 포인트 서바이벌 값을 추정하는 것일 수 있다.

[0025] 또한, 본 발명은 컴퓨터인 하드웨어와 결합되어, 전술한 방법을 실행하기 위해 매체에 저장된, 가중치 기반 통합 프로그램을 제공할 수 있다.

[0026] 본 발명의 기타 구체적인 사항들은 상세한 설명 및 도면들에 포함되어 있다.

발명의 효과

[0027] 본 발명의 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법은, 환자 수준의 물리적 데이터를 공유하지 않고 개인정보 보호 하에서 예측 모델의 일반화를 구현하여 평균 예측 성능을 향상시킬 수 있다.

[0028] 본 발명의 일 면에 따른 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법은, 기관 간의 반복적인 커뮤니케이션 없이도 다중 기관 데이터의 가중치 기반 통합 예측 모델을 구축할 수 있다.

[0029] 본 발명의 일 면에 따른 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법은, 로지스틱스 회귀 모델을 적용하여 모든 데이터가 결합된 중앙 집중식 모델 수준의 타당성을 구현할 수 있다.

[0030] 본 발명의 일 면에 따른 물리적 데이터 공유 없이 수평분할 기반 중앙화 모델을 추정하기 위한 가중치 기반 통합 방법은, 서바이벌 함수를 이용하여 각 시간 포인트에서의 포인트 서바이벌 비율을 추정할 수 있다.

[0031] 본 발명의 효과들은 이상에서 언급된 효과로 제한되지 않으며, 언급되지 않은 또 다른 효과들은 아래의 기재로부터 통상의 기술자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0032] 도 1은 본 발명의 일면에 따른 가중치 기반 통합 방법을 통해 가중치 통합 모델을 도출하는 단계를 도시한 순서도이다.

도 2는 도 1에 따른 순서도를 구체화한 도면이다.

도 3은 본 발명의 일면에 따른 가중치 기반 통합 방법을 통해 중앙 서바이벌값을 추정하는 단계를 도시한 순서도이다.

도 4는 도 3에 따른 순서도를 구체화한 도면이다.

도 5는 반복회수에 따른 가중치값을 나타낸 그래프이다.

도 6은 동일한 데이터 특성 하에서, 데이터 사이즈에 따른 가중치 패턴의 변화를 도시한 그래프이다.

도 7은 동일한 데이터 특성 하에서, 중앙 데이터와의 부합정도에 따른 가중치 패턴의 변화를 도시한 그래프이다.

도 8은 각 기관별로 10개의 모델에 대한 가중치값을 예시적으로 나타낸 도면이다.

도 9는 로지스틱 회귀 모델에 대한 예측 정도를 비교도시한 그래프이다.

도 10은 10개의 모델, 가중치 통합 모델, 중앙화 모델의 외부 검증값을 비교도시한 그래프이다.

도 11은 로지스틱 회귀 모델을 적용한 모델에 대한 OR수치를 비교도시한 그래프이다.

발명을 실시하기 위한 구체적인 내용

[0033] 본 발명의 이점 및 특징, 그리고 그것들을 달성하는 방법은 첨부되는 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나, 본 발명은 이하에서 개시되는 실시예들에 제한되는 것이 아니라 서로 다른 다양한 형태로 구현될 수 있으며, 단지 본 실시예들은 본 발명의 개시가 완전하도록 하고, 본 발명이 속하는 기술 분야의 통상의 기술자에게 본 발명의 범주를 완전하게 알려주기 위해 제공되는 것이며, 본 발명은 청구항의 범주에 의해 정의될 뿐이다.

[0034] 본 명세서에서 사용된 용어는 실시예들을 설명하기 위한 것이며 본 발명을 제한하고자 하는 것은 아니다. 본 명세서에서, 단수형은 문구에서 특별히 언급하지 않는 한 복수형도 포함한다. 명세서에서 사용되는 "포함한다(comprises)" 및/또는 "포함하는(comprising)"은 언급된 구성요소 외에 하나 이상의 다른 구성요소의 존재 또는 추가를 배제하지 않는다. 명세서 전체에 걸쳐 동일한 도면 부호는 동일한 구성 요소를 지칭하며, "및/또는"은 언급된 구성요소들의 각각 및 하나 이상의 모든 조합을 포함한다. 비록 "제1", "제2" 등이 다양한 구성요소들을 서술하기 위해서 사용되나, 이들 구성요소들은 이들 용어에 의해 제한되지 않음은 물론이다. 이들 용어들은 단지 하나의 구성요소를 다른 구성요소와 구별하기 위하여 사용하는 것이다. 따라서, 이하에서 언급되는 제1 구성요소는 본 발명의 기술적 사상 내에서 제2 구성요소일 수도 있음은 물론이다.

[0035] 다른 정의가 없다면, 본 명세서에서 사용되는 모든 용어(기술 및 과학적 용어를 포함)는 본 발명이 속하는 기술분야의 통상의 기술자에게 공통적으로 이해될 수 있는 의미로 사용될 수 있을 것이다. 또한, 일반적으로 사용되는 사전에 정의되어 있는 용어들은 명백하게 특별히 정의되어 있지 않는 한 이상적으로 또는 과도하게 해석되지 않는다.

[0036] 이하의 본 발명에 따른 가중치 기반 통합 방법은 서버와 같은 컴퓨터 장치를 통해 수행될 수 있다.

[0037] 이하, 첨부된 도면을 참조하여 본 발명의 실시예를 상세하게 설명한다.

[0038] 도 1은 본 발명의 일면에 따른 가중치 기반 통합 방법을 통해 가중치 통합 모델을 도출하는 단계를 도시한 순서도이다. 도 2는 도 1에 따른 순서도를 구체화한 도면이다.

[0039] 도 1 및 2를 참조하면, 가중치 기반 통합 방법은 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계(S100), 각 파티의 모델 파라메터를 공유하는 단계(S200), 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계(S300), 및 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계(S400)를 포함한다.

[0040] 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계(S100)는 예측 모델을 추정하고 성능을 측정하기 위해 각 파티가 두 개의 분할 데이터 세트를 생성한다.

$$n_k$$

[0041] 구체적으로는, 본 단계(S100)는 k번째 파티(크기를 n_k라 할때)를 2개의 데이터 세트로 분할한다.

[0042] 첫번째 분할 세트는 $Z^{(1)}$ 로 지칭되고, 크기를 $\frac{n_k x}{x+1}$ 를 가진다.

[0043] 두번째 분할 세트는 $Z^{(2)}$ 로 지칭되고, 크기를 $\frac{n_k}{x+1}$ 를 가진다.

[0044] $Z^{(1)}$ f 은 예측 모델 를 추정하기 위해 이용될 수 있다.

