Journal of the Korean Data & Information Science Society
2024, **00**(0), 1–11

한국데이터정보과학회지

SEC 프로세스를 통한 이종 산업 간 데이터 결합 문제 해결

정수민[1], 오현진[2], 정은혜[3], 조수현[4]

[1]경북대학교 통계학과 · [2]한국교통대학교 소프트웨어학전공 · [3]성신여자대학교 심리학과 · [4]명지대학교 AI정보과학전공

접수 0000년 0월 1일, 수정 0000년 0월 0일, 게재확정 0000년 0월 0일

요약

개인화 서비스의 중요도가 높아짐에 따라 다양한 산업 간 데이터 결합이 활성화되고 있다. 그러나 기존 데이터 결합 방식은 민감한 정보를 포함한 데이터를 직접 주고 받기 때문에 개인정보 유출 가능성이 높다는 문제점을 가지고 있다. 이에 본 연구에서는 성별 또는 연령과 같은 최소한의 고유정보를 활용해 데이터를 결합하는 방식을 제안한다. 데이터 보안을 강화하기 위해 연합 학습(Federated Learning)과 분할 학습(Split Learning)을 차용하여 모델을 학습시켰으며, 그 결과 SEC 프로세스를 통해 데이터를 결합한 경우, 결합 전 단일일 데이터에 비해 더욱 향상된 예측 성능을 보였다. 이처럼 SEC 프로세스를 통해 최소한의 고유정보를 사용한 데이터 결합은 개인정보를 침해를 예방하여 프라이버시 보호를 강화한다는 점에서 기존 데이터 결합 방식의 문제 개선과 더불어, 더욱 향상된 예측을 통해 다양한 서비스에 적용할 수 있을 것으로 기대된다.

주요용어: 개인정보 보호, 분할학습, 연합학습, 이종 산업 간 데이터 결합, SEC 프로세스, SHAP, STC 프로세스.

1. 서론

가장 대표적인 국가 경쟁력지수 측정 기관인 IMD(International Institute for Management Development)에서 발표한 자료에 따르면, 대한민국은 World Digital Ranking 6위로 매우 높은 수준의 빅데이터 활용도를 갖췄다(Bris, 2023). 이에 더하여 대한민국 정부는 다양한 분야에서 빅데이터를 활용하며 데이터 경제 시대를 선도하기 위해 노력하고 있으며, 이러한 예시로 금융위원회는 결합할 데이터를 보유하지 않은 기관이 타 기관의 데이터를 결합 및 활용할 수 있도록 제도를 개선하는 등 다양한 분야의 데이터 결합 활성화를 위해 힘쓰고 있다(금융위원회, 2022). 다양한 데이터 간 결합을 통해 만들어진 빅데이터를 활용하는 능력은 새로운 서비스나 산업 개발에 근간이 되고 있으며, 기업의 경쟁력에도 많은 영향을 주고 있다(Kim, 2020).

그러나 대부분의 데이터가 생성 과정에서 주체와 관련된 정보를 포함하고 있어 개인정보와 데이터의 동일시 여기는 문제가 발생하므로(Cho, 2017) 데이터의 직접적인 공유를 피하려는 경향으로 이어진다. 또한, 기존의 데이터 결합 방식은 민감한 정보가 포함된 데이터를 직접 공유하기 때문에 개인정보 노출,

[1] (41566) 대구광역시 북구 대학로 80, 경북대학교 통계학과. E-mail: datalover_sumin@naver.com
[2] (27469) 충청북도 충주시 대학로 50, 한국교통대학교 소프트웨어학전공. E-mail: zzzini924@gmail.com
[3] (02844) 서울특별시 성북구 보문로 34다길 2, 성신여자대학교 심리학과, 석사과정. E-mail: eunhye.choung@gmail.com
[4] 교신저자: (03674) 서울특별시 서대문구 거북골로 34, 명지대학교 AI정보과학전공, 겸임교수. E-mail: whtngus3232@gmail.com

주체의 특정 가능성 등의 위험부담을 가지고 있다는 점에서 개인정보 보호와 데이터 활용은 이해관계가 상충하는 패러독스(Paradox) 성질을 가진다(Kim, 2014). 따라서 이러한 문제를 해결하기 위해 데이터 결합 방식에 대한 연구는 필수적이다.

