Journal of Digital Convergence
Vol. 16. No. 10, pp. 445-452, 2018

ISSN 1738-1916
https://doi.org/10.14400/JDC.2018.16.10.445

프라이버시 보존형 데이터 마이닝 방법 및 척도 분석

홍은주[1], 홍도원[2], 서창호[2*]
[1]공주대학교 융합과학과 박사과정, [2]공주대학교 응용수학과 교수

Privacy Preserving Data Mining Methods and Metrics Analysis

Eun-Ju Hong[1], Do-won Hong[2], Chang-Ho Seo[2*]
[1]Ph.D Course, Dept. of Convergence Science, Kongju National University
[2]Professor, Dept. of Applied Mathematics, Kongju National University

요 약 생활의 모든 것들이 데이터화 되어가고 있는 세상에서 데이터의 양은 기하급수적으로 증가하고 있다. 이러한 데이터는 수집 및 분석을 통하여 새로운 데이터로 가공되어진다. 새로운 데이터는 병원, 금융, 기업 등 여러 분야에서 다양한 용도로 사용되고 있다. 그러나 기존의 데이터에는 개인들의 민감한 정보가 포함되어 있기 때문에 수집 및 분석과정에서 개인의 프라이버시 노출 우려가 있다. 해결 방안으로 프라이버시 보존형 데이터 마이닝(PPDM)기술이 있다. PPDM은 프라이버시를 보존하면서 동시에 데이터로부터 유용한 정보를 추출하는 방법이다. 본 논문에서는 PPDM을 조사하고 데이터의 프라이버시와 유틸리티를 평가하기 위한 다양한 측정방법을 분석한다.

주제어 : 프라이버시, 데이터 마이닝, 프라이버시 보존형 데이터 마이닝, 척도, 유틸리티

Abstract In a world where everything in life is being digitized, the amount of data is increasing exponentially. These data are processed into new data through collection and analysis. New data is used for a variety of purposes in hospitals, finance, and businesses. However, since existing data contains sensitive information of individuals, there is a fear of personal privacy exposure during collection and analysis. As a solution, there is privacy-preserving data mining (PPDM) technology. PPDM is a method of extracting useful information from data while preserving privacy. In this paper, we investigate PPDM and analyze various measures for evaluating the privacy and utility of data.

Key Words : Privacy, Data Mining, Privacy-Preserving Data Mining, Metric, Utility

1. 서론

4차 산업혁명 시대의 중심에는 데이터가 있다. 생활의 모든 것들이 데이터화 되어가는 세상에서 정보는 중요해지고 있다. 나날이 증가하는 데이터의 수집 및 분석을 통해 숨겨져 있는 유용한 정보를 발견할 수 있다. 유용한 정보는 병원의 의료, 금융 등 많은 분야에서 다양하게 사용되고 있다. 데이터를 활용하는 것은 좋지만 수집된 데이터에는 개인의 민감한 정보가 포함되어 있어 프라이버시 노출의 우려가 있다. 프라이버시는 누구나 그 개념은 알고 있지만 표준적인 정의는 없다. 본 논문에서의 정보 프라이버시를 데이터 수집 및 분석시 침해될 수 있는 정보 유출이라고 정의한다. 프라이버시의 침해를 막기 위하여 즉, 프라이버시를 보호하기 위해 원본 데이터를 수정하여 프라이버시를 보존하는 방법이 개발되었다. 원본 데이터를 수정하기 때문에 정보의 유틸리티가 떨어지고 데이터 마이닝을 통한 정보 추출이 부정확하거나 또는 수정된 데이터 자체가 쓸모 없어질 수 있다. 특정 레벨의

* The work was supported by National Research Foundation of Korea(NRF) grant funded by the Korea Government (2016R1A4A1011761, 2016R1D1A1B03931071).
* Corresponding Author : Changho Seo(chseo@.kongju.ac.kr)
Received August 7, 2018 Revised September 28, 2018
Accepted October 20, 2018 Published October 28, 2018

프라이버시를 보장하면서 동시에 데이터의 유틸리티를 극대화하는 기술로 PPDM (Privacy-Preserving Data Mining)이 있다. PPDM은 최근 몇 년 동안 연구자들 사이에서 광범위한 주목을 받았다. 서로 다른 가정과 조건하에서 프라이버시에 대한 기술을 개발했다. PPDM은 프라이버시 레벨, 데이터 유틸리티, 데이터 복잡성 등 기법을 비교하고 평가하는 척도에 초점을 두어 많은 분야에 효과적으로 적용할 수 있다. 본 논문에서는 PPDM의 방법 및 프라이버시와 데이터 유틸리티를 평가하기 위한 여러 가지 척도에 대하여 분석한다[1-6].