[0045] $Z^{(2)}$ $Z^{(1)}$ $\hat{f}$ 은 으로부터 얻어진 예측 모델 를 추정하기 위해 이용될 수 있다.

[0046] $Z^{(1)}, Z^{(2)}$ P_k 데이터 세트 ( )는 각 파티 (여기에서 k는 파티의 번호를 지칭하는 숫자를 나타낸다)에 대해서 m번 반복해서 생성(여기서 m은 2이상의 자연수)될 수 있다.

[0047] $\leq i \leq m$ m번 반복해서 생성된 데이터 세트를 지칭하기 위해 변수 를 정의할 때, 해당 범주를 만족한다(1).

[0048] $Z_{P_k,i}^{(1)}, Z_{P_k,i}^{(2)}$ ( ) 는 k번째 파티의 i번째 데이터 세트의 데이터를 나타낸다.

[0049] 각 파티의 모델 파라메터를 공유하는 단계(S200)는 각 파티가 추정한 모델 파라메터가 서로 공유된다.

[0050] P_k $\hat{f}_{P_k,i}$ $Z_{P_k,i}^{(1)}$ 구체적으로는, K번째 파티인 의 번째 모델인 가 첫번째 분할 세트 데이터인 를 통해 추정된다. 그리고, 벡터 파라메터인 $\hat{\beta}_{P_k,i}$ $\hat{f}_{P_k,i}$ 는 로부터 추정된다.

[0051] $(\hat{\beta}_{P_k,1}, \hat{\beta}_{P_k,2}, \ldots, \hat{\beta}_{P_k,m})$ $\hat{f}_{P_k,1}, \hat{f}_{P_k,2}, \ldots, \hat{f}_{P_k,m}$ 결과적으로 k개의 파티의 벡터 파라메터 세트 는 ( )를 통해 추정된다.

[0052] 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계(S300)는 모델을 각 파티의 전체 데이터 세트에 피팅하여 각 파티의 모델에 대한 손실값을 계산하는 단계이다.

[0053] 구체적으로는, 손실값을 연산하는 단계는, 각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계, 제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 전달하는 단계, 및 제2 분할 세트의 손실값을 각 파티별로 연산하는 단계를 포함할 수 있다.

[0054] 상기 각 파티별로 제1 분할 세트를 기준으로 도출된 모델을 피팅하는 단계는, 파티 별로 피팅된 $(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \ldots, \hat{f}_{P_k,i}, \ldots, \hat{f}_{P_K,i})$ 를 도출하는 단계이고, 제1 분할 세트를 기준으로 피팅한 모델을 제2 분할 세트로 $(\hat{f}_{P_1,i}, \hat{f}_{P_2,i}, \ldots, \hat{f}_{P_k,i}, \ldots, \hat{f}_{P_K,i})$ $Z_{P_k,i}^{(2)}(n = \frac{n_k}{x+1})$ 전달하는 단계는, 를 i에 대응하는 제2 분할 세트인 로 전달하는 $Loss_{P_k,i}(Z_{P_{k'},i}^{(2)})$ 단계이고, 제2 분할 세트의 손실값을 각 파티별로 연산하는 단계는 로 표현될 수 있다.

[0055] 손실값을 연산하는 것은 모델에 따라 다양한 함수가 적용될 수 있다.

[0056] 예를 들어, 이진법 분류 모델에 있어서는 로지스틱 모델을 이용하여 손실값을 연산할 수 있다.

$$\ln\left(\frac{p}{1-p}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

[0057] 예를 들어, 로지스틱 모델은 에 따르는 것일 수 있다.

$$-\sum_{i=1}^{N}\{y_i lnp_i + (1-y_i)ln(1-p_i)\}$$

[0058] 예를 들어, 로지스텍 모델을 따를 손실값 연산 함수는 -ln L(p) = 로 표

$p_i = 1/(1+\exp(-\beta^T x_i))$ β^T x_i

현되고, 여기서 일 수 있고, 는 파라메터의 벡터이며, 는 i번째 환자의 특성을

y_i

나타내는 벡터이며, 는 i번째 환자의 바이너리 출력값일 수 있다.

$Loss_{P_{k'},i}$

[0059] 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계(S400)는 의 역수인

$E_{P_{k'},i}$ $W_{P_{k'},i}$

를 정의하여 가중치 를 정의할 수 있고, 통합 모델을 가중치를 통해 연산할 수 있다. 손실값의 역수를

$$W_{P_{k'},i} = \frac{E_{P_{k'},i}}{\sum_{k=1}^{K} E_{P_{k'},i}}, (i = 1,2,\ldots,m)$$

$E_{P_{k'},i}$

기준으로 각 파티의 가중치를 연산은 를 통해 연산되고, 여기서 는

$Loss_{P_{k'},i}$

의 역수로 정의되며, 상기 통합 모델을 구축하는 단계는,

$\hat{f}_{IM} = \hat{f}_{P_1} \times \widehat{W}_{P_1} + \hat{f}_{P_2} \times \widehat{W}_{P_2} + \cdots + \hat{f}_{P_K} \times \widehat{W}_{P_K}$

를 통해 연산되는 것일 수 있다.

[0060] 이렇게 연산된 통합 모델의 가중치는 2가지 요소에 의해 결정될 수 있다. 첫번째는 파티의 데이터 크기로 중앙 데이터에 대한 데이터 사이즈의 비율에 해당한다. 두번째는 다른 파티에 대해서 파티 모델이 얼마나 잘 부합하는지 정도에 대한 것으로, 각 파티별로 형성한 모델이 다른 파티들에 대해서 얼마나 부합(fitting)되는지에 해

$\sum_{k'=1}^{K'} Z_{P_{k'},i}$ (2)

당한다. 만일 k번째 파티의 데이터 크기가 매우 큰 경우, 전체 파티에 대한 데이터 크기 에 대한

$Loss_{P_{k'},i}(i = 1,2,\ldots,m)$

데이터의 비율이 커지는 것이며, k번째 파티에 대한 값 즉 손실값은 작아진다. 이렇게

$W_{P_{k'},i}(i = 1,2,\ldots,m)$

될 경우, 가중치 값 은 다른 파티 대비 커질 수 있다. 이의 의미는, 큰 데이터 세트를 가지는 파티의 가중치는 작은 데이터 세트를 가지는 파티의 가중치 대비 더 크다는 것을 의미한다.

[0061] 도 3은 본 발명의 일면에 따른 가중치 기반 통합 방법을 통해 중앙 서바이벌값을 추정하는 단계를 도시한 순서도이다. 도 4는 도 3에 따른 순서도를 구체화한 도면이다.

