개인정보보호와 가명·익명화를 위한 데이터 마이닝 기술

이기용 숙명여자대학교 소프트웨어학부 교수

1. 머리말

데이터 마이닝(Data Mining)은 대규모 데이터로부터 숨겨진 패턴이나 정보 또는 지식을 찾아내는 기술을 말한다. 데이터 마이닝의 주요 세부 분야로는 분류(Classification), 회귀(Regression), 클러스터링(Clustering), 연관 규칙 탐색(Association rule discovery), 이상 탐지(Anomaly detection) 등이 있다[1].

- 분류: 기존 데이터의 클래스(Class)를 바탕으로 새로운 데이터의 클래스를 예측하는 작업
- 회귀: 분류와 유사하게, 새로운 데이터의 클래스 대신 해당 데이터와 관련된 어떤 수치를 예측하는 작업
- 클러스터링: 주어진 데이터들을 유사한 데이터끼리 묶어 그룹화하는 작업
- 연관 규칙 탐색: 데이터 간에 빈번히 발생하는 패턴을 찾아내는 작업
- 이상 탐지: 주어진 데이터 중 다른 데이터와 유난히 다른 성질을 가진 데이터를 찾아내는 작업

최근 마케팅, 금융, 과학, 쇼핑, 의료, 환경, 소셜 네트워크 등 다양한 분야에서 데이터 크기가 지속적으로 증가하고 있다. 이에 많은 조직 또는 기관이 데이터 마이닝 기술을 사용해 데이터로부터 유용하고 의미 있는 정보를 찾아내거나, 미래를 예측하는 데 힘을 쏟고 있다.

하지만 데이터 마이닝의 이면엔 개인정보 노출 및 사생활 침해라는 심각한 사회적 문제가 도사리고 있다[2]. 예를 들어, 개인 금융거래 정보는 데이터 마이닝을 통해 맞춤형 금융상품 제공이나 위험경고 등에 활용될 수 있으나, 반면 개인 재정 상태·수입과 같은 민감한 정보가 외부에 노출될 수 있다.

개인이 페이스북, 인스타그램 등 소셜 네트워킹 서비스에 올리는 수많은 데이터도 마찬가지다. 데이터 마이닝을 통해 맞춤형 광고, 생활정보 등에 활용될 수도 있지만, 개인 쇼핑 패턴, 정치 성향, 주거 지역 등 민감한 사생활 정보가 제3자에게 노출될 가능성이 있다.

최근엔 개인의 건강진단 결과, 진료기록과 같은 의료정보에 데이터 마이닝 기술을 적용해 맞춤형 의료 제공, 의료영상 자동 판독, 질병 예측 등 차세대 의료서비스를 제공하려는 연구가 활발히 진행되고 있다. 이 역시 개인의 병력과 건강 상태 등 민감한 개인정보가 외부로 노출될 수 있다는 위험이 존재한다.

마지막 예로, 개인 휴대폰을 통해 전송되는 위치정보도 위험하다. 이는 데이터 마이닝을 통해 사람들의 동선 파악, 이동 경로 예측 등에 활용될 수 있지만, 개인의 행동 패턴, 방문지 등 원치 않는 정보가 노출될 수 있다.

이러한 개인정보 노출은 사회문제로 나타날 수 있다. 이는 수입, 나이, 병력, 구매 패턴처럼 타인에게 노출하기 꺼림칙한 정보가 본인 의사와 관계없이 노출된다는 불쾌함, 지나친 마케팅이나 과도한 생활 침투처럼 일상생활에 불편함을 주는 문제, 보이스 피싱·금융사기·스토킹·개인정보 도용 범죄처럼 직접적이고 심각한 손해에 이르기까지 다양한 형태를 지닌다[2].

따라서 데이터 마이닝 기술의 활용에선, 데이터 패턴·정보·지식을 찾아내는 것만큼이나, 개인정보 유출을 막는 것도 의미 있는 연구 분야라 볼 수 있다.

