원격 의료 서비스를 위한 EHR 데이터 비식별화 기법 제안

윤준호*, 김현성*,**
*경일대학교 사이버보안학과
**말라위대학교 수학과
e-mail: *yoonjunho62@gmail.com, **kim@kiu.ac.kr

Deidentification Method Proposal for EHR Data on Remote Healthcare Service

Junho Yoon*, Hyunsung Kim*,**
*Dept. of Cyber Security, Kyungil University
**Mathematical Science Dept., University of Malawi

요 약

최근 인공지능과 빅데이터 등 최첨단 기술이 빠른 속도로 의료 정보시스템에 도입됨에 따라 환자정보를 포함한 민감한 개인정보에 대한 사이버 공격이 급증하고 있다. 다양한 개인정보 비식별화에 대한 표준이 제안되었지만, 데이터의 범주에 따른 기법 적용에 대한 연구가 미비하다. 본 논문에서는 EHR 데이터를 위한 심근경색을 대상으로 하는 원격 의료 시스템을 위한 개인정보들에 대한 민감도를 4단계로 분류하고 이에 따른 비식별화 기법에 대해 제안한다. 본 논문에서 제안한 EHR 데이터에 대한 분류 및 비식별화 기법은 다양한 의료 정보 서비스를 위한 프라이버시 보호에 활용될 수 있다.

1. 서론

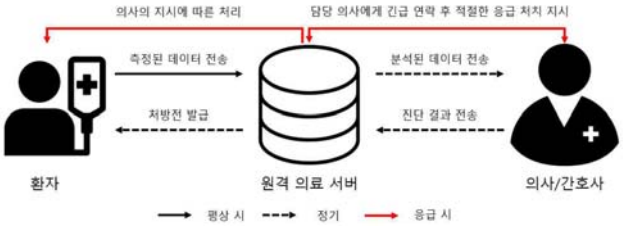
원격 의료는 의료기관이 없는 산간·오지 등이나 거동이 불편한 환자를 멀리 떨어져 있는 의사와 환자 사이에 통신수단을 활용해 진단과 처방이 이뤄지는 첨단의료 시스템이다[1-4]. 다양한 출처로부터 데이터를 결합하여 더 많은 새로운 정보를 도출할 수 있다는 것은 빅데이터의 대표적 활용이다. 하지만 빅데이터화와 다양한 시스템 간 다양한 데이터의 이동에 따라 다양한 보안 및 프라이버시 문제가 대두되고 있다[5-6].

대용량 데이터 기술은 데이터 보존 정책과 더불어 데이터 처리 과정에서 개인 식별 정보 및 민감한 데이터의 보호를 요구한다[7-10]. 다양한 데이터의 결합은 데이터 경제 환경에서 빅데이터 활용 가치를 극대화시킬 수 있는 수단이다. 하지만, 개인정보 처리자 간의 정보집합에 대한 데이터 결합으로 인한 개인정보 재식별의 위험이 항상 존재한다. 따라서 데이터 결합 시 더 높은 수준의 보안 및 프라이버시가 요구된다[11]. 이런 이유로 다양한 개인정보 보호에 대한 표준이 제시되고 있다.

본 논문에서는 EHR 데이터를 위한 심근경색을 대상으로 하는 원격 의료 시스템을 위한 개인정보들에 대한 민감도를 분류하고 이를 위한 구체적인 비식별화 기법에 대해 제안한다. 본 논문에서 제안한 EHR 데이터에 대한 분류 및 비식별화 기법에 대한 제안은 다양한 의료 정보 서비스를 위한 프라이버시 보호에 활용될 수 있을 것으로 기대한다.

2. EHR 데이터 개요

원격 의료 서비스는 원거리에서 건강 관리를 제공하기 위한 통신과 정보 기술의 결합 서비스이다. 그림 1은 본 논문에서 고려하는 원격 의료 서비스의 기본 개요를 보여준다. 이러한 서비스를 제공하기 위해서는 센서 노드가 배치된 환자와 원격 의료 서버, 그리고 의사/간호사로 구성되는 3가지 구성요소가 필요하다.