[0062] 도 3 및 4를 참조하면, 가중치 기반 통합 방법은 각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계(S110), 각 파티의 모델 파라메터를 공유하는 단계(S210), 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계(S310), 및 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계(S410), 각 파티의 이벤트 타임 데이터에 대해서 서바이벌 함수를 연산하는 단계(S500), 각 파티에 대해서 서바이벌 함수값을 더해서, 중앙 타임 포인트에서의 중앙 서바이벌값을 추정하는 단계(S600)를 포함한다.

[0063] 2개로 분할한 세트를 m번(여기서 m은 2이상의 자연수) 생성하는 단계(S110)는, 각 파티별로 이벤트 타임 데이터

$\{t_{j1}, t_{j2}, \ldots, t_{jn_{dj}}\}$ n_j

를 함께 생성하는 단계이다. 각 파티별로 이벤트 타임 데이터의 생성은 로 표현되고, 는

특정 사이트에서의 이벤트의 개수를 나타낼 수 있다.

[0064] 각 파티의 모델 파라메터를 공유하는 단계(S210)는, 각 파티별로 분할된 세트에서 추정한 파라메터 벡터값 및 이벤트 타임 데이터값을 서로 다른 분할된 세트로 보내어 공유하는 단계일 수 있다.

[0065] 예를 들어, 파티 A와 파티 B를 가정할 때, A의 m개 $Z^{(1)}$ 에서 추정한 m개의 $\hat{\beta}_{A1}, \hat{\beta}_{A2}, \ldots, \hat{\beta}_{Ap}$ ( )와 $\{t_{A1}, t_{A2}, \ldots, t_{An_{dA}}\}$ 를 B에 보내고, B의 m개 $Z^{(1)}$ 에서 추정한 m개의 $\hat{\beta}_{B1}, \hat{\beta}_{B2}, \ldots, \hat{\beta}_{Bp}$ ( )와 $\{t_{B1}, t_{B2}, \ldots, t_{Bn_{dB}}\}$ 를 A로 보낼 수 있다.

[0066] 각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계(S310)는 콕스 모델(Cox model)의 손실값 함수를 이용하여 연산하는 것일 수 있다. 예를 들어, 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계 이후에, 각 파티의 이벤트 타임 데이터에 대해서 서바이벌 함수를 연산하는 단계를 더 포함할 수 있다.

[0067] 손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계(S410)는 손실 값역수를 기준으로 각 파티 의 가중치를 계산하고, m개의 평균으로 최종 가중치 를 도출 후, 통합 모델을 구축하는 단계이다.

[0068] 각 파티의 이벤트 타임 데이터 $\{t_1, t_2, \ldots, t_{N_d}\}$ 에 대해서 서바이벌 함수를 연산하는 단계(S500)는 각 파티별 $\sum_{j \in R(t)} \exp(x'_i \hat{\beta}_{IM})$ 를 연산할 수 있다.

[0069] 각 파티에 대해서 서바이벌 함수값을 더해서, 중앙 타임 포인트에서의 중앙 서바이벌값을 추정하는 단계(S600)는 $\{t_1, t_2, \ldots, t_{N_d}\}$ 에서 각 파티별 $\sum_{j \in R(t)} \exp(x'_i \hat{\beta}_{IM})$ 를 더해서 중앙 서바이벌 값을 추정하고, 추정된 중앙 서바이벌 값의 추정치를 각 타임 포인트에서의 포인트 서바이벌 값을 추정하는 것일 수 있다.

[0070] 예를 들어, 이벤트가 발생한 타임 포인트마다 복수 개(예를 들어, 200개 이상)의 통합 모델을 기반으로 복수 개의 중앙 서바이벌이 추정되고, 200개의 평균으로 포인트 서바이벌 값을 최종 추정할 수 있다.

[0071] 도 5는 반복회수에 따른 가중치값을 나타낸 그래프이다. 3개의 파티를 대상으로 반복회수를 200,400,600,800 및 1000번으로 한 경우의 가중치 양상이 도시되었다. 수진 점선은 200번 반복한 경우를 나타내는 도시선이다.

[0072] 일반적으로 반복회수가 200에 도달하는 경우에 가중치가 포화되는 양상을 보이고 있으므로, 연산의 효율성을 위해서는 200번 부근, 예를 들어, 170번 내지 230번을 반복하는 것이 효율적일 수 있으나 이에 한정되는 것은 아니다.

[0073] 도 6은 동일한 데이터 특성 하에서, 데이터 사이즈에 따른 가중치 패턴의 변화를 도시한 그래프이다. 도 6에 도시된 바와 같이, 시나리오 1에 해당하는 C의 크기를 1000, 2000, 3000, 000, 5000으로 변화함에 따른 해당하는 파티 A, B, C에 대한 중앙 데이터의 부합 정도를 나타내고 있다.

[0074] 도 7은 동일한 데이터 사이즈 하에서, 중앙 데이터와의 부합정도에 따른 가중치 패턴의 변화를 도시한 그래프이다. 시나리오 2에 해당하는 크기 1000에 해당하는 파티 A, B, C에 대한 중앙 데이터의 부합 정도를 나타낸 그래프이다.

[0075] 도 8은 각 기관별로 10개의 모델에 대한 가중치값을 예시적으로 나타낸 도면이다. 병원 1번부터 10번 까지 200번 반복한 경우의 손실값, 가중치, AUC 값, n값 등이 도시된다.

[0076] 도 9는 로지스틱 회귀 모델에 대한 예측 정도를 비교도시한 그래프이다. 도 9를 참조하면 총 2,845개의 ICU모델 입원을 기준으로 한 중앙 집중식 LR 모델과 본 발명에 따른 가중치 통합 모델의 ROC 곡선, AUC 곡선 등을 통한 로지스틱 회귀 모델의 예측력이 비교된다.

[0077] 도 10은 10개의 모델, 가중치 통합 모델, 중앙화 모델의 외부 검증값을 비교도시한 그래프이다.

[0078] 도 10을 참조하면, 중앙 집중식 모델, WIM 및 각 병원의 10 개 모델에 대한 외부 검증의 AUC 결과 (오차 막대 : 95 % CI). 검은 색, 짙은 회색, 밝은 회색은 각각 WIM, 중앙 집중식 모델 및 각 병원의 10 개 모델을 나타낸다.