이번 원고에선 데이터 마이닝 기술이 데이터 분석뿐만 아니라 개인정보보호와 가명·익명화에 어떻게 사용될 수 있을지, 앞으로의 가능성을 살펴보고자 한다. 이를 통해 데이터 마이닝의 적용 분야를 넓히는 한편, 개인정보를 더욱 안전하게 보호할 수 있는 기술이 연구되기를 기대한다.

2. 개인정보보호 및 가명·익명화를 위한 데이터마이닝 기술

2.1 개인정보보호 및 가명·익명화를 위한 분류 기술

데이터 마이닝의 분류 기술은 훈련 데이터에 부착된 클래스 또는 라벨(Label) 정보를 보고, 이를 바탕으로 새로운 데이터의 클래스 또는 라벨을 예측한다. 이를 위한 데이터 마이닝의 일반적 목표는 훈련데이터 x에 대해 그의 클래스 y를 잘 묘사하는 모델 y=f (x)를 찾는 것이다. 최근에는 좀 더 정확하고 정교한 모델을 찾기 위해 CNN, RNN, LSTM, 트랜스포머(Transformer) 등 다양한 딥러닝 모델 또는 그들의 조합이 활용되고 있다[3].

이러한 분류 기술은 크게 두 가지 방법으로 개인정보보호 및 가명·익명화에 활용될 수 있다. 첫째, 개인정보 노출 위험도가 높은 데이터를 파악할 수 있다. 예컨대, k -익명성, l -다양성, t -근접성과 같은 데이터 익명화 기술을 통해 익명화된 데이터일지라도, 각 데이터는 항상 개인정보가 노출될 가능성이 존재한다[2]. 이때 익명화된 데이터가 개인정보 노출 위험도가 높은지 또는 낮은지를 데이터 마이닝의 분류 모델을 활용해 예측할 수 있다. 이 경우 훈련 데이터로는 데이터에서 개인정보가 보호돼야 할 민감 속성을 제외한 부분을 x로, 개인정보가 보호돼야 할 민감 속성의값을 y(예: 병명)로 사용할 수 있다. 만약 이러한 훈련 데이터로 훈련된 모델이 주어진 x에 대해 y의 값을 제대로 예측하지 못할수록, x는 개인정보 노출 위험도가 낮다고 볼 수 있다.

또한 이러한 방법을 통해 데이터셋 전체에 대한 분류 모델의 정확도를 측정하면, 데이터셋 전체의 개인정보 노출 위험도 혹은 데이터셋에 적용된 익명화 기법의 효과를 평가할 수 있다. 예를 들어, [그림 1](a)는 원본 데이터를, [그림 1](b)는 k-익명화로 익명화된 데이터를 나타낸다. 익명화된 데이터에서 연령, 성별, 지역코드 값으로 병명을 예측하기 어려울수록 익명화기법의 효과가 더 높다고 할 수 있다.

두 번째, 분류 모델이 민감 속성의 값을 예측하는데 가장 영향을 많이 미치는 속성들을 XAI(설명가능인공지능, eXplainable AI) 기법으로 파악할 수 있다. 이 경우, 해당 속성들을 좀 더 우선적으로 삭제하거나 익명화함으로써 개인정보를 더욱 효과적으로 보호할 수 있다.