2. 기본적인 데이터 마이닝 기술

많은 데이터가 수집되고 있다. 수집된 데이터의 분석을 통해 의료, 금융, 기업 등 여러 분야에서 많은 이익을 얻어낼 수 있다. 데이터에서 정보를 발견하는 것과 데이터 마이닝은 비슷한 맥락이라고 할 수 있다. KDD(Knowledge Discovery from Data)는 데이터 클리닝, 데이터 통합, 데이터 수집, 데이터 변형, 데이터 마이닝, 패턴 평가, 지식 표현의 여러 단계로 구성되어 있다[7]. 본 절에서는 데이터 마이닝 단계의 고전적인 데이터 마이닝 기술에 대하여 살펴본다.

데이터 마이닝이란 빅데이터로부터 패턴을 추출하는 단계이다. 패턴이란 데이터의 부분집합을 설명하는 표현식 또는 부분집합에 적용할 수 있는 모델로 정의된다. 데이터 마이닝 방법은 패턴 발견 및 추출이 목적이기 때문에 패턴인식 기술이 종종 사용된다[8]. 게다가 패턴인식과 머신러닝은 두 가지 측면에서 동일한 분야에 있다고 볼 수 있다. 데이터 마이닝의 주요 목적은 데이터로부터 예측 모델을 형성하는 것이다. 모델은 지도학습과 비지도학습을 사용하여 형성된다. 지도학습기법은 트레이닝 집합이 이미 레이블 되어있고 테스트집합에서 입력데이터와 출력데이터 사이의 관계가 형성되어 있어 기계가 데이터를 구별하는 방법을 배우고 모델을 구성하게 된다. 비지도학습은 레이블이 없기 때문에 데이터에서 관계를 찾으려고 하고 트레이닝집합이 사용되지 않는다. 연관규칙마이닝, 분류, 클러스터링은 데이터 마이닝에서 가장 일반적인 방법들 이며 첫 번째와 두 번째는 지도학습이고 세 번째는 비지도학습이다[6,9].

2.1 연관규칙 마이닝

연관규칙 마이닝 알고리즘은 데이터 집합의 변수들 사이에서 서로 관련 있는 관계를 발견하기 위해 설계되었다. 연관성의 형태는 조건과 결과로 구성되어 있다. 연관규칙은 발생 확률을 가지며 조건을 만족하면 결과가 발생할 확률이 있다. 연관규칙은 다음과 같이 공식화할 수 있다[9].

$I = I_1, \ldots, I_m$ 은 아이템으로 불리는 바이너리 속성의 집합이고 D는 트랜젝션 T의 데이터베이스이다. 각각 트랜잭션 T는 공집합이 아닌 아이템 집합이고, $T \subseteq I$ 이다. A와 B가 공집합이 아니며, I의 부분집합이라고 하자. 즉 $A \subset I$이고, $B \subset I$이며 $A, B \neq \varnothing$ 이다. $A \Rightarrow B$ 형식을 연관규칙으로 나타낸다. 연관규칙은 최소 지지도(support)와 최소 신뢰도(confidence)로 구성되어 있다. 규칙의 지지도는 D에서 A와 B가 동시에 발생한 확률이다.

$$\text{support}(A \Rightarrow B) = P(A \cap B)$$

규칙의 신뢰도는 D에서 A가 발생한 후에 B가 발생한 확률이다.

$$\text{confidence}(A \Rightarrow B) = P(B \mid A)$$

지지도와 신뢰도 척도를 사용하여 연관규칙 마이닝은 다음 두 단계를 실행한다[9].

1) D에서 모든 아이템 집합에서 주어진 최소 지지도의 임계값보다 크거나 같은 아이템 집합을 찾는다.
2) 1에서 찾은 아이템 집합으로부터 미리 주어진 최소 신뢰도의 임계값 만족하는 아이템 집합을 찾는다.

2.2 분류

분류는 지도학습문제이다. 분류기는 알려지지 않은 데이터에 클래스 라벨을 식별해줄 수 있다. 다시 말해서 분류기는 트레이닝 집합으로부터 알려지지 않은 데이터를 분류할 수 있다[9]. 분류에는 두 단계의 접근 문제가 있다. 트레이닝과 분류 단계이다. 먼저 클래스를 잘 분류하는 함수를 정의하는 것이다. 직관적으로 함수의 매핑은 수학적인 계산, 분류규칙, 결정트리에 의해 나타낼 수 있다[9].

매핑함수를 가지면 분류단계에 속성을 분류할 수 있다. 분류기를 평가하기 위해서 정확하게 분류한 확률인 정확도를 계산한다. 트레이닝 집합은 사용될 수 없으므로 최적의 정확도의 추정하기 위해 테스트 집합을 대신

사용한다.