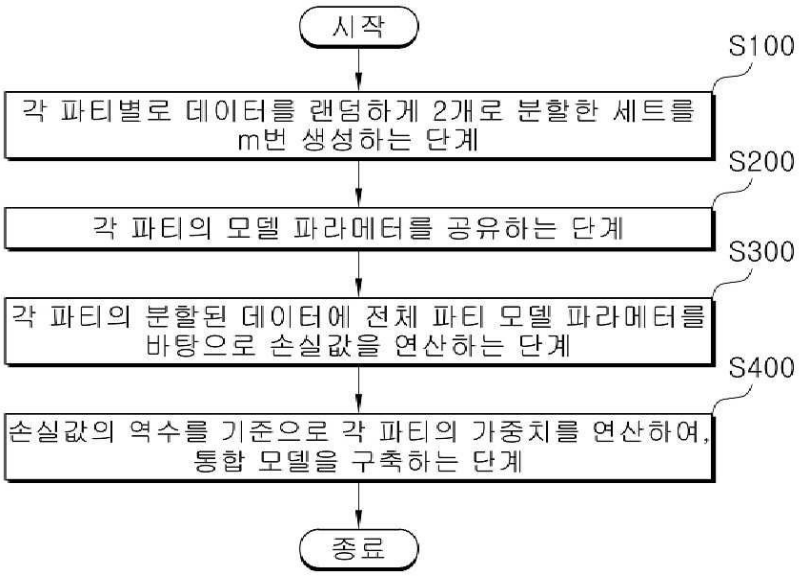
[0079] 도 11은 로지스틱 회귀 모델을 적용한 모델에 대한 OR수치를 비교도시한 그래프이다. 도 11을 참조하면, 첫 번째 로지스틱 회귀 모델의 11 개 특성에 대한 추정 OR 및 95 % CI 비교 데이터가 도시된다.

[0080] 본 발명의 실시예와 관련하여 설명된 방법 또는 알고리즘의 단계들은 하드웨어로 직접 구현되거나, 하드웨어에 의해 실행되는 소프트웨어 모듈로 구현되거나, 또는 이들의 결합에 의해 구현될 수 있다. 소프트웨어 모듈은 RAM(Random Access Memory), ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리(Flash Memory), 하드 디스크, 착탈형 디스크, CD-ROM, 또는 본 발명이 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터 판독가능 기록매체에 상주할 수도 있다.

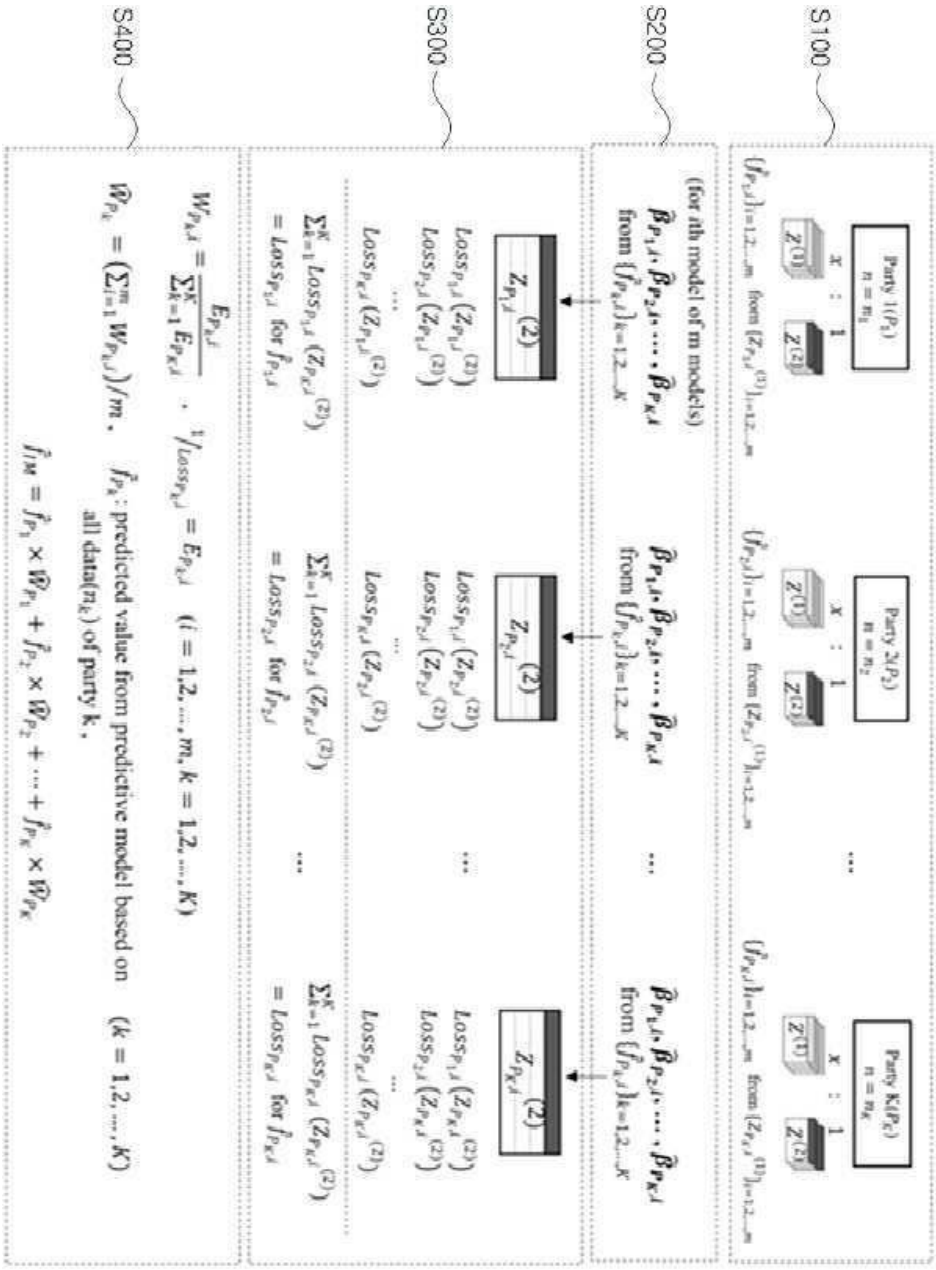
[0081] 이상, 첨부된 도면을 참조로 하여 본 발명의 실시예를 설명하였지만, 본 발명이 속하는 기술분야의 통상의 기술자는 본 발명이 그 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 그러므로, 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며, 제한적이 아닌 것으로 이해해야만 한다.