식별자	준식별자			민감속성
성명	연령	성별	우편번호	병명
김민준	20	남	02148	당뇨
이소연	35	여	02053	고혈압
박기혁	30	남	02148	간암
최민수	22	남	02067	독감
윤달수	29	남	02148	간암
문수빈	30	여	02067	폐암
조민희	21	여	02053	간염

(a) 원본데이터

연령	성별	지역코드	병명
[20-30]	남	02***	당뇨
[20-30]	남	02***	간암
[20-30]	남	02***	독감
[20-30]	남	02***	간암
[20-35]	여	02***	고혈압
[20-35]	여	02***	폐암
[20-35]	여	02***	간염

(b) 익명화된 데이터

[그림 1] 데이터 익명화의 예(k-익명화)[4]

2.2 개인정보보호 및 가명·익명화를 위한 회귀 기술

데이터 마이닝의 회귀 기술은 훈련 데이터에 존재하는 속성들의 값을 보고 이를 바탕으로 새로운 데이터의 특정 속성값을 예측한다. 이를 위한 데이터마이닝의 일반적인 목표는 분류와 유사한데, 훈련 데이터 x에 대해 어떤 특정 속성의 값 y를 잘 묘사하는 모델 y=f (x)를 찾는 것이다. 회귀 모델과 분류 모델의 차이는, 분류 모델이 클래스 혹은 라벨을 예측하는 한편 회귀 모델은 임의의 숫자 값을 예측한다는 것이다. 회귀 모델도 분류와 마찬가지로 좀 더 정확한 모델을 찾기 위해 다양한 딥러닝 모델 또는 그들의 조합이 활용되고 있다[3].

이러한 회귀 기술은 개인정보보호 및 가명·익명화에 다음과 같이 활용될 수 있다. 첫째, 분류 기술과 유사하게, 개인정보 노출 위험도가 높은 데이터 파악에 사용될 수 있다. 이 경우, 훈련 데이터로는 데이터에서 개인정보가 보호돼야 할 민감 속성을 제외한 데이터를 x로, 개인정보가 보호돼야 할 민감 속성의 값을 y로 사용할 수 있다. 단, 이 경우 y는 병명과 같은 클래스 또는 라벨이 아니라 연봉 혹은 몸무게와 같이 임의의 숫자가 올 수 있는 속성의 값이 된다. 만약 이러한 훈련 데이터로 훈련된 모델이 주어진 x에 대해 y의 값을 제대로 예측하지 못하는 경우, x는 개인정보 노출 위험도가 낮다고 볼 수 있다.

둘째, 데이터 가명 처리 혹은 익명화 과정에서 숫자 값을 가지는 어떤 속성의 값을 변형하고 싶을 때, 회귀 모델이 예측한 값을 대신 사용할 수 있다. 현재는 가명 처리를 위한 숫자 속성에 대해선, <표 1>에 나열된 바와 같이 주로 마스킹, 총계처리, 부분총계, 라운딩, 상하단코딩 등의 기법이 사용되고 있다. 이는 해당 속성의 값을 지나치게 변형해 데이터 분석을 어렵게 만든다는 단점이 있다. 따라서 이 경우 회귀 모델이 예측한 값을 사용하면 데이터 특징을 잘 유지하면서도 원 값을 다른 값으로 대체할 수 있다.

2.3 개인정보보호 및 가명·익명화를 위한 클러스터링 기술

데이터 마이닝의 클러스터링 기술은 데이터 객체들이 주어졌을 때 유사한 객체들끼리 모아 그룹 또는 클러스터를 생성한다. 이를 위한 데이터 마이닝의 일반적 목표는, 유사한 객체를 서로 동일한 클러스터에, 유사하지 않은 객체를 서로 다른 클러스터에 속하도록 하는 것이다. 이를 위해 k-평균, 계층적 클러스터링(Hierarchical clustering), DB-SCAN 등 다양한 클러스터링 알고리즘들이 존재하며, 이들은 클러스터링을 위해 각기 다른 전략을 사용한다. 최근에는 객체 간 유사도를 좀 더 정확히 측정하기 위해 딥러닝을 사용한 임베딩(Embedding) 기술이 많이 사용되고 있으며,

<표 1> 개인정보의 가명처리 기술 종류[5]