최근 국내에서 실제 데이터의 결합을 바탕으로 진행된 연구들의 경우, 모바일 통신 기지국 데이터와 교통카드 데이터 등의 다양한 모빌리티 데이터를 가상으로 결합하여 특정 인원의 위치를 파악하고 분석하는 방법론과 같이 특정 분야에 국한되어 진행하는 연구나, 대부분 법정책적 연구가 주를 이룬다는 한계점을 보이고 있다(Cho 등, 2022). 이러한 한계점을 극복하고자, 본 연구는 성별, 나이와 같은 최소한의 고유정보만을 사용하여 데이터의 특성을 고려하지 않고 데이터를 결합하기 위한 방법론을 제시한다.

본 논문에서는 최소한의 고유정보로 성별과 나이만을 포함하는 익명의 신용카드 연체 여부 데이터와 건강보험 관심도 데이터를 활용한다. 두 가지의 데이터 결합 방법론을 제안하며, 전체적인 구성은 Feature 1.1과 같다. 모델의 학습 과정에서는 데이터 보안을 강화하기 위한 목적으로 연합학습(Federated Learning)과 분할학습(Split Learning) 방식을 사용하였으며, 제안한 방법론을 활용한 분석 결과를 설명하기 위해 XAI(eXplainable Artificial Intelligence)기법을 적용한다.

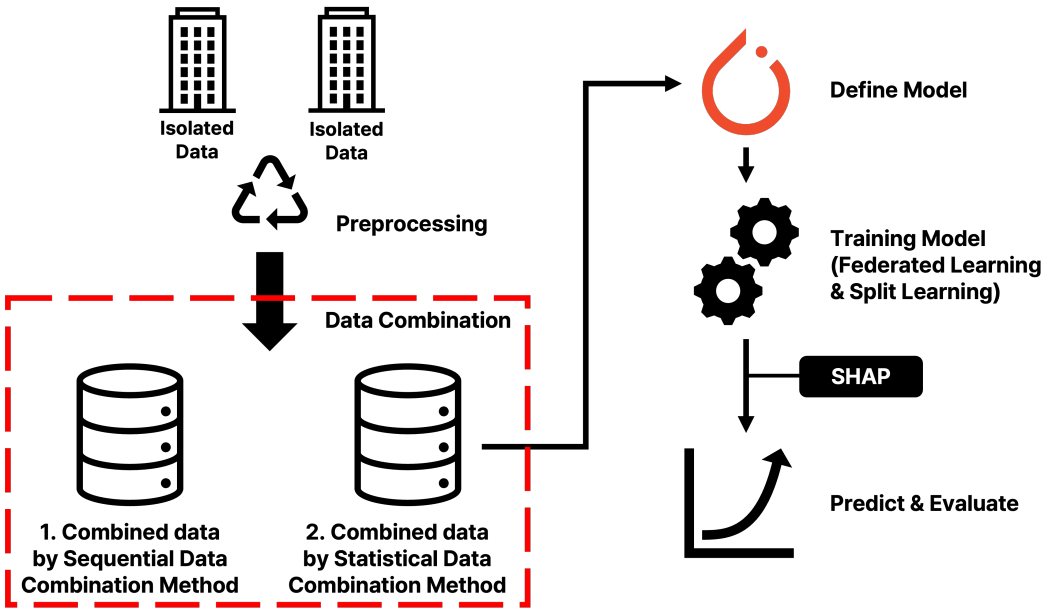


Figure 1.1: workflow of the research

논문의 구성은 다음과 같다. 먼저 2절에서 연합학습(Federated Learning)과 XAI의 SHAP 기법에 대해 간략하게 소개하고, 3절에서는 본 연구의 분할학습(Split Learning)과 데이터 결합 방법론으로 제안한 SEC 프로세스와 STC 프로세스에 대해 설명한다. 4절에서는 딥러닝 기반 분류 모형에 대한 설명과 성능을 비교하고, 마지막으로 5절에서 본 연구의 결과를 요약하고 연구의 시사점을 제시한다.

2. 이론적 배경

2.1. 연합학습 (Federated learning)

Jakub 등(2016)은 훈련 데이터를 하나의 기계나 중앙 서버에 모아 머신러닝을 진행하는 기존의 방법에서 더 나아가 데이터의 탈중앙화를 제시하는 연합학습(Federated Learning)을 제시하였다. 연합학습은 데이터를 외부로 노출시키지 않고 모델을 학습시킨다는 장점을 가지고 있어 개인정보 유출 문제를 완화할 수 있다.