2.3 군집

군집화 또는 군집분석은 데이터의 그룹화의 과정이다. 각각의 군집들은 라벨이 없는 클래스로 이루어지고 군집화는 자동 분류로 나타내어진다. 군집은 비지도학습이기 때문에 데이터에서 알려지지 않은 새로운 관계를 발견할 수 있다. 군집은 다음과 같은 성질을 기반으로 한다[9].

- 분할기준 : 군집은 계층적으로 생성되거나 모든 군집이 동일한 레벨 일 수 있다.
- 분리 : 군집은 겹치거나 혹은 겹쳐지지 않을 수 있고, 겹치는 경우 객체들은 많은 군집에 속할 수 있는 반면 겹쳐지지 않는 경우 군집은 상호 배타적이다.
- 유사도 측정 : 객체들 사이를 나타내는 척도는 거리 기반 또는 연결 기반으로 측정한다.
- 군집 공간 : 전체 데이터 공간 또는 정확성을 위해 불필요한 속성을 제거하여 차원을 줄인 부분 공간에서 군집을 찾을 수 있다.

3. 프라이버시와 데이터마이닝

데이터 수집과 데이터 마이닝 기술은 여러 응용에 적용된다. 이러한 기술을 다루다 보면 개인의 민감한 데이터를 공개할 때가 있다. 즉 개인의 정보 노출에 대한 우려를 야기한다. PPDM 기술은 민감한 정보의 노출을 막으면서 동시에 큰 데이터에서 정보를 추출할 수 있게 개발됐다[1,5,6,10,11]. PPDM의 주요 과정은 프라이버시 보존을 위해 오리지널 데이터의 일부를 제거하거나 수정한다. 데이터의 유틸리티와 프라이버시 레벨 사이의 균형 유지는 잘 알려져 있는 사실이다. PPDM 방법은 효과적인 마이닝을 통하여 데이터의 유틸리티는 최대화하면서 적절한 프라이버시 레벨을 보장할 수 있게 설계되었다. PPDM은 다음과 같이 데이터 수집 단계, 데이터 공개 단계, 마이닝 결과 단계로 나누어 서술할 수 있다.

3.1 데이터 수집시 프라이버시

데이터를 수집하는 개체를 신뢰할 수 없다고 가정한다. 데이터의 프라이버시를 보존하기 위하여 원본 데이터에 랜덤 값을 추가하는 데이터 변환이 필요하다. 프라이버시 노출을 막기 위해서 원본 데이터는 저장할 수 없고 변형된 과정을 사용한다. 결과적으로 랜덤화는 개개인에 각각의 값으로 수행된다. 랜덤화 방법에는 가장 많이 알려진 노이즈를 추가하는 방법이 사용된다[1,12]. 데이터 마이닝 알고리즘을 사용하여 원본 데이터의 분포를 재 복원할 수 있지만 여전히 원본의 값은 알 수 없다. 데이터 마이닝의 랜덤화 단계는 데이터 수집시 랜덤화, 분포 재구성, 재구성된 데이터의 마이닝으로 이루어진다. 데이터 수집가는 원본 데이터의 분포를 추정할 수 있다. 즉 노이즈 분포를 가지고 분포 재구성에 사용된다. Table 1은 노이즈 추가 방식에 대한 설명 및 장단점을 정리하였다.

Table 1. Randomization Method

Randomization Method	Additive Noise	Multiplicative Noise
Description	Data is randomized with a known statistical distribution.	
	adding noise	multiplying noise
Advantage	Performs independently for each captured value. Preserves statistical properties after reconstruction of the original distribution.	
		More effective than additive noise at preserving privacy, since the reconstruction is more difficult.
Disadvantage	Masking extreme values require great quantities of noise, severely degrading data utility.	

3.2 데이터 공개시 프라이버시

데이터 유저는 데이터가 수집되거나 공개될 때 민감한 정보의 노출 없이 데이터가 분석되길 바란다. 따라서 데이터 공개 시 프라이버시 보존은 수집된 데이터를 공개하기 전에 레코드를 익명화하는 것이다. PPDM에서 데이터 공개는 PPDP(Privacy-Preserving Data Publishing)라고 알려져 있다. PPDP는 데이터에서 유저를 바로 식별할 수 있는 명백한 속성을 제거하고, 준 식별자와 민감한 속성은 제거하지 않는다. 준 식별자란 민감한 속성도 아니고 유저의 명백한 속성도 아니지만 외부 다른 데이터와 조합했을 때 레코드의 유저를 식별할 수 있는 가능성이 있는 속성이다. 외부 데이터와 조합하여 유저를 식별할 수 있는 공격을 속성연결 공격이라 한다. 민감한 속성은 개인의 구체적인 속성을 나타내고 이는 비공개되어야 하고 개인과 연결되어서도 안된다. 2002년 Sweeney 논문에 따르면 미국인의 87%는 고유한 특징을 가지고

있어 외부 데이터와 조합했을 때 개인이 재 식별화됨을 보였다[13]. 이 논문에서 시사하는 바는 데이터 공개 시 명백한 속성을 제거한다고 해서 개인을 식별할 수 없는 것은 아니라는 것을 보여준다. 따라서 Aggarwal은 익명화 알고리즘의 중요 핵심을 민감 속성이 아닌 준 식별자로 꼽았다[1,13-15].