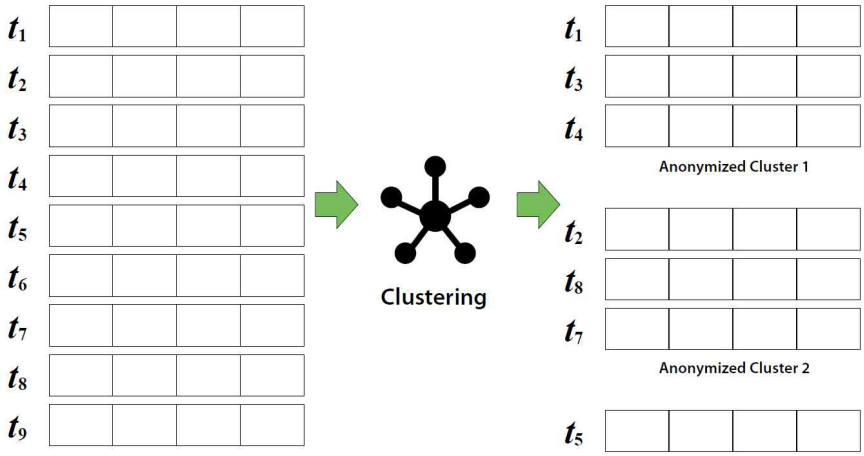
분류	기술	세부기술	설명
개인정보 삭제	삭제기술	삭제 (Suppression)	• 원본정보에서 개인정보를 단순 삭제
		부분삭제 (Partial suppression)	• 개인정보 전체를 삭제하는 방식이 아니라 일부를 삭제
		행 항목 삭제 (Record suppression)	• 다른 정보와 뚜렷하게 구별되는 행 항목을 삭제
		로컬 삭제 (Local suppression)	• 특이정보를 해당 행 항목에서 삭제
개인정보 일부 또는 전부 대체	삭제기술	마스킹 (Masking)	• 특정 항목의 일부 또는 전부를 공백 또는 문자('*', '_'등이나 전각 기호)로 대체
	통계도구	총계처리 (Aggregation)	• 평균값, 최댓값, 최소값, 최빈값, 중간값 등으로 처리
		부분총계 (Micro Aggregation)	• 정보집합물 내 하나 또는 그 이상의 행 항목에 해당하는 특정열 항목을 총계처리. 즉, 다른 정보에 비하여 오차 범위가 큰 항목을 평균값 등으로 대체
	일반화 (범주화) 기술	일반 라운딩 (Rounding)	• 올림, 내림, 반올림 등의 기준을 적용하여 집계 처리하는 방법으로, 일반적으로 세세한 정보보다는 전체 통계정보가 필요한 경우 많이 사용
		랜덤 라운딩 (Random rounding)	• 수치 데이터를 임의의 수인 자리 수, 실제 수 기준으로 올림(round up) 또는 내림(round down) 하는 기법
		제어 라운딩 (Controlled rounding)	• 라운딩 적용 시 값의 변경에 따라 행이나 열의 합이 원본의 행이나 열의 합과 일치하지 않는 단점을 해결하기 위해 원본과 결과가 동일하도록 라운딩을 적용하는 기법
		상하단코딩 (Top and bottom coding)	• 정규분포의 특성을 가진 데이터에서 양쪽 끝에 치우친 정보는 적은 수의 분포를 가지게 되어 식별성을 가질 수 있음 • 이를 해결하기 위해 적은 수의 분포를 가진 양 끝단의 정보를 범주화 등의 기법을 적용하여 식별성을 낮추는 기법

이를 통해 정형 데이터뿐만 아니라 텍스트, 이미지, 그래프 등 다양한 형태 데이터에 대한 클러스터링 기법이 개발되고 있다[6].

