TTA Standard

정보통신단체표준(국문표준)

TTAx.xx-xx.xxxx 제정일: 2019년 12월 31일

데이터 비식별화 용어 정의 및 처리 과정

Data de-identification terminology definition and its process

한국정보통신기술협회
Telecommunications Technology Association

표준초안 검토 위원회 개인정보보호/ID관리, 블록체인 보안(PG502)

표준안 심의 위원회 정보보호기술위원회(TC5)

	성명	소 속	직위	위원회 및 직위	표준번호
표준(과제) 제안	염흥열	순천향대학교	교수	PG502 위원	
표준 초안 작성자	염흥열	순천향대학교	교수	PG502 위원	
	김지혜	한국정보기술단	연구원	PG502 참관위원	
	박정인	순천향대학교	연구원	PG502 참관위원	
사무국 담당	황예지	TTA		TC 5 간사	

본 표준 발간 이전에 접수된 지식재산권 확약서 정보는 본 표준의 '부록(지식재산권 확약서 정보)'에 명시하고 있으며, 이후 접수된 지식재산권 확약서는 TTA 웹사이트에서 확인할 수 있습니다.
본 표준과 관련하여 접수된 확약서 외의 지식재산권이 존재할 수 있습니다.

발행인 : 한국정보통신기술협회 회장
발행처 : 한국정보통신기술협회
13591, 경기도 성남시 분당구 분당로 47
Tel : 031-724-0114, Fax : 031-724-0109
발행일 : 2019.12

서 문

1 표준의 목적

비식별화는 조직이 수집, 사용, 보관 및 다른 조직과 공유하는 데이터에서 개인 정보를 제거하는 데 사용할 수 있는 도구다. 비식별화는 하나의 기술로 실현되는게 아니라 다양한 수준의 비식별화 효과를 갖는 데이터에 적용 할 수 있는 여러 접근 방법의 모음이다. 본 표준은 비식별화와 연관된 용어 정의, 주요 기술, 비식별화 과정을 제시한다. 본 표준은 데이터 비식별화 기법을 적용해 비즈니스 기회를 창출하고자 하는 개인정보 처리자에 의해 사용될 수 있다.

2 주요 내용 요약

이 표준은 비식별화와 관련되 개념, 용어, 비식별화 기법의 분류에 대해 기술한다. 또한 다양한 비식별화 기법과 비식별화 과정을 제시한다.

3 인용 표준과의 비교

3.1 인용 표준과의 관련성

본 표준은 다음 2개의 국제표준을 참조해 개발된다.

- ISO/IEC 29100:2011, Information technology – Security techniques – Privacy framework
- NISTIR 8053, De-Identification of Personal Information, October, 2015.

3.2 인용 표준과 본 표준의 비교표

Preface

1 Purpose of Standard

De-identification is a tool that organizations can use to remove personal information from a set of data that is collected, used, archived, and shared with other organizations. De-identification is not a single technique, but a collection of approaches that can be applied to data with different levels of effectiveness.

This standard provides concept and the terminology and a classification of approaches associated with de-identification, its process, approaches for reducing the risk of re-identification. This standard can be used by PII controller to create business opportunities by applying data de-identification techniques.

2 Summary of Contents

This standard describes a concept of de-identification and an associated terminology and a classification of approaches related to de-identification and its process. It also provides various de-identification approaches.

3 Relationship to Reference Standards

This standard is developed based on the following two international standards, but reflects the requirements of the domestic privacy laws and regulations.

- ISO/IEC 29100:2011, Information technology – Security techniques – Privacy framework
- NISTIR 8053, De-Identification of Personal Information, October, 2015.

목 차

데이터 비식별화 용어 및 처리 과정

(Data de-identification terminology and its process)

1 개요

비식별화는 조직이 수집, 사용, 보관 및 다른 조직과 공유하는 데이터에서 개인 정보를 제거하는 데 사용할 수 있는 도구다. 비식별화는 하나의 기술로 실현되는게 아니라 다양한 수준의 비식별화 효과를 갖는 데이터에 적용할 수 있는 여러 접근 방법의 모음이다. 본 표준은 비식별화와 연관된 용어 정의, 주요 기술 개요, 그리고 비식별화 과정을 제시한다.

2 표준의 구성 및 범위

이 표준은 비식별화와 관련되 개념, 용어, 비식별화 기법의 주요 개요를 기술한다. 또한 다양한 비식별화 기법과 비식별화 과정을 제시한다.