데이터베이스에서 레코드의 익명화는 다양한 프라이버시 모델로 구성되어 있다. 프라이버시 모델은 레코드 소유자의 식별을 보존하기 위해 하나 또는 여러 연산을 조합하여 적용한다.

- 일반화(generalization): 더 일반화된 값으로 대체하는 것으로 숫자 데이터의 경우 구간으로 정의하고, 카테고리컬 속성은 트리의 계층적 구조에 의해 대체된다.
- 범주화(suppression): 정보 노출을 막기 위해 일부 속성값을 제거한다. 이 연산은 또한 행이나 열을 수정할 수 있다.
- 분리화(Anatomization): 준 식별자와 민감 속성의 연결을 더 어렵게 하기 위해서 두 속성을 분리한 테이블에서 준 식별자와 민감 속성의 연관성을 막는다.
- 섭동(perturbation): 동일한 확률적 정보를 가지는 변형된 값에 대하여 오리지널 데이터를 대체한다. 예로 랜덤화 방법이 있다. 데이터 교환과 변형 데이터 일반화 방법도 포함된다. 데이터 교환은 민감한 속성을 레코드 연결을 막기 위해 다른 데이터 집합과 교환하는 것이고 변형 데이터 일반화는 통계적 모델로 원본 데이터를 변형하고 모델로부터 변형 값을 얻게 된다.

이런 연산을 기반으로 하는 프라이버시 모델들은 다음과 같다. 이 중에서 가장 잘 알려진 프라이버시 모델은 k-익명성 모델로 Samarati와 Sweeney가 제안하였다[13]. 외부 데이터를 이용하여 공격할 수 있는 레코드 연결 공격을 막기 위한 모델이다. 이 모델의 핵심 개념은 임의의 데이터베이스에서 준 식별자에 대한 하나의 레코드와 나머지 k-1개의 레코드들을 서로 구별할 수 없다는 것이다. 이러한 데이터 집합을 k-익명성이라고 한다. 구별할 수 없는 k레코드의 집합을 동치류라고 한다[1]. k-익명성 모델에서는 k값이 프라이버시의 척도가 된다. k값이 높을수록 레코드가 재 식별되기 어렵지만, k값이 증가하면 높은 수준의 일반화가 발생하기 때문에 데이터의 유틸리티가 감소한다. 이론적으로 k-익명성에서 재 식별될 레코드의 확률이 1/k이지만, 각 동치류의 민감한 값이 모두 동일할 경우 100%로 개인의 민감한 값이 노출된다. 이 공격을 속성공격이라고 한다. 속성공격은 민감한 값의 다양성이 부족할 때 발생하는 공격이다. 민감한 값의 다양성을 추가하여 k-익명성의 문제점을 해결하는 l-다양성 모델이 제안되었다[14]. k-익명성에서 민감한 속성들의 다양성을 1가지로 하여 위의 문제를 해결할 수 있다. l-다양성의 l값을 좀 더 구체적으로 나타내기 위해서 엔트로피의 개념을 이용한다. l-다양성은 동치류 내에서 민감한 값의 다양성을 증가시켰지만, 전체적인 분포는 고려하지 않았다. 민감한 값이 왜곡되어 분포되어 있을 때 프라이버시 침해를 유발할 수 있다. 이 공격을 비대칭 공격이라고 한다. 또한 l-다양성은 민감한 값들 사이에 유사도가 있을 때 유사도 공격이 발생할 수 있다. 민감한 값을 알 수 없지만 값들 사이에 유사성이 존재하여 공격자는 충분히 민감한 값을 예상할 수 있다. 위의 공격들을 막기 위해 Li는 t-근접성 모델을 제안하였다[15]. 동치류의 민감한 값의 분포와 원본의 분포를 임계값 t로 제한한다. t-근접성의 원리는 거리를 측정하는데 사용되는 다양한 거리함수에 의존한다. 일반적인 거리함수는 variational distance, *Kullback-Leibler* distance, Earth Mover's distance 이다.

k-익명성, l-다양성, t-근접성 프라이버시 모델은 모든 레코드를 같은 조건으로 일반화하여 프라이버시를 보호한다[1,13-15]. 그러나 이상적으로 데이터 소유자는 자신의 레코드에 필요한 프라이버시 레벨을 지정할 수 있다. 위의 프라이버시 모델은 과다한 프라이버시 제어를 통하여 데이터 마이닝 결과의 유용성을 감소시킨다. 과다한 프라이버시 제어를 해결하기 위해 xiao와 Tao는 각각의 개인마다 프라이버시 레벨을 정의하는 개인 프라이버시 개념을 제시하였다[16]. 목적은 개인별 프라이버시를 설정하면서 최대한의 유틸리티를 보존하는 것이다. 개인 프라이버시는 민감한 값의 트리를 생성하여 레코드 소유자가 보호 노드를 정의하게 된다.