도면

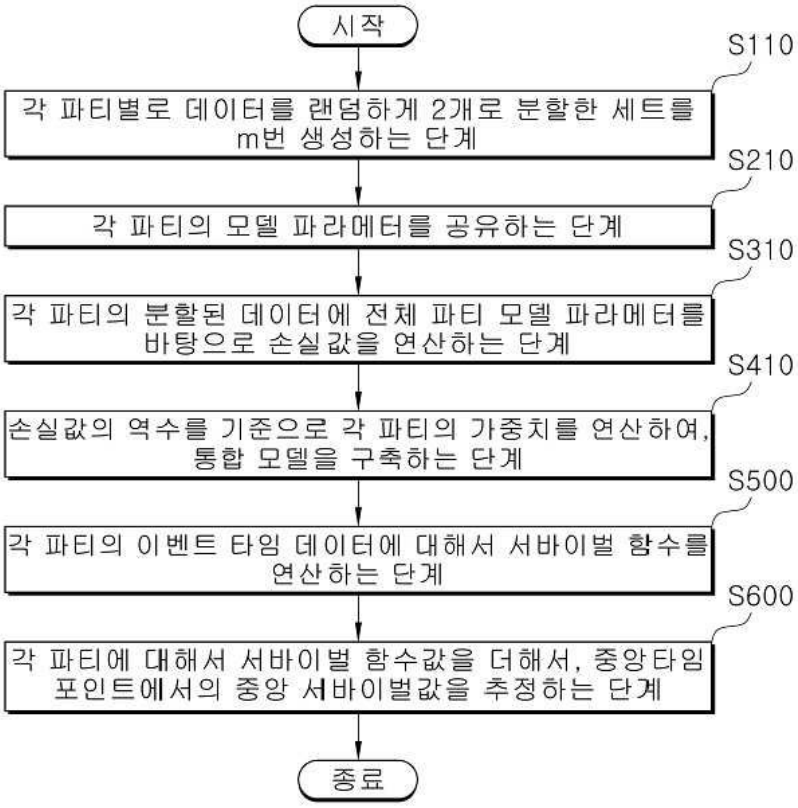
도면1



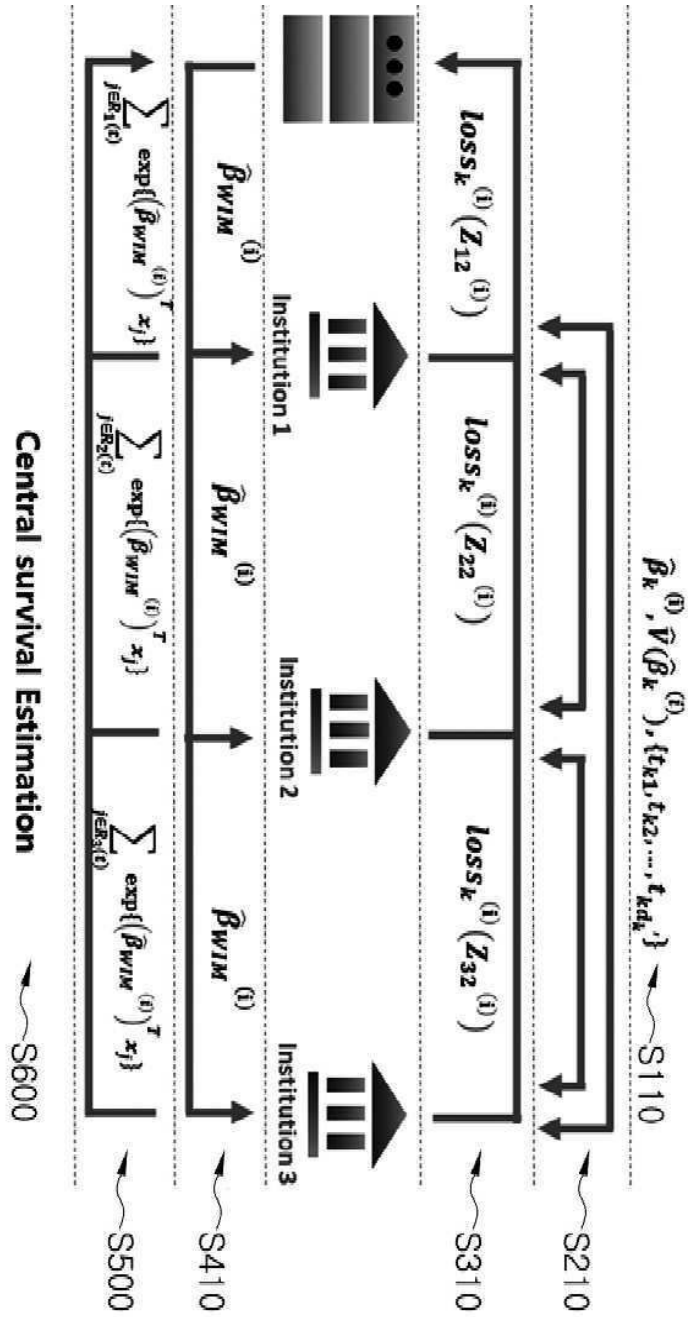
도면2


S100
S200
S300
S400
Party 1(P_1)
Party 2(P_2)
Party K(P_K)
(for ith model of m models)
predicted value from predictive model based on all data(n_k) of party k,
$i = 1, 2, \ldots, m, k = 1, 2, \ldots, K$
$k = 1, 2, \ldots, K$