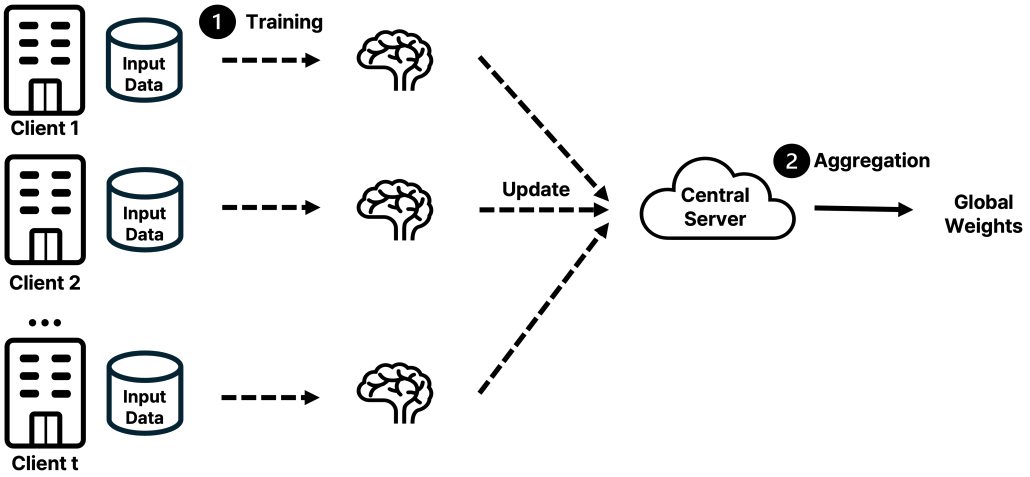


Figure 2.1: Federated Learning

Figure 2.1과 같이 연합학습은 로컬에서 자신의 데이터로 학습한 후 모델의 개별 가중치 업데이트와 중앙 서버로 전달하여 집계하는 통합 업데이트의 두 가지 과정이 있다(Kim 등, 2023; Kong 등, 2022; Goetz 등, 2019).

연합학습에서 개별 가중치 업데이트에 대한 계산식은 아래와 같다.

$$U_i^t := W_i^t - W^t \tag{2.1}$$

식 2.1에서 W는 학습 모델의 가중치, t와 i는 라운드와 로컬이다. 이렇게 계산된 H_i를 중앙 서버로전달하면 진행하는 통합 업데이트 계산식은 아래와 같다.

$$W^{t+1} := W^t + n_t \sum \frac{1}{n} U_i^t \tag{2.2}$$

식 2.2에서 n_t는 학습률로 단순히 평균만을 계산하기 위해 1로 설정한다(Jakub 등, 2016; Dhakal 등, 2020).

2.2. SHAP 분석

Lundberg와 Lee (2017)에 의해 발표된 SHAP(SHapley Addictive exPlanations)는 다중공선성이 존재하는 선형 모델의 Shapley value를 통해 예측 모델의 출력에 대한 각 feature의 중요도를 제공하는

기법이다. SHAP은 특정 예측에서 feature들에 중요도를 부여하며, Shapley value는 식 2.3을 통해 계산된 ϕ_i에 해당하는 값이다(Kim, 2024; Van den Broeck 등, 2022).

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - fS_{(x_S)}]. \tag{2.3}$$

식 2.3에서 F는 모든 변수의 집합, x_s는 S에 포함된 변수만 포함한 input, f_s는 학습모델을 의미한다. 계산된 Shapley value가 양의 값을 가지면 해당 변수가 예측에 긍정적인 기여를, 음의 값을 가지면 해당 변수가 예측에 부정적인 기여를 한다는 것을 의미한다.

본 논문에서는 Shapley Value와 DeepLIFT(Deep Learning Important FeaTures)의 결합을 통해 딥러닝 예측 모델의 출력을 설명하는 Deep SHAP 기법을 사용하여(Han 등, 2023), 기존 데이터와 결합된 데이터를 사용해 학습한 딥러닝 예측 모델에 대한 Shapley Value를 비교함으로써 데이터 결합이 예측에 미치는 영향을 확인하고자 한다.

$$m_{x_j f_3} = \frac{\phi_i(f_3, x)}{x_j - E[x_j]} \tag{2.4}$$

$$\forall_{j \in \{1,2\}}\ m_{y_i f_j} = \frac{\phi_i(f_j, y)}{y_i - E[y_i]} \tag{2.5}$$

$$m_{y_i f_3} = \sum_{j=1}^{2} m_{y_i f_j} m_{x_j f_3} \qquad \text{chain rule} \tag{2.6}$$

$$\phi_i(f_3, y) \approx m_{y_i f_3}(y_i - E[y_i]) \qquad \text{linear approximation} \tag{2.7}$$

DeepLIFT 수행 과정에 대한 수식을 바탕으로 Deep SHAP의 작동 방식을 알 수 있다. 식 2.4, 2.5에서는 레이어 f_3의 이전 레이어인 f_1, f_2에서 얻은 x_1, x_2의 기여도 점수 $\phi_i(f_3, x)$를 구한다. 이후 식 2.6, 2.7에서 DeepLift의 chain rule을 통해 전체 모델에 대한 기여도 점수를 결과로 얻게 된다.