대부분의 개인 정보보호 모델은 레코드 소유자의 신원을 보호하거나 익명화된 레코드에서 민감한 값의 추론을 보호하려고 한다. 대부분 모델들은 레코드의 존재가 소유자의 프라이버시에 어떤 영향을 미치는지는 측정하지 않는다. 즉, 소유자의 존재로는 정보가 유출되지 않는다고 생각한다. Dwork은 개인의 레코드 존재 여부와 개

인의 프라이버시 공개의 차이를 측정하기 위해 차분 프라이버시라는 개념을 제시했다[17]. 단일 레코드가 데이터세트에 대한 분석 결과에 영향을 미치지 않도록 보장하는 프라이버시 모델이다. 강력하고 공식적인 프라이버시 개념이지만 차분 프라이버시에는 엡실론을 설정하는 것과 같은 몇 가지 제한점이 있다. 현재 차분 프라이버시에 대한 많은 연구가 진행되고 있다.

3.3 데이터 마이닝 결과 프라이버시

데이터 마이닝 알고리즘의 결과는 원본 데이터에 직접적으로 접근하지 않아도 극단적으로 드러난다. 공격자는 쿼리를 통해 원본 데이터에 대한 중요한 정보를 추측할 수 있다.

- 연관규칙숨기기(Association Rule Hiding) - 연관규칙 데이터 마이닝에서 일부 규칙은 개인 또는 개인의 집합에 대한 개인 정보를 명확하게 공개할 수 있다. 연관규칙 숨기기는 민감하지 않은 규칙은 모두 밝혀내고 민감한 규칙은 숨기는 프라이버시 보호 기술이다. 최적이 아닌 솔루션은 민감한 규칙이 숨겨진 방식으로 데이터 항목을 혼란시키지만, 처리 과정에서 상당수의 민감하지 않은 규칙을 잘못 숨길 수 있다[1,18].
- 분류기 효율성 감소(Downgrading Classifier Effectiveness) - 분류기는 공격자에게 사용자의 정보를 누출 시킬 수 있다. 예로 레코드가 트레이닝 데이터 집합에 있는지를 공격자가 판단하는 멤버십 추론공격이 있다. 분류기의 프라이버시를 보존하기 위해서 분류기의 정확도를 낮추는 기술이 사용된다. 일부 규칙을 기반으로 하는 분류기는 연관규칙 마이닝 방법을 사용하기 때문에 연관 규칙 숨기기 방법이 적용되어 분류기의 효율성을 낮춘다[10].
- 질의 감사 및 추론 제어(Query Auditing and Inference Control) - 엔티티가 원본 데이터 집합에 대한 접근을 제공하여 데이터에 대한 통계적 질의만 허용할 수 있다. 사용자는 개별 또는 그룹 레코드가 아닌 데이터 집합의 집계 데이터만 질의 할 수 있다[10]. 그러나 일부 질의는 여전히 개인정보를 나타낸다. 질의 감사에는 두 가지 접근방식이 있다. 원본데이터 또는 질의의 출력값이 변형되어 있는 경우인 질의 추론제어와 하나 또는 그 이상의 질의가 일련의 질의로부터 거부되는 질의감사가 있다. 질의 감사 및 추론제어 기술은 상황에 따라 통계 데이터베이스 보안에서 광범위하게 연구되고 있다.

위의 방법들은 어떤 애플리케이션을 사용하는지에 따라 영향을 받는다. 애플리케이션 자체가 다운그레이드 되거나 데이터에 대한 접근이 제한되어 있기 때문이다. 따라서 프라이버시와 유틸리티 사이의 절충이 항상 존재한다.

4. 프라이버시와 유틸리티 측정

프라이버시에 대한 표준 정의가 없기 때문에 프라이버시를 정량화하는 것은 어려운 일이다. PPDM 관점에서 몇 가지 척도가 제안되었다. 프라이버시 관점에서 데이터 공개 시 얼마나 안전한지를 측정하는 프라이버시 레벨 측정과 유틸리티 관점에서 공개된 데이터에서 정보의 손실량을 측정하는 척도와 마지막으로 다양한 기술에 대하여 효율성과 확장성을 측정하는 복잡성 척도가 있다[6].