(그림 1) 원격 의료 시스템 개요도

시스템 구성요소들의 기본적인 개요는 다음과 같다.

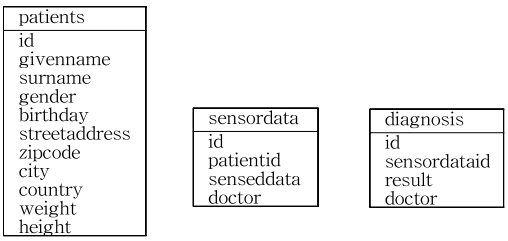
- 환자 : 데이터의 주체이며 센서노드를 통해 센싱된 바이오 데이터가 원격 의료 서버로 주기적으로 전송이 된다. 특히, 측정된 데이터에는 개인정보 등의 민감한 정보가 포함되어 있으므로 이러한 정보가 노출되지 않도록 각별한 주의가 필요하다.
- 원격 의료 서버 : 데이터 관리자이며 동시에 데이터 처리를 수행한다. 데이터 수집의 주체가 되고, 수집된 데이터에 따라 환자에게 처방전을 전달하고 응급 상황에 따른 적절한 처리를 담당한다. 특히, 환자의 센서 노드로부터 수집된 임상 데이터를 데이터베이스에 저장하는 역할을 주로 담당한다. 이렇게 수집

된 내용은 다양한 툴을 이용하여 의사와 환자 간에 필요한 다양한 건강 관리 정보를 제공한다. 특히, 서버에 저장되는 데이터베이스에 대한 프라이버시 제시가 필요하다. 이를 위해 다양한 데이터 보안 관련 표준에서 비식별화에 대해 권고한다.

- 의사/간호사 : 원격 의료 서버로부터 분석된 의료 데이터에 대한 분석과 처방전 작성을 수행한다. 서버로부터 수신한 데이터 중 잘못 분석된 데이터가 있는지 또는 추가적인 분석이 필요한지를 판단하며 응급 상황에 대비한 적절한 처방전을 제시한다.

프라이버시(Privacy, 개인정보) 데이터는 직접 또는 간접적으로 개인을 실별 가능한 데이터를 의미한다. 이러한 데이터에는 통신 참여자를 식별할 수 있는 식별자 또는 자신의 신체적, 정신적, 경제적, 사회적 또는 문화적 정체성과 관련된 하나 이상의 속성을 포함할 수 있다. 데이터 관리자는 개인정보 관련 데이터에 대해 처리 목적과 수단을 결정함으로써 데이터를 임의로 활용하는 것을 방지해야 한다. 데이터 관리자는 자신이 보유한 특정 개인의 프라이버시 데이터에 대한 처리 권한이 자신에게 유지되는 경우에만 다른 참여자가 데이터를 활용할 수 있도록 제공한다.

특히, 본 논문에서는 환자의 정보 데이터베이스인 EHR에 초점을 맞춘 개인정보의 비식별화에 대한 분류 및 방법을 제시한다. 다양한 EHR 구성에 대한 연구들이 진행되었지만 본 논문에서는 그림 2에 제시된 EHR 테이블을 기반으로 하는 데이터를 연구 대상으로 한다.



(그림 2) EHR 데이터의 구성

patients		sensordata	
id	1	id	1
givenname	Shannon	patientid	1
surname	French	senseddata	gfxdwmv.xml
gender	female	doctor	jansen
birthday	3/24/1953		
streetaddress	45 Iolaire Road	id	2
zipcode	NP4 2PT	patientid	1
city	NEW INN	senseddata	bszwrsl.xml
country	GB	doctor	jansen
bloodtype	A+		
weight	55.8	id	3
height	162	patientid	1
		senseddata	gsjadfv.xml
		doctor	jansen

diagnosis	
id	1
sensordataid	3
result	Epilepsy
doctor	jansen

(그림 3) EHR 데이터 예시[12]

EHR 관련 데이터들은 patients와 sensordata, 그리고 diagnosis 총 3개의 테이블에 저장된다. 이 테이블들은 id를 주요키(Primary Key)로 이용한다. 각 테이블의 속성은 다음과 같다.