3. 연구 방법론

3.1. 분할학습 (Split learning)

Praneeth (2018)가 발표한 내용에 따르면, 분할학습(Split learning)은 MIT에서 분산 딥러닝 메서드 SplitNN을 통해 민감한 데이터를 공유하지 않고도 기관들이 협업하여 딥러닝 모델을 훈련할 수 있도록 하는 기법이다. 대부분의 분할학습은 연합학습과 함께 연구되고 있다(Ryu 등, 2021; Thapa 등, 2022; Gao 등, 2020).

분할학습 과정은 각 클라이언트가 잘라진 레이어까지 모델을 훈련시키고 가중치를 서버에 보내는 것으로 시작된다. 그런 다음 서버는 나머지 레이어에 대해 모델을 훈련하며, 이것으로 전방향 전파가 완료된다. 그 후 서버는 최종 레이어의 그래디언트를 생성하고 오차를 잘라진 레이어까지 역전파한다. 그런 다음 그래디언트는 상대 클라이언트로 전달되며, 클라이언트에 의해 나머지 역전파가 완료된다. 이 과정은 모델이 훈련될 때까지 반복된다(Poirot 등, 2019; Kim 등, 2020).

본 연구에서는 기존의 분할학습 과정에서 데이터의 보안 및 예측 성능을 향상시키고자 서버에 데이터의 매칭 키를 전송한 후, 해당 키를 바탕으로 불러온 데이터를 서버에서 결합하는 방법을 고안하였다.

Algorithm 1: Split learning using SplitNN. The K clients are indexed by k, *key* is the data matching key, B is the local minibatch size, and η is the learning rate.

Server executes at round $t \geq 0$:
  for each client k $\in S_t$ **in parallel**
    $\mathbf{A}_t^k,\ key \leftarrow$ ClientUpdate(k, t)
    $\mathbf{B}_t^k = \phi$
    data = Map(key)
    Concatenate $f(data, \mathbf{H}_t^k)$ to $\mathbf{B}_t^k$
    $\mathbf{SEC}_t^k$ = Concatenate$(\mathbf{A}_t^k, \mathbf{B}_t^k)$
    Compute $\mathbf{W}_t \leftarrow \mathbf{W}_t - \eta \nabla \ell(\mathbf{W}_t; \mathbf{SEC}_t)$
    for each client p $\in S_t$ - {k} **in parallel**
      Send $\nabla \ell(\mathbf{SEC}_t; \mathbf{W}_t)$ to client p for OtherClientsBackProp(p, t)

ClientUpdate(k, t): // *Run on client k*
  $\mathbf{A}_t^k = \phi$
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      Concatenate $f(b, \mathbf{H}_t^k)$ to $\mathbf{A}_t^k$
      Send $\mathbf{A}_t^k,\ key$ to server

OtherClientsBackprop$(p, t, \nabla \ell(\mathbf{A}_t; \mathbf{W}_t))$: // *Run on clients except k*
  for batch $b \in \mathcal{B}$ **do**
    $\mathbf{H}_t^p = \mathbf{H}_t^p - \eta \nabla \ell(\mathbf{A}_t; \mathbf{W}_t; b)$

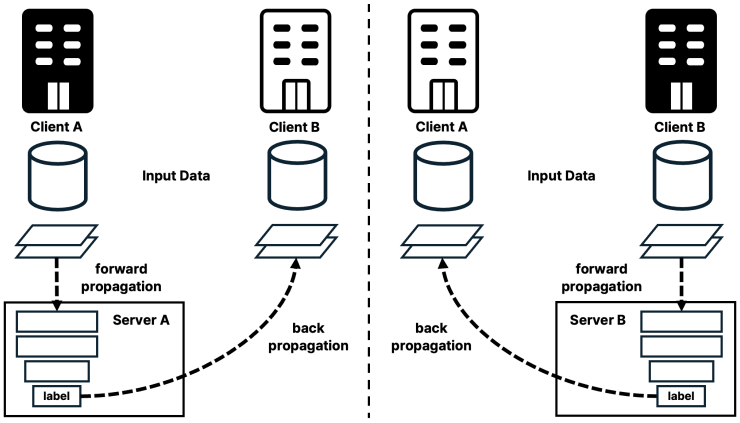


Figure 3.1: Split learning for vertically partitioned data

Figure 3.1은 본 연구에서 사용한 분할학습 과정을 나타내며, A는 신용카드 연체 여부 데이터, B는 건강보험 관심도 데이터라 가정한다. A는 자신이 보유한 데이터를 가지고 잘라진 레이어에 해당하는 부분까지