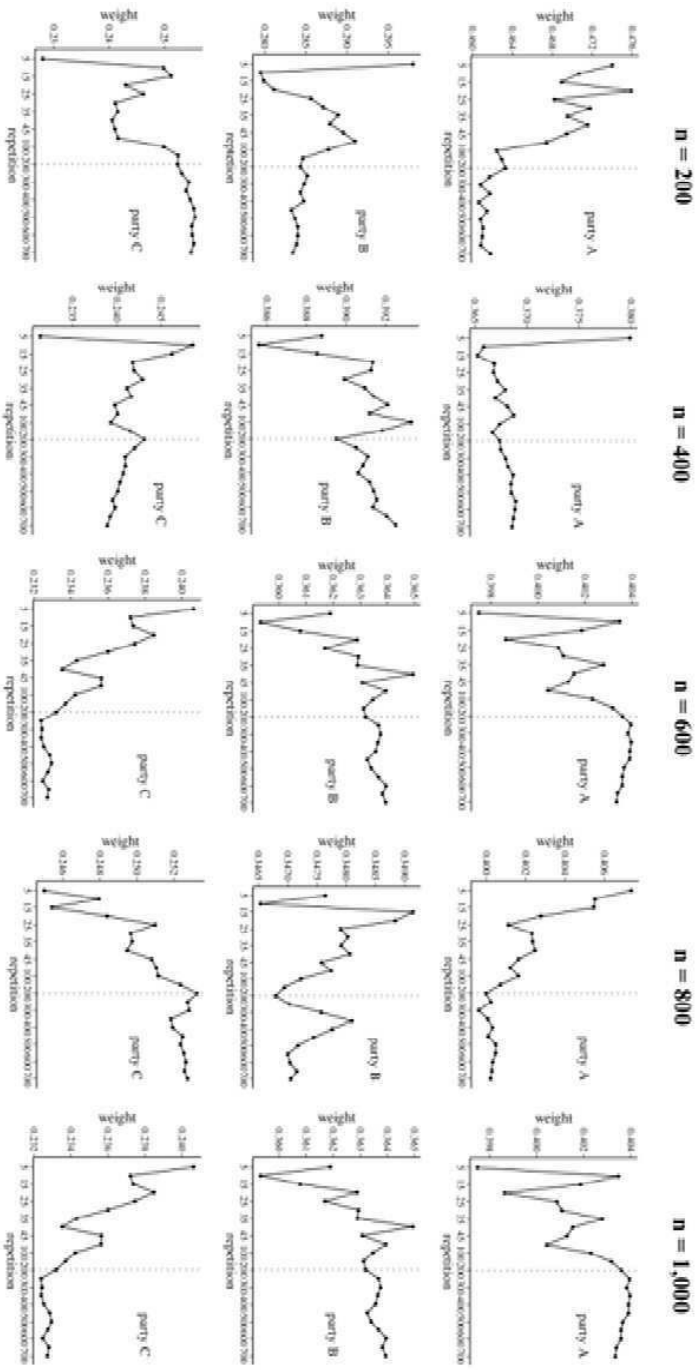
도면3


시작
S110
각 파티별로 데이터를 랜덤하게 2개로 분할한 세트를 m번 생성하는 단계
S210
각 파티의 모델 파라메터를 공유하는 단계
S310
각 파티의 분할된 데이터에 전체 파티 모델 파라메터를 바탕으로 손실값을 연산하는 단계
S410
손실값의 역수를 기준으로 각 파티의 가중치를 연산하여, 통합 모델을 구축하는 단계
S500
각 파티의 이벤트 타임 데이터에 대해서 서바이벌 함수를 연산하는 단계
S600
각 파티에 대해서 서바이벌 함수값을 더해서, 중앙타임 포인트에서의 중앙 서바이벌값을 추정하는 단계
종료

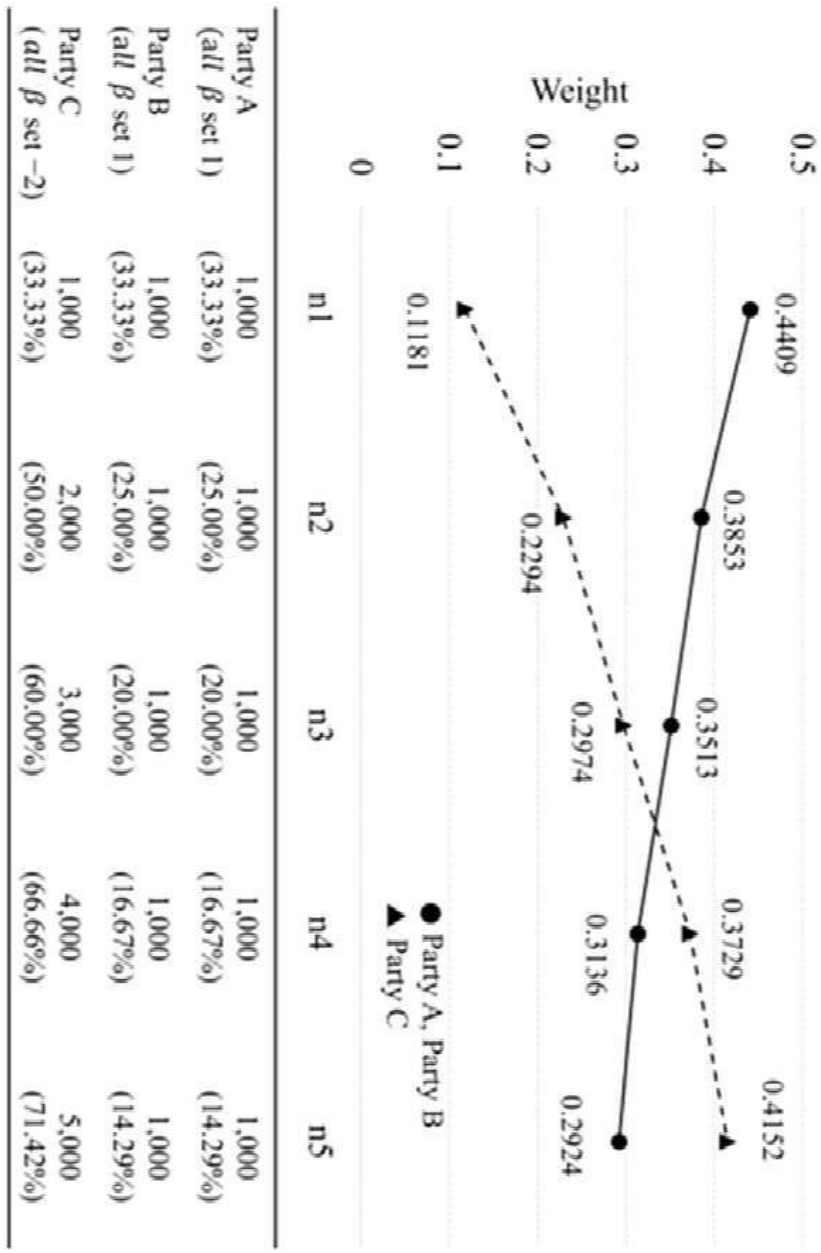
도면4


S110
S210
S310
S410
S500
S600
Institution 1
Institution 2
Institution 3
Central survival Estimation