- patients 테이블: 환자의 개인정보를 저장한다. 이 테이블은 이름, 성별, 지번, 국적 등의 데이터를 저장한다.
- sensordata 테이블: 한 환자로부터 센싱된 데이터의 이름과 주치의 이름을 저장한다. patientid 필드는 참조하고 있는 patients 테이블의 id 값을 저장한다. senseddata 필드는 센서 노드를 통해 전달받은 생체 정보를 xml 파일로 저장한다. doctor 필드는 해당 환자의 담당 의사의 이름을 저장한다.
- diagnosis 테이블: 생체 데이터에 대한 분석을 기반으로 한 의사의 진단 관련 데이터를 저장한다. sensordataid 필드는 참조하고 있는 sensordata 테이블의 id 값을 저장한다. result 필드에는 병명을, doctor 필드에는 담당 의사의 이름을 저장한다.

3. 비식별화 표준

개인정보 비식별 조치를 위해서는 ISO/IEC WD20889, NIST 비식별 처리 가이드라인, 익명화 프레임워크, 빅데이터 프라이버시 설계와 같이 다양한 표준들이 존재한다[8]. 본 논문에서는 이러한 표준들 중 개인정보 비식별 조치 가이드라인 표준을 고려한다[13-14]. 표 1은 개인정보 비식별 조치 가이드라인 표준에서 정의한 비식별 조치 기법이다.

<표 1> 개인정보 비식별 조치 가이드라인[14]

처리기법	개념	세부기술
가명처리 (Pseudonymization)	개인 식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법	휴리스틱 가명화, 암호화, 교환 방법
총계처리 (Aggregation)	통계값(전체 혹은 부분)을 적용하여 특정 개인을 식별할 수 없도록 함	총계처리, 부분총계, 라운딩, 재배열
데이터 삭제 (Data reduction)	개인 식별이 가능한 데이터 삭제 처리	식별자 삭제/부분삭제/전부삭제, 레코드 삭제
데이터 범주화 (Data suppression)	특정 정보를 해당 그룹의 대푯값으로 변환하거나 구간 값으로 변환하여 개인 식별을 방지	감추기, 랜덤 라운딩, 범위 방법, 제어 라운딩
데이터 마스킹 (Data masking)	데이터의 전부 또는 일부분을 대체값(공백, 노이즈 등)으로 변환	임의 잡음 추가, 공백과 대체

개인정보 비식별 조치 가이드라인은 5가지 분류와 17가지의 세분화를 제시한다. 본 절에서는 EHR 데이터를 중심으로 세부기술들의 특징을 제시하고 이에 대한 적정성을 평가한다.

가명처리는 개인정보를 다른 값으로 대체하는 방법이다. 데이터의 변형이 적지만 식별 가능한 고유 속성이 유지되는 장점이 있다. 가명처리는 모든 문자 데이터를 다음 형태로 비식별 처리할 수 있다.

- 휴리스틱 가명화: 식별자의 분포를 고려하거나 수집된 자료의 사전 분석을 하지 않고 모든 데이터를 동일한 방법으로 가공
- 암호화: 정보 가공 시 일정한 규칙의 암호시스템을 적용하여 암호화와 복호화가 가능하도록 함
- 교환 방법: 기존 데이터베이스의 레코드를 사전에 정해진 외부의 변수 값과 연계하여 교환

총계처리는 통계값을 적용하는 비식별 처리기법이다. 민감한 수치 정보에 대하여 비식별 처리가 가능하며 통계분석용 데이터

셋 작성에 유리하지만 정밀 분석이 어려우며 집계 수량이 적을 경우 추론에 의한 식별 가능성이 있다. 총계처리는 수치 데이터에 대한 비식별화 기법이다.