4.1 프라이버시 측정

PPDM의 목적은 데이터 유용성을 최대화하면서 일정 수준의 프라이버시를 보존하는 것이다. 프라이버시 레벨은 가능한 프라이버시 침해로부터 얼마나 안전한지를 나타낸다. 프라이버시 레벨 측정은 데이터 프라이버시 측정과 결과에 대한 프라이버시 측정으로 분류될 수 있다. 데이터 프라이버시 측정은 프라이버시 보존 방법을 적용한 결과 변형된 데이터로부터 원본 데이터의 민감한 정보를 추론할 수 있는 방법을 측정하고, 결과 프라이버시 척도는 데이터 마이닝 결과가 원본 데이터에 대한 정보를 얼마나 주는지 측정한다.

Confidence Level은 랜덤화 데이터에 대해 원본 데이터를 얼마나 잘 추정했는지 측정하는 척도이다[19]. 이 척도는 원본 데이터의 분포를 고려하지 않는 문제점이 있다. 따라서 문제 해결을 위해 Average Conditional Entropy가 제안되었다. 곱셈 노이즈 랜덤화에서는 프라이버시를 측정하기 위해 원본과 섭동된 데이터 사이의 Variance을 사용한다. Variance의 비율에 따라 원본 값을 얼마나 정확히 추정하였는지 나타낸다.

데이터 공개 프라이버시에서 k-익명성, l-다양성, t-근접성 및 ϵ-DP 모델이 제시되었다[13-15,17]. 이 모델

들 각각의 변수 k,l,t,ϵ가 프라이버시 레벨에 대해 제한을 갖는다. 그러나 이러한 측정 기준은 해당하는 모델에만 특화되어 있어 활용 범위가 좁다.

결과 프라이버시 척도는 프라이버시와 정보 발견 사이의 균형을 측정하는 HF(Hidden failure)가 있다. HF는 프라이버시 보존 방법을 가지고 숨겨진 민감한 패턴과 원본 데이터에서 발견되는 민감한 패턴 사이의 비율로 정의된다[20].

$$HF = \frac{*R_P(D')}{*R_P(D)}$$

D'과 D는 변형된 데이터 집합과 원본 데이터집합이고, $*R_P(\,)$는 민감한 패턴의 개수이다. HF=0이면 모든 민감한 패턴이 숨겨지면서 더 많은 민감하지 않은 정보가 손실될 수 있다. 이 척도는 패턴 인식 데이터 마이닝 기술에서 사용될 수 있다. HF는 프라이버시 레벨만 측정하고 손실된 정보의 양에 대해서는 측정하지 않는다. Table 2는 프라이버시 레벨 척도를 정리하였다.

Table 2. Privacy Level Metrics

Data Metrics	Results Metrics
- Confidence Level - Average Conditional Entropy - Variance - Privacy Model Specific(K,L,T)	- Hidden Failure

4.2 유틸리티 측정

프라이버시를 보존하는 기술은 데이터의 유틸리티를 떨어트린다. 데이터 유틸리티 척도는 데이터의 손실을 계산하려고 한다[2]. 일반적인 측정으로 원본 데이터와 변환된 데이터의 결과를 비교한다. 데이터 유틸리티를 평가할 때 다음과 같은 중요한 매개변수가 측정된다. 원본 데이터와 변형된 데이터가 얼마나 가까운지 측정하는 정확도가 있고, 프라이버시 레벨 척도와 마찬가지로 데이터 유틸리티 관점과 데이터 마이닝 결과에 대한 유틸리티를 측정하는 관점이 있다. Table 3은 유틸리티 척도를 정리하였다. 익명화와 억제 기법에서 MD(Minimal Distortion)와 LM(Loss Metric) 및 ILOSS(Information Loss) 척도가 있다. MD는 값이 상위 값으로 생성될 때마다 카운터가 증가되는 간단한 방법이다. 값이 높을수록 데이 터가 더 일반화되고 결과적으로 많은 정보가 손실된다. LM과 ILOSS 측정은 분류트리의 원본 리프 노드의

Table 3. Data Quality Metrics

Data Metrics	Results Metrics
- Minimal Distortion - Loss Metric - Information Loss - Discernibility Metric	- Misses Cost - Artifactual Patterns - Misclassification Error

총 수를 고려하여 모든 레코드에 대한 평균 정보손실을 측정한다. ILOSS는 평균에 대하여 속성마다 다른 가중치를 적용한다. 동일한 클래스 알고리즘일 경우엔 DM으로 측정한다. 이 척도는 일반화로 인해 주어진 레코드와 동일한 레코드 수를 측정한다. 값이 더 높을수록 더 많은 정보 손실이 있다.

재구성 알고리즘의 정확도를 측정함으로써 유틸리티를 계산하는 방법이 있다. 재구성된 분포와 원본 분포를 비교하여 정보손실을 측정한다. 정보손실은 다음과 같이 정의된다.

$$I(f_X(x),\widehat{f_X}(x)) = \frac{1}{2}E[\int_{\Omega_X} |f_X(x) - \widehat{f_X}(x)|dx]$$

$f_X(x)$는 원본 밀도함수이고 $\hat{f}_X(x)$는 재구성된 밀도함수이다.