딥러닝 모델을 훈련시킨다. 마찬가지로 B 또한 자신이 보유한 데이터를 가지고 잘라진 레이어까지에 해당하는 부분 딥러닝 모델을 훈련시킨다. 각 클라이언트의 잘라진 레이어에서의 출력들은 연결되어 서버로 전송되어 forward propagation이 진행된다. 기존 모델과는 다르게 나머지 부분의 딥러닝 모델을 훈련시킨 다음, 상대 클라이언트의 잘라진 레이어로 back propagation이 진행된다. 이 과정에서 A와 B는 서로의 데이터를 공유하지 않고 분산된 딥러닝 모델의 훈련을 완료하기 위해 계속해서 진행한다.

3.2. SEC 프로세스

SEC(SEquential data Combination) 프로세스는 성별과 나이와 같은 고유정보를 키로 사용하여, 기준이 되는 데이터의 키와 동일한 키를 갖는 결합 대상 데이터를 일대일 방식을 통해 순서대로 연결해 가상의 데이터를 만들어내는 프로세스이다.

Client A			
A	남	21세	100
B	남	21세	200
C	남	21세	300
D	남	21세	400

Client B			
가	남	21세	50
나	남	21세	500

Virtual Data				
A	남	21세	100	50
B	남	21세	200	500
C	남	21세	300	50
D	남	21세	400	500

Figure 3.2: Sequential Data Combination Method

Figure 3.2는 신용카드 연체 여부 데이터(A)를 기준 데이터로 하여 건강보험 관심도 데이터(B)를 SEC 프로세스를 통해 결합하는 예이다. A과 B 모두 21세 남성의 데이터를 가지고 있지만 데이터의 개수가 다르다는 차이가 존재한다. SEC 프로세스를 사용하면 동일한 성별과 나이를 기준으로 B의 '가'와 '나' 데이터를 A의 'A'와 'B'에 일대일로 연결하게 된다. 마찬가지로 A의 'C'와 'D'도 B의 '가'와 '나'를 반복하여 순서대로 연결한다.

이처럼 SEC 프로세스를 사용하면, 결합하는 두 데이터의 수가 다르더라도 기준이 되는 데이터의 수를 유지할 수 있다. 또한, 성별과 나이라는 최소한의 고유정보를 사용하기 때문에 주체가 특정될 위험이 적다.

Algorithm 2: The way data combined through the SEC Process.

```
SEC Process(A, B):
  for each group in A:
    for each data in group:
      Concatenate(data, B.group[i])
      IF B.group fully matched THEN
        Match again from index 0 of B.group
      ELSE Match next index
  return A
```

3.3. STC 프로세스

STC(STatistical data Combination) 프로세스는 결합 대상 데이터의 성별과 나이와 같은 고유정보를 기준으로 그룹화 한 후, 각 그룹별로 평균, 중앙값, 최빈값 등 데이터의 특성에 맞는 다양한 통계를 이용

하여 집계한 값을 각 컬럼으로 만드는 과정을 거친 후, 그룹화된 통계 데이터를 기준 데이터와 결합하여 가상의 데이터를 만들어내는 프로세스이다.

Figure 3.3: Statistical Data Combination Method

Figure 3.3은 신용카드 연체 여부 데이터(A)를 기준 데이터로 하여 건강보험 관심도 데이터(B)를 STC 프로세스를 통해 결합하는 예이다. 먼저, B를 성별과 나이를 기준으로 그룹화한 후, 연속형 변수의 특성을 살려 평균, 중앙값, 최빈값을 계산하여 통계 데이터를 만든다. 동일한 기준을 적용하여 A의 21세 남성인 데이터 'A'와 'B'에는 B의 데이터 '1'을, A의 21세 여성인 데이터 'C'는 B의 데이터 '2'를, A의 데이터는 'D'는 24세 여성이므로 B의 데이터 '3'을 연결한다.