- 총계처리: 데이터 전체 또는 부분을 집계
- 부분총계: 데이터 셋 내 일정부분 레코드만 총계 처리
- 라운딩: 집계 처리된 값에 대하여 라운딩(올림, 내림, 사사오입) 기준 적용
- 재배열 : 기존 정보는 유지하되 개인정보가 식별되지 않도록 데이터를 재배열

데이터 삭제는 개인 식별 데이터를 삭제하는 기법이다. 개인식별요소의 전부 및 일부 삭제 처리가 가능하지만 분석의 다양성과 분석 결과의 유효성 및 신뢰성이 저하된다. 비식별 처리된 데이터 복구가 불가능하기 때문에 이 방식은 본 논문에서 다루는 EHR 비식별화로는 적합하지 않다.

데이터 범주화는 특정 데이터를 그룹의 대푯값이나 구간 값으로 변환하는 처리기법이다. 통계형 데이터 형식이므로 다양한 분석을 제시할 수 있지만, 정확한 분석 결과 도출이 어렵고 데이터 범위 구간이 적을 시 추론 가능성이 존재한다. 데이터 범주화는 수치 데이터에 대한 비식별화 방법이다.

- 감추기: 수치 데이터의 정확한 값을 숨기기 위하여 데이터의 평균 또는 범주 값으로 변환
- 랜덤 라운딩: 수치 데이터를 임의의 수 기준으로 올림 또는 내림
- 범위 방법: 수치 데이터를 해당 값의 범위 또는 구간으로 표현
- 제어 라운딩: 랜덤 라운딩 방법에서 어떠한 특정 값을 변경할 경우 행과 열의 합이 일치하지 않는 단점을 해결하기 위해 행과 열이 맞지 않는 것을 제어하여 일치시키는 기법

데이터 마스킹은 데이터의 전부 또는 일부분을 대체 값으로 변환하는 기법이다. 이 기법은 다른 기법들에 비해서 강도 조절에 대해 민감한 기법이다. 데이터 마스킹은 모든 문자 데이터를 비식별 처리할 수 있다.

- 임의 잡음 추가: 불필요한 데이터를 삽입하는 방법
- 공백과 대체: 공백 또는 대체문자로 바꾸는 기법

4. EHR 데이터 비식별 처리 기법

본 장에서는 EHR 데이터의 개인정보 민감도에 따른 데이터 분류를 제시하고 이러한 분류에 따른 비식별화 기법을 제안한다.

4.1 EHR 데이터 레벨 분류

다양한 프로젝트에서 EHR 데이터의 특성을 구현과 응용에 따라 다양하게 정의하여 활용하고 있다. 본 논문에서는 비식별 기법의 효율적인 제안을 위해서 다음과 같이 EHR 데이터를 정의한다.

- EHR 데이터베이스는 충분히 많은 환자의 데이터를 저장한다.
- 저장한 데이터는 환자의 치료를 목적으로 할 때만 사용된다.
- 데이터베이스는 3개 이상의 테이블로 구성되고 추가적인 데이터가 존재할 수 있다. 모든 개인정보 데이터는 비식별 처리 후 저장한다.
- 환자의 개인정보에 대한 추가적인 확인이 필요한 데이터의 경우, 재식별이 가능한 비식별 처리기법을 사용한다.

본 논문에서는 효율적인 EHR 데이터에 대한 비식별 처리 기법을 제안하기 위해서 개인정보를 민감도에 따라 레벨 0부터 4까지 단계별로 나눈다. 표 2는 EHR 데이터에 대한 민감도에 따른 비식별화 기법에 대한 분류이다. 효율적인 데이터 분류를 위해 본 논문에서는 심근경색(Heart Attack)인 환자들에 대한 비식별 처리를 대상으로 기법을 제안한다.

레벨 0는 가장 민감한 개인정보를 포함한 데이터들의 집합이다. 이러한 데이터는 시스템에서 데이터 교환 시 데이터의 훼손을 최소화하는 비식별 기법을 적용하여 처리해야 한다. 이에 해당하는 필드는 id, patientid, streetaddress, zipcode, givenname, surname 필드들이다.

레벨 1은 레벨 0과 비슷한 민감도를 가지지만 환자의 건강관련 정보들로 수치화된 데이터들을 고려한다. 수치 데이터에는 레벨 0과는 다른 비식별 처리를 적용하는게 적합하다. 이에 해당하는 필드는 sensordataid, senseddata, result 필드이다.