결과의 유틸리티를 측정하기 위한 측정 기준은 어떤 데이터 마이닝 기술을 사용했는지에 따라 달라진다. 기준은 변형된 데이터와 원본 데이터의 각각의 마이닝 결과 비교를 기반으로 한다. 패턴인식 알고리즘의 결과에서 데이터 유틸리티 손실을 측정하는 측정 기준은 MC(Misses Cost)와 AP(Artifactual Patterns)이다[20]. MC는 다음과 같이 잘못 숨겨진 패턴 수를 측정한다.

$$MC = \frac{**R_P(D) - **R_P(D')}{**R_P(D)}$$

$**R_P(X)$는 데이터베이스 X에서 발견된 비 제한적 패턴의 수를 의미한다.

AP는 원본 데이터 D에는 존재하지 않았지만 D'에서 새로 생성된 패턴을 측정한다. 다음 방정식은 AP 척도를 정의한다.

$$AP = \frac{|P'| - |P \cap P'|}{|P'|}$$

P와 P'은 D와 D'에 속하는 모든 패턴의 집합이다. $|*|$는 카디널리티를 나타낸다.

클러스터링 기술의 경우에는 ME(Misclassification Error)가 있다. ME는 왜곡된 데이터베이스에서 잘 분류

되지 않은 데이터 포인트의 백분율을 측정한다. 즉, 동일한 클러스터 안에서 원본 데이터 및 위생 데이터로 그룹화되지 않은 포인트 개수이다. ME는 다음과 같이 정의된다.

$$ME = \frac{1}{N}\sum_{i=1}^{k}(|Culster_i(D)| - |Cluster_i(D')|)$$

N은 데이터베이스의 총 포인트 개수이고, k는 클러스터의 개수, $|Cluster_i(X)|$는 데이터베이스 X안에 i번째 클러스터의 데이터 포인트의 개수이다.

분류와 클러스터에 대한 결과의 유틸리티를 평가하기 위한 추가적인 측정기준과 데이터 마이닝 결과의 유틸리티 측정을 위한 일반적인 정량적 접근법도 있다[2,20,21].

4.3 계산 복잡성 측정

PPDM 기술의 복잡성은 대부분 구현된 알고리즘의 효율성과 확장성에 관련이 있다. 효율성을 측정하기 위해 시공간과 같은 특정 리소스 사용에 대한 척도를 사용할 수 있다. 시간은 CPU 시간 또는 계산 비용으로 측정할 수 있고, 공간은 알고리즘을 실행하는데 필요한 메모리양을 정량화할 수 있다. 분산 컴퓨팅에서 시간이나 교환된 메시지 수와 통신비용 등을 측정할 수 있다[21].

확장성은 데이터양이 점점 증가할 때 기술이 얼마나 잘 수행되는지를 나타낸다. 분산 계산에서는 입력이 늘어나면 통신량이 급격하게 증가할 수 있다. 따라서 PPDM 알고리즘은 확장 가능한 방식으로 설계되어야 한다. 확장성은 시스템마다 서로 다른 로드를 주기 때문에 경험적으로 평가할 수 있다. PPDM알고리즘의 확장가능성을 테스트 하려면 입력데이터에서 몇 가지 실험을 수행하여 효율성 손실을 측정한 후 이에 맞게 확장성을 측정할 수 있다.

5. 결론

병원, 은행, 기업 외 여러 단체에서는 기존에 제공된 서비스를 개선하기 위해 끊임없이 데이터를 수집하고 활용한다. 서비스를 위해 필요한 데이터들은 민감성을 가지고 있다. 따라서 프라이버시 문제가 발생된다. PPDM 방법은 개인의 프라이버시를 보존하면서 데이터에서 정보를 추출할 수 있도록 하는 기술이다. PPDM 방법은 데이터의 수집, 게시, 배포 및 출력과 같은 데이터 활용 단계에 따라 다르게 설명된다. 이러한 기술은 데이터의 프라이버시 및 유틸리티를 평가하는 척도를 분석함으로써 해결된다. 추후 PPDM에서 수행해야 할 연구는 다음과 같다. 첫째, 프라이버시의 레벨을 제어하여 프라이버시 보호의 개념을 구현하는 시스템이 필요하다. 둘째, 개인 레코드 소유자의 권리와 책임을 위한 프라이버시 보존 툴이 필요하다. 이 문제에 대해서는 이전 관련된 연구로 personalized privacy가 좋은 출발을 제공하였지만, 구현 측면에서 효율성을 얻지 못하여 다음 연구가 이루어지지 않았다[16]. 마지막으로 데이터의 양이 증가함에 따라 공격자는 충분한 배경지식을 가질 수 있다. 따라서 공격자의 배경지식을 모델링 하여 공격을 방어할 수 있는 새로운 프라이버시 보호 메커니즘의 연구가 필요하다.