도면5

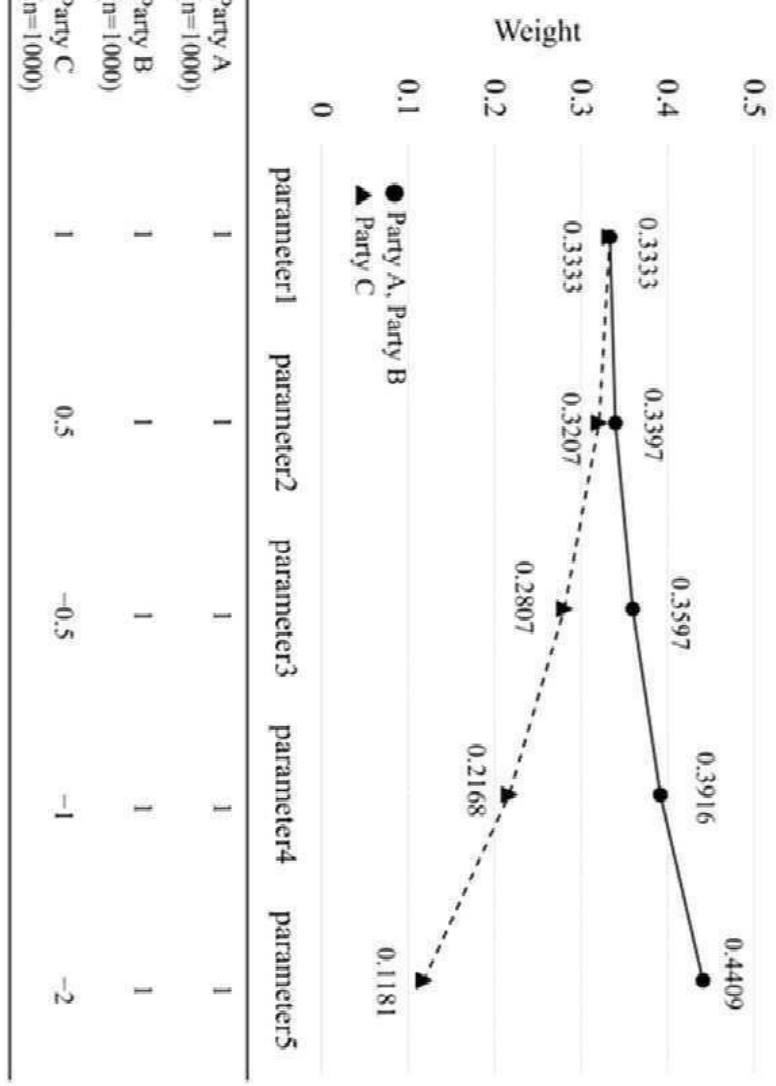

n = 200
n = 400
n = 600
n = 800
n = 1,000
party A
party B
party C
weight
replication

도면6



Party A (*all β* set 1)	1,000 (33.33%)	1,000 (25.00%)	1,000 (20.00%)	1,000 (16.67%)	1,000 (14.29%)
Party B (*all β* set 1)	1,000 (33.33%)	1,000 (25.00%)	1,000 (20.00%)	1,000 (16.67%)	1,000 (14.29%)
Party C (*all β* set −2)	1,000 (33.33%)	2,000 (50.00%)	3,000 (60.00%)	4,000 (66.66%)	5,000 (71.42%)

도면7

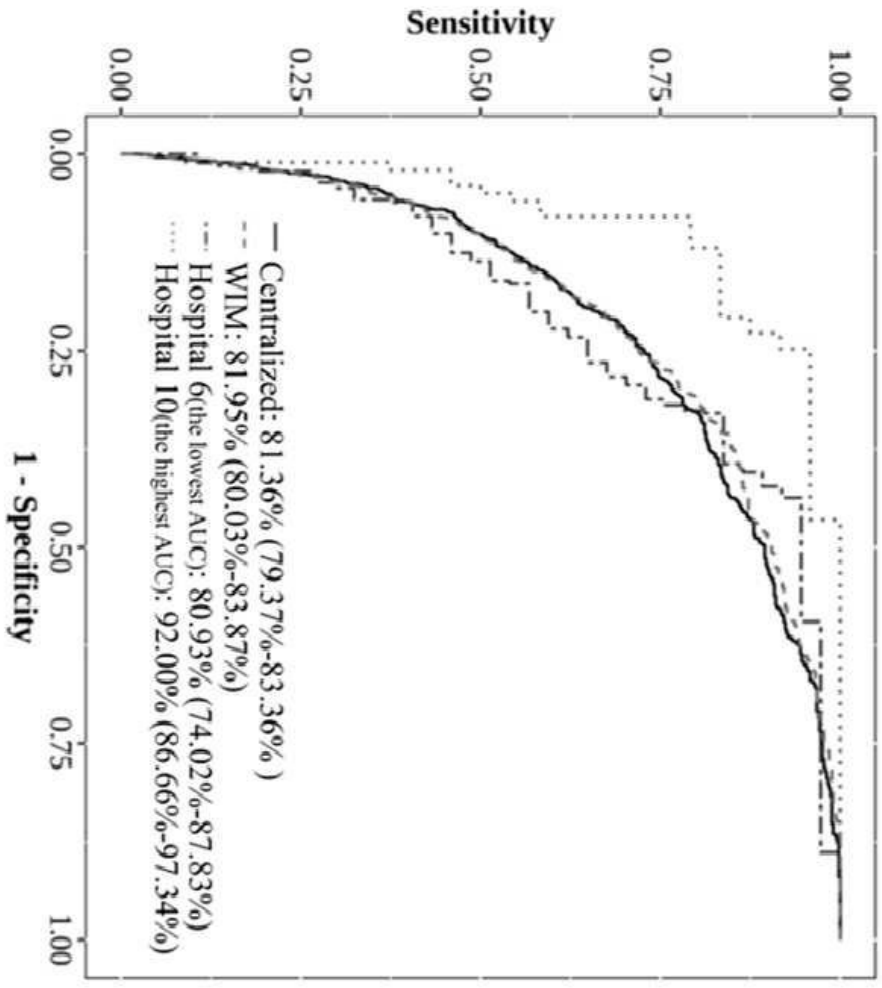


	parameter1	parameter2	parameter3	parameter4	parameter5
Party A (n=1000)	1	1	1	1	1
Party B (n=1000)	1	1	1	1	1
Party C (n=1000)	1	0.5	−0.5	−1	−2