STC 프로세스는 기준이 되는 컬럼을 그룹화하여 통계량을 계산하므로 이상치에 대한 영향이 감소한다. 또한, 기준으로 그룹화된 데이터이기 때문에 개인정보 노출 문제에 비교적 자유롭다.

Algorithm 3: The way combined through the STC Process.

```
Make_Stats(k):
  for each group in K:
    temp.group.append([group.avg, group.mod ...]) // statistic values
  return temp

STC Process(A, B):
  temp = Make_Stats(B)
  for each group in A:
    Concatenate(group, temp.group)
  return A
```

4. 분류 모델 및 결과

4.1. 딥러닝 기반 예측 모델

딥러닝 모델 구성에는 Pytorch 라이브러리를 사용했다. 두 개의 데이터 모두 예측하려는 값이 고르지 않았기 때문에 데이터의 불균형 비율을 고려하여 학습에 사용하였고, 예측 대상에 따라 이진 분류와 다중 분류 모델을 만들었다. 딥러닝 모델의 레이어 구성은 데이터별로 레이어가 달라지도록 레이어의 크기를 데이터의 feature 개수에 비례하도록 설정한 경우와, 데이터별로 동일한 레이어를 사용할 수 있도록 레이어의 크기를 상수 값으로 고정하는 경우를 비교하였다. 결과적으로, 모델의 레이어 크기를 상수 값으로 고정하는 방법이 더 적절하다고 판단하여, 해당 방법을 통해 딥러닝 예측 모델을 구성하였다.

4.2. 분석 결과

Table 4.1 Evaluation Metrics for Classification Model

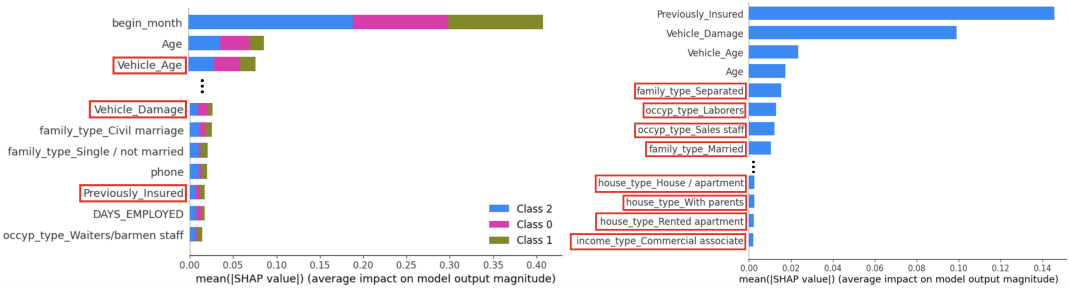
Data		Accuracy	Precision	Recall	F1 Score
Credit card delinquency prediction data	Original data	0.49	0.16	0.33	0.22
	SEC processed data	0.50	**0.44**	**0.39**	**0.35**
	STC processed data	**0.51**	0.32	0.38	0.31
Health insurance interest prediction data	Original data	0.68	0.40	0.00	0.00
	SEC processed data	**0.76**	0.58	**0.76**	**0.66**
	STC processed data	**0.76**	**0.60**	0.71	0.65

동일한 딥러닝 모델에 기존의 단일 데이터와 SEC 프로세스, STC 프로세스를 통해 결합한 데이터를 학습시킨 후 비교하였으며, 그 결과는 Table **??**과 같다. Accuracy만으로 예측 결과를 비교한다면, 결합 전 단일 데이터를 학습시킨 경우에 비해 결합 후 데이터를 학습시킨 경우의 정확도가 크게 상승하지 않았다.

그러나 실제 예측 결과를 직접 비교해본 결과, 결합 전 데이터는 예측이 특정 y값에 편향되어 정확도가 매우 떨어졌다. 반면 결합된 데이터는 하나의 값에 편향되는 현상을 보이지 않았으며, 특히 SEC 프로세스를 통해 데이터를 결합한 경우에서 예측한 y값이 조금의 편향도 없이 매우 고르게 분포하는 모습을 보였다. 이는 결합 전 데이터의 경우 feature의 갯수 뿐 아니라 데이터 자체의 갯수가 부족하다는 점과 STC 프로세스를 통해 결합한 데이터는 같은 그룹에 동일한 데이터가 결합된다는 점에 비해, SEC 프로세스를 통해 결합한 데이터는 결합을 통해 feature의 수가 충분히 증가했을 뿐 아니라, 같은 그룹의 데이터라도 서로 다른 데이터가 결합된다는 점이 주요하게 나타난 것으로 보인다.