REFERENCES

[1] C. C. Aggarwal. (2015) Data Mining: The Textbook. *New York, NY, USA:Springer.*

[2] S. Fletcher & M. Z. Islam. (2015) Measuring information quality for privacy preserving data mining. *Int. J. Comput. Theory Eng, 7(1),* 2128.

[3] Y. A. A. S. Aldeen, M. Salleh & M. A. Razzaque.(2015) A comprehensive review on privacy preserving data mining. *SpringerPlus, 4(1),* 694.

[4] S. Yu. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE Access, 4,* 2751-2763.

[5] A. Shah & R. Gulati. (2016) Privacy preserving data mining: Techniques, classication and implicationsA survey. *Int. J. Comput. Appl, 137(12),* 40-46.

[6] R. Mendes & J. P. Vilela.(2017) Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access, 5,* 10562-10582.

[7] U. Fayyad, G. Piatetsky-Shapiro & P. Smyth. (1996) From data mining to knowledge discovery in databases. *AI Mag, 17(3),* 3754.

[8] C. M. Bishop. (2006) Pattern Recognition and Machine Learning. *vol. 4. New York, NY, USA: Springer-Verla.*

[9] J. Han, M. Kamber & J. Pei. (2012) Data Mining: Concepts and Techniques. *Amsterdam, The Netherlands: Elsevier.*

[10] C. C. Aggarwal & P. S. Yu. (2008) A general survey of privacy-preserving data mining models and algorithms. in *Privacy-Preserving Data Mining. New York, NY, USA: Springer,* 1152.

[11] L. Xu, C. Jiang, J. Wang, J. Yuan & Y. Ren. Information security in big data: Privacy and data mining. *IEEE Access, 2,* 1149-1176.

[12] K. Liu, H. Kargupta & J. Ryan. (2006) Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans. Knowl. Data Eng, 18(1),* 92-106.

[13] L. Sweeney. (2002) K-anonymity: A model for protecting privacy. *Int. J.Uncertainty, Fuzziness Knowl.-Based Syst, 10(5),* 557-570.

[14] A. Machanavajjhala, D. Kifer, J. Gehrke & M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discovery Data, 1(1),* 3.

[15] N. Li, T. Li & S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. in *Proc. IEEE 23rd Int. Conf. Data Eng, (ICDE),* Apr, 106-115.

[16] X. Xiao & Y. Tao. (2006) Personalized privacy preservation. in *Proc, VLDB,* 139-150.

[17] C. Dwork. (2006) Differential privacy. in *Automata, Languages and Programming, 4052.* Venice, Italy: Springer-Verlag, Jul. 1-12.

[18] V. S. Verykios. (2013) Association rule hiding methods. *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, 3(1),* 28-36.

[19] R. Agrawal & R. Srikant. (2000) Privacy-preserving data mining. *ACM SIGMOD Rec, 29(2),* 439-450.

[20] S. R. Oliveira & O. R. Zaiane. (2002) Privacy preserving frequent itemset mining. in *Proc. IEEE Int. Conf. Privacy, Secur. Data Mining, 14 Dec,* 43-54.

[21] E. Bertino, D. Lin & W. Jiang. (2008) A survey of quantication of privacy preserving data mining algorithms. in *Privacy-Preserving Data Mining.* New York, NY, USA: Springer, 183-205.

홍 은 주(Hong, Eun Ju) [정회원]

- 2013년 2월 : 공주대학교 응용수학과(이학사)
- 2015년 2월 : 공주대학교 융합과학과(이학석사)
- 2015년 2월 ~ 현재 : 공주대학교 융합과학과 박사과정
- 관심분야 : 데이터마이닝, 정보보호
- E-Mail : baby0708@kongju.ac.kr

홍 도 원(Hong, Do Won) [정회원]

- 1994년 2월 : 고려대학교 수학과(이학사)
- 2000년 2월 : 고려대학교 수학과(이학박사)
- 2000년 4월 ~ 2012년 2월 : 한국전자통신연구원 팀장, 책임연구원
- 2012년 2월 ~ 현재 : 공주대학교 응용수학과 교수
- 관심분야 : 암호기술, 프라이버시 보호기술 등
- E-Mail : dwhong@kongju.ac.kr

서 창 호(Seo, Chang Ho) [정회원]

- 1990년 2월 : 고려대학교 수학과(이학사)
- 1992년 2월 : 고려대학교 수학과(이학석사)
- 1996년 2월 : 고려대학교 수학과(이학박사)
- 2000년 3월 ~ 현재 : 공주대학교 응용수학과 교수
- 관심분야 : 암호알고리즘, PKI, 무선인터넷 보안 등
- E-Mail : chseo@kongju.ac.kr