이러한 클러스터링 기술은 개인정보보호 및 가명·익명화에 다음과 같이 활용될 수 있다. 첫째, 데이터를 가명 처리하거나 익명화할 때 클러스터링 기술을 먼저 적용해 유사한 객체들을 그룹화한 후, 객체 수가 매우 적거나 기준 숫자 이하인 그룹들을 우선적으로 가명 처리 또는 익명화할 수 있다. 객체 수가 적은 그룹은 해당 속성 값을 가진 혹은 그들과 유사한 데이터의 수가 적다는 것을 의미하고, 이는 곧 해당 그룹에 속한 객체들의 개인정보 노출 위험도가 높다는 것을 뜻한다. 따라서 이 경우, 클러스터링 기술은 개인정보 노출 위험도가 높은 그룹 또는 객체를 우선적으로 파악하는 데 사용될 수 있으며, 위험도가 높은 그룹만 선별적으로 가명 처리하거나 익명화함으로써 데이터 품질이 크게 저하되는 것을 막을 수 있다.

둘째, k -익명화와 같은 익명화 기술들은 준식별자 값들의 각 조합에 대해서, 그에 해당하는 레코드들이 최소 k 개 이상이 되도록 하는 것이 필요한 경우가 많다. 이때 객체 수가 k 개 이상인 그룹을 빠르고 효과적으로 형성하기 위해 클러스터링 기술을 사용할 수 있다. 이 경우, 클러스터

링의 목표는 k 개의 그룹을 찾는 것이 아니라 객체 수가 k 개 이상인 그룹을 찾는 것이된다. 이러한 경우에도 목표는 조금 다르지만, 기존 클러스터링 기술을 변형해 데이터 익명화에 활용할 수 있다.



[그림 2] 클러스터링 기술을 활용한 익명화[7]

2.4 개인정보보호 및 가명·익명화를 위한 연관 규칙 탐색 기술

데이터 마이닝의 연관 규칙 탐색 기술은 데이터 객체들이 주어졌을 때 객체 간 빈번히 발생하는 패턴을 탐색하는 것이다. 가장 대표적인 예로는 빈번히 같이 나타나는 객체들을 찾거나, 빈번히 어떤 순서로 나타나는 객체들의 시퀀스(Sequence)를 찾는 것이다. 이를 위해 데이터 마이닝은 Apriori, FP-Growth 등 다양한 알고리즘을 사용해 빈번히 같이 발생하는 객체들 또는 그들의 시퀀스를 찾는다. 연관 규칙 탐색 기술은 빈번히 같이 구매되는 상품, 특정 순서에 따라 자주 발생하는 이벤트, 사용자의 웹 브라우징 패턴, 서로 다른 지역 간 연관성을 탐색하는 데 널리 사용된다[1].