정리 4.1 실제 예측 결과를 직접 비교해본 결과, 모델에 결합 전의 데이터를 학습시킨 경우에 비해 결합 후의 데이터를 학습시킨 경우가 훨씬 정확한 예측 성능을 보였다.

정리 4.2 SEC 프로세스를 통한 결합의 경우가 STC 프로세스를 통한 결합의 경우보다 더욱 정확한 예측 성능을 보였다.



(a) Credit card delinquency prediction data (b) Health insurance interest prediction data

Figure 4.1: SHAP values of SEC processed data

SEC 프로세스를 통해 결합한 데이터에 대한 SHAP values는 위의 Figure 4.1a, 4.1b와 같다. 두 그래프에서 공통적으로 기존 데이터의 features 뿐 아니라, 결합한 데이터의 features 또한 상당히 중요한 요소로 나타난다는 것을 확인할 수 있다. 이는 데이터 결합 이후 기존 데이터의 feature에 더해 결합한 데이터의 feature들 또한 상당히 중요한 요소로 사용되어 학습에 영향을 미치기 때문에, 예측 성능에 있어 SEC 프로세스를 통한 데이터 결합은 매우 유의미하다는 점을 시사한다.

5. 결론

본 연구에서는 최소한의 고유정보만을 사용하여 데이터를 결합하고 딥러닝 모델을 통해 가장 성능이 뛰어난 데이터 결합 방식을 찾아보고자 하였다. 결합하기 전 단일 데이터와 SEC 프로세스, STC 프로세스를 통해 결합한 데이터를 사용해 모델을 학습시킨 후 각 모델을 통해 예측을 수행한 결과, 결합 전 단일 데이터보다 결합된 데이터를 학습한 모델이 더욱 정확한 예측을 수행하였다. 특히 두 가지 데이터 결합 방법 중에서도 STC 프로세스를 사용한 경우에 비해 SEC 프로세스를 사용한 경우가 보다 나은 예측을 수행하였다. 이에 본 연구자들은 이종 산업 간 데이터 결합 문제를 해결하기 위한 방법으로 SEC 프로세스를 제시한다. SEC 프로세스를 사용한 데이터 결합은 개인정보 침해 문제의 해결과 더불어 예측의 정확도 또한 향상시킨다는 점을 시사한다. 또한 본 연구에서 사용한 데이터 결합 프로세스는 성별 및 나이와 같은 최소한의 고유정보만을 활용한다는 점에서, 이종 산업 뿐 아니라 다양한 산업 간 데이터 결합에 활용이 가능하기 때문에 활발한 데이터 결합을 통한 다양한 산업군의 교류 활성화 및 서비스 품질 향상이 기대된다.

참고문헌

Bris, A. (2023). IMD World Digital Competitiveness Ranking 2023. *IMD World Competitiveness Center*, 118-119.

Financial Services Commission. (2022). *Revised Rules on Credit Information Business to Enhance Convenience and Efficiency in Data Use*. Available from: `https://www.fsc.go.kr/eng/pr010101/78041`

Kim, S. G. and Kim, S. K. (2020). An Exploration on Personal Information Regulation Factors and Data Combination Factors Affecting Big Data Utilization. *Journal of the Korea Institute of Information Security and Cryptolog*, **30(2)** 287-304.

Cho, S. E. (2017). Improvement Strategy of Information Usability under Personal Information Protection Legal System. *Premium Report of Korea Information Society Development Institute*, 1-28.

Kim, S. K. (2014). A Study of the Personalization Service and Privacy Paradox in the Big Data Era-Focus on the Socio-technical Perspective. *Journal of the Korean Cadastre Information Association*, **16(2)**, 193-207.

Cho, B. C., An, D. B. and Kwon, K. H. (2022). Development of Virtual Fusion Methodology for Analysis Via Mobility Bigdata. *The Korea Journal of BigData*, **7(2)**, 75-90.

Jakub, K., H.Brendan, M., Felix, X. Y., Peter, R., Ananda, T. S. and Dave, B.(2016). Federated Learning: Strategies for Improving Communication Efficiency. *NIPS Workshop on Private Multi-Party Machine Learning (2016).*

Kim, J. S., Yang, S. M., Lee, K. Y. and Lee, K. K. (2023). Advances and Issues in Federated Learning Open Platforms: A Systematic Comparison and Analysis. *Journal of Internet Computing and Services*,**24(4)**, 1-13.