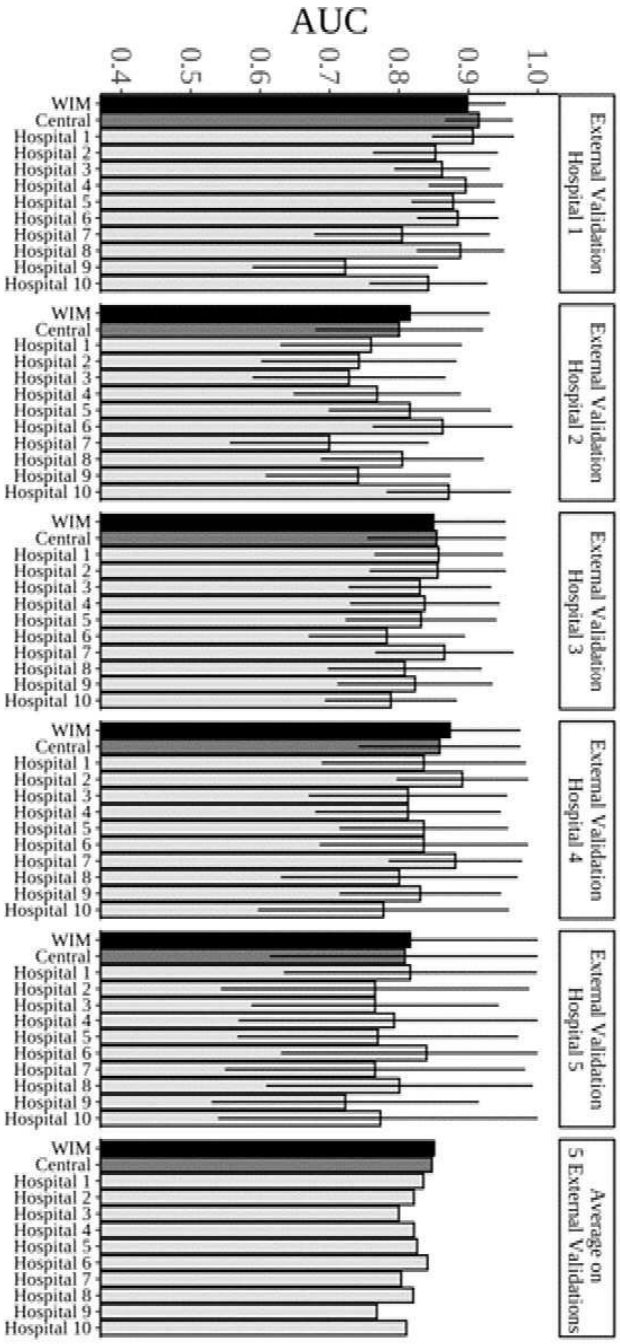
도면8

hospital num	n (%)	AUC (95% CI)	log loss from 200 repetitions (min, max)	log loss from 200 repetitions median	weight
1	510 (17.93%)	83.81% (79.99%-87.63%)	(535.45, 668.13)	575.18	0.1188
2	387 (13.6%)	82.14% (76.82%-87.47%)	(536.59, 754.68)	577.40	0.1181
3	268 (9.42%)	86.67% (81.57%-91.78%)	(547.65, 755.15)	616.63	0.1109
4	338 (11.88%)	86.48% (81.43%-91.53%)	(552.61, 787.62)	617.14	0.1109
5	231 (8.12%)	86.29% (80.19%-92.4%)	(572.31, 1814)	723.90	0.0929
6	316 (11.11%)	80.93% (74.02%-87.83%)	(539.71, 978.16)	626.65	0.1076
7	308 (10.83%)	85.95% (78.23%-93.67%)	(561.92, 1071.16)	665.89	0.1024
8	197 (6.92%)	83.81% (75.88%-91.73%)	(569.31, 7280.35)	712.29	0.0912
9	165 (5.8%)	86.63% (79.2%-94.05%)	(566.39, 1774.99)	758.66	0.0890
10	125 (4.39%)	92% (86.66%-97.34%)	(634.35, 13722.49)	1008.64	0.0583

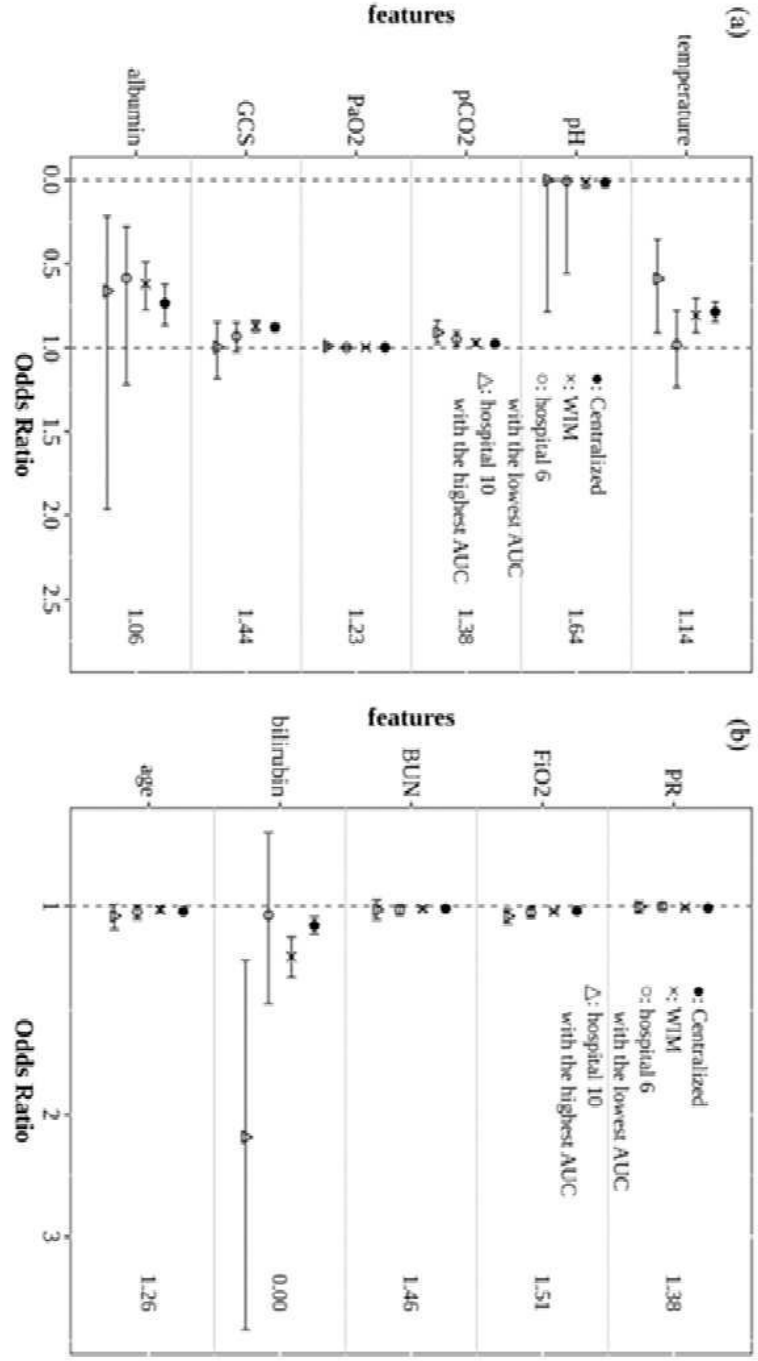
도면9


Sensitivity
0.00
0.25
0.50
0.75
1.00
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Centralized: 81.36% (79.37%-83.36%)
WIM: 81.95% (80.03%-83.87%)
Hospital 6(the lowest AUC): 80.93% (74.02%-87.83%)
Hospital 10(the highest AUC): 92.00% (86.66%-97.34%)

도면10


AUC
0.4
0.5
0.6
0.7
0.8
0.9
1.0
WIM
Central
Hospital 1
Hospital 2
Hospital 3
Hospital 4
Hospital 5
Hospital 6
Hospital 7
Hospital 8
Hospital 9
Hospital 10
External Validation Hospital 1
External Validation Hospital 2
External Validation Hospital 3
External Validation Hospital 4
External Validation Hospital 5
Average on 5 External Validations

도면11


(a)
features
temperature
pH
pCO2
PaO2
GCS
albumin
1.14
1.64
1.38
1.23
1.44
1.06
●: Centralized
×: WIM
○: hospital 6
with the lowest AUC
△: hospital 10
with the highest AUC
Odds Ratio
0.0
0.5
1.0
1.5
2.0
2.5
(b)
features
PR
FiO2
BUN
bilirubin
age
1.38
1.51
1.46
0.00
1.26
●: Centralized
×: WIM
○: hospital 6
with the lowest AUC
△: hospital 10
with the highest AUC
Odds Ratio
1
2
3