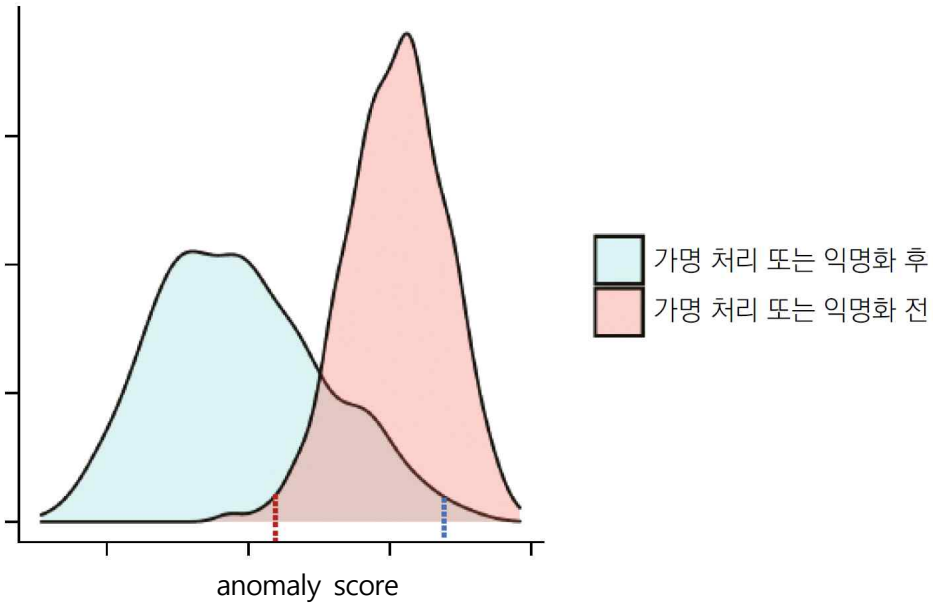
이러한 연관 규칙 탐색 기술은 개인정보보호 및 가명·익명화에 다음과 같이 활용될 수 있다. 첫째, 어떤 민감 속성의 값과 빈번히 같이 나타나는 준식별자의 값을 탐색하는 데 사용될 수 있다. 이 경우, 해당 준식별자 값을 변형함으로써 해당 민감 속성값의 노출 위험도를 크게 낮출 수 있다. 예를 들어, n개의 준식별자에 대한 값 x1, x2, ..., xn과 민감 속성의 값 y에 대해 연관 규칙 {x1, x2, ..., xn}→y가 탐색됐다고 하자. 이 경우, x1, x2, ..., xn의 값을 우선적으로 변형함으로써, 민감 속성의 값이 y인 레코드의 개인정보 노출 위험도를 크게 낮출 수 있다. 이때 준식별자 값을 변경할 때는 마스킹, 총계처리, 부분총계, 라운딩, 상하단 코딩 등 기존 가명처리 기법을 이용할 수도, 앞에서 설명한 분류·회귀 모델을 사용해 준식별자의 값을 모델이 예측한 값으로 대체할 수도 있다.

둘째, 클러스터링 기술과 유사하게 빈번히 같이 발생하는 준식별자의 값 {x1, x2, ..., xn}을 찾고, 준식별자의 값으로 이들을 가지는 데이터를 제외한 데이터에 대해 가명 처리 또는 익명화를 수행할 수 있다. 빈번히 같이 발생하는 준식별자의 값을 가진 데이터는 개인정보 노출 위험도가 낮으므로, 이들을 제외한 데이터에 대해 가명 처리 또는 익명화를 수행하면, 좀 더 빠르게 가명처리 또는 익명화를 하면서도 데이터 품질이 크게 떨어지는 것을 막을 수 있다.

2.5 개인정보 보호 및 가명·익명화를 위한 이상 탐지 기술

데이터 마이닝의 이상 탐지 기술은 데이터 객체들이 주어졌을 때 다수의 객체와 그 성질이 크게 다른 객체를 찾아내는 것이다. 이를 위한 데이터 마이닝의 일반적 목표는 다수의 객체가 보이는 패턴을 모델링 한 뒤, 이 모델이 묘사하는 패턴에 크게 어긋나는 객체를 찾는 것이다. 전통적으로 다수의 객체가 보이는 통계적 분포를 모델로 사용하는 방법, 다수의 객체와 유사도가 현저히 낮은 객체를 찾는 방법 등이 사용됐으나, 최근엔 오토인코더(Autoencoder)와 같은 딥러닝 기술을 사용해 다수 객체가 보이는 패턴을 좀 더 정확하고 정교하게 모델링하는 방법들이 연구되고 있다. 대부분 이상 탐지 기술은 각 객체에 대해 이상 점수(Anomaly score)를 매긴 후, 이 점수가 사용자가 지정한 점수를 넘는 객체를 이상 객체로 탐지한다[1].