레벨 2는 레벨 0과 레벨 1에 비해서 민감도가 떨어지는 데이터들의 집합이다. 다만, 국적과 같은 필드를 이용하여 특정 인물을 파악하는 데에 도움을 주는 것과 같이 다른 데이터들과 결합하여 특정 개인의 정보가 간접적으로 드러날 수 있으므로 비식별 처리가 필요하다. 이에 해당하는 필드는 birthday, weight, height 필드로 한다.

레벨 3은 민감하지 않은 데이터들의 집합이다. 레벨 2에 비해 데이터의 민감도는 떨어지지만, 특정 개인을 연상케 하는 단서가 될 수 있으므로 비식별 처리가 필요하다. 이에 해당하는 필드는 city, country 필드로 한다.

레벨 4는 비식별 처리가 필요 없는 데이터들의 집합이다. 개인정보와 관련이 없으므로 보호가 필요 없는 데이터들이다. 이에 해당하는 필드는 gender, doctor 필드로 한다.

데이터 내 수치 데이터의 포함 여부에 따라 분류한 이유는 비식별 처리 기법 중 수치 데이터만을 대상으로 하는 처리기법이 존재하기 때문이다. 재식별이 가능한 비식별 처리기법만으로 비식별 처리를 해야 하는 환경이므로 재식별이 될 가능성이 높다. 이러한 이유로 다양한 비식별 처리기법을 혼용해야 한다.

<표 2> 개인정보 민감도에 따른 데이터 분류

특성 / 레벨	대상 필드	세부기술
레벨 0	id, patientid, streetaddress, zipcode, givenname, surname	암호화
레벨 1	sensordataid, senseddata, result	범위 방법
레벨 2	birthday, weight, height	교환 방법
레벨 3	city, country	감추기
레벨 4	gender, doctor	-

4.2 비식별화 기법

본 절에서는 EHR 데이터베이스 테이블의 각 대상 필드 데이터 민감도를 고려하여 표 2의 레벨에 따라 구체적 기법의 특성을 제안한다.

[암호화] 비식별 처리된 정보에 대한 원래 값의 확인이 필요한 경우 암호화가 적용된다. 암호화 기법에서는 대칭키 암호 시스템을 기본 기법으로 고려한다. 이는 연산에 있어서 효율성에 대한 고려로 인함이고, 서버의 과도한 부하를 방지하는데 그 목적이 있다. 이러한 이유로 민감한 개인정보를 비식별 처리를 함과 동시에 데이터의 훼손을 최소화 할 수 있다. 따라서 개인정보 민감도가 가장 높은 레벨 0의 비식별화 기법으로 암호화를 적용한다.

[범위방법] 범위 방법은 특정 데이터를 특정 범위나 구간으로 대체하여 재식별이 불가능하도록 적용하는 기법이다. 이로 인해 서로 공통점이 명확히 드러나고 개수가 여럿인 필드 또는 구간을 구분 지을 수 있는 필드를 대상으로 한다. 환자의 생체정보가 데이터의 집합으로 저장되고 그 데이터의 수 들은 임의적일 수 있다. 따라서 환자의 생체정보 집합 중 하나의 데이터를 가리키는 공통점을 갖는 데이터인 sensordataid, senseddata 필드 그리고 병명을 위한 result 필드를 대상으로 한다.

[교환방법] 교환 방법은 사전에 정해진 외부 변수 값과 연계하므로 참조할 외부 변수 값에 대한 설정이 추가로 필요하다. 특히, 이때 참조하는 외부 변수 값이 노출될 경우 시스템에 치명적 영향을 미치므로 이에 대한 보안 방안이 추가로 고려되어야 한다. 외부 변수 값은 서버 내에 하나의 테이블로 저장되고 이 테이블은 비식별화에 활용한다. 따라서 낮은 보안 수준을 요구하는 레벨 2에 이 기법을 적용한다.