3 참조 표준(권고)

- ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and classification of techniques
- NISTIR 8053, De-Identification of Personal Information, October, 2015.
- Information and Privacy Commissioner of Ontario, De-identification Guidelines for Structured Data, June 2016

4 용어정의

4.1 가명 (pseudonym)

데이터 주체에 대한 식별자를 대체하기 위해 데이터 주체에 대해 생성된 별도 고유 식별자

4.2 가명화 (pseudonymization)

데이터 주체의 식별자를 가명으로 대체하여 해당 데이터 주체의 신원을 숨기는 비식별 기법

4.3 간접 식별자 (indirect identifier)

데이터 세트 내 또는 외부에 있는 다른 속성과 함께 특정 운영 환경에서 데이터 주체를 고유하게 식별하게하는 속성

4.4 고유 식별자 (unique identifier)

데이터 세트에서 데이터 주체를 골라 내는 데이터 세트내 속성

4.5 공격자 (adversary)

데이터 세트에서 하나 이상의 개인을 식별하려고 시도하는 개인 또는 단체

4.6 난수화 기법 (randomization technique)

속성 값이 새로운 값으로 무작위 적으로 변경되도록 속성 값을 수정하는 비식별 기법

4.7 데이터 세트 (data set)

데이터 모음

4.8 데이터 주체 (data principal)

데이터와 관련된 실체

4.9 등가 클래스(equivalence class)

특정 속성 집합에 대해 동일한 값을 갖는 데이터 세트에서 레코드의 집합

4.10 레코드 (record)

단일 데이터 주체에 관한 속성의 집합

4.11 배포 모델 (release model)

데이터 세트의 수신자에게 접근 권한이 제공되는 방식

4.12 마스킹 (masking)

데이터 주체의 직접 식별자를 제거하거나 가명 또는 암호 값으로 바꾸는 프로세스

4.13 마이크로데이터 (microdata)

개별 데이터 주체와 관련된 레코드로 구성된 데이터 세트

4.14 매크로데이터 (macrodata)

총합 데이터로 구성된 데이터 세트

4.15 민감 속성 (sensitive attribute)

운영 환경에 따라 속성 값의 노출, 속성 값의 존재, 또는 어떤 데이터 주체와 연관을 가능하게 하는 잠재적 재식별 공격으로부터 특화되고 높은 수준의 보호를 받을 가치가 있는 데이터 세트에서 속성

4.16 변수 (variable)

속성 집합을 나타내는 데이터 세트의 열의 값

4.17 비식별화(de-identification)

일련의 식별 데이터와 데이터 주체 간의 연관성을 제거하는 프로세스

4.18 비식별화 과정 (de-identification process)

일련의 식별 속성과 데이터 주체 사이의 연관을 제거하는 과정

4.19 비식별화 기법 (de-identification technique)

정보가 개별 데이터 주체와 연관 될 수 있는 정도를 줄이기 위할 목적으로 데이터 세트를 변형하는 방법

4.20 비식별화된 데이터 세트 (de-identified dataset)

비식별화 과정의 결과로 나타난 데이터 세트

4.21 속성 (attribute)

고유 특성

4.22 순열 (permutation)

값을 수정하지 않고 데이터 세트의 레코드 전반에 걸쳐 선택된 속성의 값을 재정렬하는 비식별 기법

4.23 식별자 (identifier)

특정 데이터 처리 환경에서 데이터 주체의 고유 식별을 가능하게 하는 데이터 세트 내 속성들의 집합

4.24 식별 속성 (identifying attribute)

특정 데이터 처리 환경에서 데이터 주체를 고유하게 식별하는 데 기여할 수 있는 속성의 데이터 세트

4.25 신원 노출 (identity disclosure)

데이터 주체의 신원을 올바르게 할당하게 하는 재식별화 이벤트

4.26 연결 (linking)

데이터 주체에 관한 레코드를 별도의 데이터 세트에서 동일한 데이터 주체에 관한 레코드와 연결시키는 행위

4.27 연결성 (likability)

데이터 주체에 관한 레코드를 별도 데이터 세트에 존재하는 동일한 데이터 주체에 관한 레코드와 연관시킬 수 있는 데이터 세트에 대한 속성

4.28 일반화 (generalization)

선택된 속성에 포함된 정보의 정확성을 줄이는 비식별 기법

4.29 일방향 해시 함수 (one-way hash function)

암호화 된 값에서 입력 데이터를 다시 생성하는 것이 사실상 불가능한 암호화 매핑 함수

4.30 잡음 부가

선택된 속성의 값에 임의의 값을 추가하여 데이터 세트를 수정하는 비식별 기법

4.31 전수 공격 (brute force attack)

가능한 모든 조합을 시도하는 시행 착오적 공격

4.32 준 식별자 (quasi identifier)

데이터 세트에서 다른 속성과 함께 고려될 때 데이터 주체를 선택하는 데이터 세트내 속성

4.33 재식별 (re-identification)

비식별된 데이터 세트의 데이터를 원래 데이터 주체와 연관시키는 과정

4.34 재식별 공격(re-identification attack)