이러한 이상 탐지 기술은 개인정보 보호 및 가명·익명화에 다음과 같이 활용될 수 있다. 첫째, 이상 탐지 기술로 데이터에 포함된 각 레코드의 이상 점수를 매긴 후, 이 점수가 높은 레코드를 우선적으로 가명처리하거나 익명화한다. 이상 점수가 높은 레코드는 다른 레코드에 비해 그와 동일하거나 유사한 값이 적게 나타나는 레코드를 의미하므로, 이들은 개인정보 노출 위험도가 높다고 할 수 있다. 따라서 이상 점수가 높은 레코드를 우선적으로 가명 처리하거나 익명화함으로써, 데이터의 품질을 크게 떨어트리지 않고 개인정보 노출 위험도를 효과적으로 낮출 수 있다. 둘째, 가명 처리 또는 익명화 전 데이터와 가명 처리 또는 익명화 후 데이터의 개인정보 노출 위험도를 비교할 때, 이상 탐지 기술을 사용할 수 있다. 이 경우, 가명 처리 또는 익명화 전 데이터에 대한 이상 점수 분포와 가명 처리 또는 익명화 후 데이터에 대한 이상 점수 분포를 비교한다. 가명 처리 또는 익명화 작업의 효과가 높을수록 이상 점수가 높은 객체들의 숫자들이 이상 점수의 전체 분포에서 크게 줄어들어야 한다. 이것은 개인정보 노출 위험도가 높은 데이터의 준식별자 값들이 가명 처리 또는 익명화 작업에 의해 다른 데이터의 값들과 유사해져야 하기 때문이다. 즉, 이상 점수의 분포가 [그림 3]과 같은 형태로 바뀌는 것이 바람직하다고 할 수 있다. 따라서 이상점수 분포의 형태를 비교해 가명 처리 또는 익명화의 효과를 확인해 볼 수 있다.



[그림 3] 가명 처리 또는 익명화 전과 후 이상 점수 분포 비교[8]

3. 맺음말

최근 들어 다양한 분야에서 데이터가 크게 증가하면서, 그로부터 숨겨진 패턴, 정보, 지식을 추출하는 데이터 마이닝 수요가 꾸준히 증가하고 있다. 하지만 지금까지 데이터 마이닝 기술 연구는 주로 개인의 패턴 또는 행동을 예측하거나 선호도를 파악하는 등 개인정보의 '보호'보단 개인정보를 '파악'하는 방향으로 진행됐다. 따라서 이번 원고에선 지금까지 개발되트리지 않으면서도 데이터를 효과적, 효율적으로 가명 처리하거나 익명화하는 데 사용될 수 있다. 이번 원고를 통해 데이터 마이닝 기술이 개인정보보호 연구에 적극적으로 활용되기를 기대한다.

※ 본 논문은 2022년 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00634).

[참고문헌]

[1] Pang-Ning Tan, Michael Steinbach, Anuj Karapatne, and Vipin Kumar, "Introduction to Data Ming," 2nd Edition, Pearson, 2019.

[2] 이기용, "빅데이터 활용을 위한 개인정보 보호 및 관리기술 동향," 한국통신학회지, vol. 37, no. 1, pp. 32-39, 2019.

[3] Shi Dong, Ping Wang, and Khushnood Abbas. "A survey on deep learning and its applications." Computer Science Review, vol. 40, pp. 100379, 2021.

[4] 김종선, 이혁기, 정기정, 정연돈, "데이터 익명화 - 개념 이해 및 최신 기술 동향," 휴먼싸이언스, 2018.

[5] 가명정보 처리 가이드라인, https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=9900

[6] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, S. Yu Philip, and Lifang He, "Deep clustering: A comprehensive survey," IEEE Transactions on Neural Networks and Learning Systems, 2024.

[7] Abdul Majeed, Safiullah Khan, and Seong Oun Hwang, "Toward Privacy Preservation Using Clustering Based Anonymization: Recent Advances and Future Research Outlook", IEEE Access, vol. 10, pp. 53066-53097, 2022.

[8] Si Liu, Risheek Garrepalli, Dan Hendricks, Alan Fern, Debashis Mondal, and Thomas G. Dietterich, "PAC Guarantees and Effective Algorithms for Detecting Novel Categories," Journal of Machine Learning Research, vol. 23, no. 44, pp. 1-47, 2022.

※ 출처: TTA 저널 제214호