[감추기] 감추기는 특수한 성질을 지닌 단체 데이터의 평균이나 범주 값에 적용하는 기법이다. 이 정보들을 통해 그 집단에 속한 개인의 정보를 쉽게 추론할 수 있다. 그래서 데이터의 일부가 노출되어도 경우의 수가 많아 추론하기 어려운 레벨 3에 적용한다.

5. 결론

본 논문에서는 원격 의료 진료 시스템의 개인정보에 대한 프라이버시 보호를 위한 비식별화 기법을 제안하였다. 효율적인 비식별화 기법을 제안하기 위해서 EHR 데이터를 개인정보의 민감도에 따라 4레벨로 분류하였다. 특히, 비식별화 표준에서 제안하고 있는 다양한 기법들 가운데 이들 데이터의 민감도를 효과적으로 반영할 수 있는 비식별화 기법에 대한 선정은 실질적인 응용을 개발하기 전에 선결되어야 하는 필수 과제이다. 본 논문에서 제안한 EHR 데이터에 대한 분류 및 비식별화 기법 제안은 다양한 의료 정보 서비스를 위한 프라이버시 보호에 활용될 수 있을 것이다.

사사(Acknowledgement)

본 연구는 중소벤처기업부에서 지원하는 2019년도 산학연 Collabo R&D사업(S2754028)의 연구수행으로 인한 결과물임을 밝힙니다. 또한 부분적으로 본 연구는 2017년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017R1D1A1B04032598).

참고문헌

[1] M. M. Dhanvijay, S. C. Patil, “Internet of Things: A survey of enabling technologies in healthcare and its applications,” Computer Networks, vol. 153, no. 22, pp. 113-131, 2019.

[2] H. Kim, “Freshness-Preserving Non-Interactive Hierarchical Key Agreement Protocol over WHMS,” Sensors, vol. 14, pp. 23742-23757, 2014.

[3] H. Kim, “Research Issues on Data Centric Security and Privacy Model for Intelligent Internet of Things based Healthcare,” ICSES Transactions on Computer Networks and Communications, vol. 5, no. 2, pp. 1-3, 2019.

[4] S. Cho, H. Kim, “Hash Chain based Authenticated Secure Communication for Healthcare System,” International Journal of Advances in Science Engineering and Technology, vol. 7, no. 2, pp. 41-46, 2019.

[5] S. W. Lee, T. Vallent, H. Kim, “Security and Privacy Measures on Data Mining for Internet of Things,” International Journal of Applied Engineering Reserach, vol. 13, no. 14, pp. 11648-11652, 2018.

[6] D. Ku, H. Kim, “Enhanced User Authentication with Privacy for IoT-based Medical Care System,” International Journal of Computer Theory and Engineering, vol. 10, no. 4, pp. 125-129, 2018.

[7] 박도휘, 강민영, 이명구, 박문구, “데이터 중심의 도시 운영, Data-Driven 스마트 시티를 주목하라,” 삼정 KPMG Issue Monitor, vol. 103, pp. 1-23, 2019.

[8] 임형진, “빅데이터 환경에서의 개인정보 비식별 처리 방법 분석,” 전자금융과 금융보안, vol. 8, pp. 9-37, 2017.

[9] 전웅렬, “개인정보 관리를 위한 PDS의 구조 및 보안기능 연구,” Journal of Information Technology and Architecture, vol. 15, no.1 3, pp. 345-356, 2018.

[10] 차연철, “데이터 경제와 개인정보 비식별 기술 동향,” 주간기술동향, pp. 15-29, 2019.

[11] H. Kim, “Data Centric Security and Privacy Research Issues for Intelligent Internet of Things,” ICSES Interdisciplinary Transactions on Cloud Computing, IoT, and Big Data, vol. 1, no. 1, pp. 1-2, 2017.

[12] Lastdrager, E.E.H. “Securing Patient Information in Medical Databases,” Master’s thesis, University of Twente, 2011.

[13] 김재한, “의료데이터 활용을 위한 개인정보 비식별화 기술 및 프로그램 동향”, 보건산업브리프, vol. 268, pp. 4-6, 2018.

[14] 이원석, “익명화 데이터의 익명 결합 방법”, 전자금융과 금융보안, vol. 15, pp. 67-102, 2019.