Kong, H. S., Yang and Kim, K. S. (2022). Self-supervised Meta-learning for the Application of Federated Learning on the Medical Domain. *Journal of Intelligence and Information Systems*,**28(4)**, 27-40.

Goetz, J., Malik, K., Bui, D., Moon, S., Liu, H. and Kumar, A. (2019). Active Federated Learning. *arXiv preprint, arXiv, 1909.12641v1.*

Dhakal, S., Prakash, S., Yona, Y., Talwar, S. and Himayat, N. (2020). Coded Federated Learning. *arXiv preprint, arXiv, 2002.09574v2.*

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, vol. 30, 2017.*

Kim, Y. G. (2024). Model Interpretation through LIME and SHAP Model Sharing. *Journal of The Institute of Internet, Broadcasting and Communication*,**24(2)**, 177-184.

Van den Broeck, G., Lykov, A., Schleich, M. and Suciu, D. (2022). On the tractability of SHAP explanations. *Journal of Artificial Intelligence Research*,**74**, 851-886.

Han, H. J., Kim, Y. S., Shim, J. W. and Jung, H. Y. (2023). Noise robustness analysis of Shapley value for Deep SHAP. *Journal of Korean Institute of Intelligence Systems*,**33(1)**, 23-28.

Vepakomma, P., Gupta, O., Swedish, T. and Raskar, R. (2019). Split learning for health: Distributed deep learning without sharing raw patient data. *ICLR AI for social good workshop 2019.*

Ryu, J., Won, D. and Lee, Y. (2021). A Study of Split Learning Model to Protect Privacy. *Jouranl of Information and Security.*,**21(3)**, 49-56.

Thapa, C., Arachchige, P. C. M., Camtepe, S. and Sun, L. (2022). Splitfed: When federated learning meets split learning. *In Proceedings of the AAAI Conference on Artificial Intelligence*,**36(8)**, 8485-8493.

Gao, Y., Kim, M., Abuadbba, S., Kim, Y., Thapa, C., Kim, K., Camtepe, S. A., Kim, H. and Nepal, S. (2020). End-to-end evaluation of federated learning and split learning for Internet of Things. *arXiv preprint, arXiv, 2003.13376v2.*

Poirot, M. G., Vepakomma, P., Chang, K., Kalpathy-Cramer, J., Gupta, R. and Raskar, R. (2019). Split learning for collaborative deep learning in healthcare. *arXiv preprint, arXiv, 1912.12115.*

Kim, J. W., Shin, S. H., Yu, Y. U., Lee, J. S. and Lee, K. B. (2020). Multiple Classification with Split Learning. *arXiv preprint, arXiv, 2008.09874v3.*

Journal of the Korean Data & Information Science Society
2024, **00**(0), 1–11

한국데이터정보과학회지

Solving Problems While Combining Data Between Different Industries Through SEC(SEquential data Combination) Process

Sumin Jeong[1]· Hyeonjin Oh[2]· Eunhye Choung[3]· Suhyun Cho[4]

Department of Statistics, Kyungpook National University
Department of Software, Korea National University of Transportation
Department of Psychology, Sungshin Women's University
Department of AI Information Science, Myongji University

Received 1 0000, revised 0 0000, accepted 0 0000

Abstract

As the importance of personalized services increases, data combination across different industries is becoming more active. However, existing data combining methods have the problem of high possibility of personal information leakage because of the direct exchange of data containing sensitive information. So in this study, we suggest a data combination method using minimal unique information such as gender and age. We used Federated Learning and Split Learning to enhance data security when training the model. As a result, combining data through the SEC process showed improved prediction performance compared to raw data. Data combination using minimal unique information through the SEC process enhances privacy protection by preventing infringement of personal information. So we expect this method can not only improving problems with existing data combination methods, but also be applied to various services through improved predictions.

Keywords: Data combination across different industries, Federated learning, Privacy protection, SEC process, SHAP, Split learning, STC process

[1] Department of Statistics, 80, Daehak-ro, Buk-gu, Daegu, Korea.
E-mail: datalover_sumin@naver.com
[2] Department of Software, 50, Daehak-ro, Daesowon-myeon, Chungju-si, Chungcheongbuk-do, Korea.
E-mail: zzzini924@gmail.com
[3] Department of Psychology, 2, Bomun-ro 34da-gil, Seongbuk-gu, Seoul, Korea.
E-mail:eunhye.choung@gmail.com
[4] Professor, Department of AI Information Science, 34 , Geobukgol-ro, Seodaemun-gu, Seoul
Email: whtngus3232@gmail.com