재식별을 목적으로 공격자가 비식별 데이터에 대해 수행하는 행위

4.35 재식별 위험 (re-identification risk)

재식별 공격이 성공할 위험

4.36 직접 식별자 (direct identifier)

특정 운영 환경 내에서 데이터 주체의 고유 식별을 가능하게 하는 속성

4.37 차등 프라이버시 (differential privacy)

특정 데이터 주체가 입력 데이터 세트에 나타나는지 여부에 무관하게, 통계 분석의 출력 확률 분포가 지정된 값보다 작게 다르도록 보장하는 공적 프라이버시 측정 모델

4.38 총계 데이터(aggregated data)

정보 주체의 그룹을 나타내는 데이터 (예, 그룹의 통계적 특성의 모음)

4.39 추론

하나 이상의 속성 값을 사용하거나 외부 데이터 소스를 상호 연관시킴으로써 무시할 수 없는 확률로 알려지지 않은 정보를 추론하는 행위

4.40 K 익명성

데이터 세트의 각 식별자에 대해 적어도 k개의 레코드를 포함하는 대응 등가 클래스가 존재하는 것을 보장하는 공식적 프라이버시 측정 모델

4.41 L 다양성

선택된 속성에 대해 각 등가 클래스가 최소 L 개 이상의 잘 표현 된 값을 가지도록 보장하는 공식적 프라이버시 측정 모델

4.42 T 유사성

등가 클래스에서 선택된 속성의 분포와 전체 테이블에서 이 속성의 분포 사이의 거리가 임계 값 T 이하가 됨을 보장하는 공식적 프라이버시 측정 모델

5 비식별화 개요

비식별화는 레코드 또는 데이터 세트에서 개인 정보를 삭제하는 프로세스이다. 개인 정보는 식별 가능한 개인에 관한 정보이다. 비식별화를 구체적으로 정의하면 (i) 개인을 식별하는 정보 또는 (ii) 개인이 식별 할 수 있는 합리적인 기대가 존재하는 정보를 제거하는 과정이라고 볼 수 있다. 개인을 식별하기 위해 데이터 단독으로 또는 다른 데이터와 함께 사용된다. 비식별화는 일련의 과정, 고려 사항 및 가능한 결과, 그리고 연관된는 비식별 과정을 언급 할 때 사용된다.

데이터 세트를 비식별할 때는 다음과 같은 여러 가지 이슈를 고려해야 한다.

데이터 배포 모델. 비식별 데이터 세트는 공개적으로, 세미- 공개적으로 (유사 공개), 또는 비공개적으로 배포될 수 있다. 공개 데이터 배포는 비식별된 데이터 세트를 아무 조건없이 다운로드하거나 사용할 수 있다. 이러한 종류의 배포는 최대한의 가용성을 제공하지만 수신자에 의한 최소한의 보호조치만을 제공한다.

비공개 데이터 공개는 데이터 집합의 가용성을 소수의 수신자로 제한한다. 데이터 수신 조건으로 수신자는 데이터의 개인정보 보호 및 보안과 관련된 이용 약관 (일반적으로 데이터 공유 계약에 명시되어 있음) 에 동의해야 한다. 이러한 종류의 배포는 가용성을 최소화하지만보다 수신자에 의한 많은 보호조치를 제공 할 수 있다.

데이터 세트는 세미 공개로 배포될 수 있다. 공개 및 비공개 옵션의 요소를 포함한다. 세미 공개 데이터 배포에서는 누구나 다운로드 할 수 있다. 그러나 데이터 수신 조건으로 수신자는 데이터 세트를 공개하는 조직에 등록하고 데이터 처리 및 공유에 대한 제한 (일반적으로 사용 약관 형식) 에 동의해야 한다.

세미 공개 데이터 배포에 대한 이용 약관에 추가적인 개인정보 보호 및 보안 조치가 포함될 수 있지만, 이러한 보호조치는 배포 공개 성으로 인해 담보하기가 어렵다. 따라서 이러한 방법으로 배포된 데이터 세트는 제공 할 수 있는 보호 수준에 제한적이다. 사용되는 배포 모델에 따라 비식별화 량이 다를 수 있다.

다른 종류 식별자. 비식별화는 개인을 직접 식별하는 정보와 정보를 단독으로 또는 다른 정보와 함께 사용하여 개인을 식별 할 수 있는 "합리적 기대"가 있는 정보를 제거해야 한다. 첫 번째 유형의 식별자는 "직접 식별자" 라고 하고, 두 번째 유형은 "간접 식별자"또는 "준 식별자"라고 한다.

다른 재식별 공격. 데이터 세트에 적용되어야 하는 비식별의 양은 공격자가 데이터 세트에서 하나 이상의 개인을 재식별하려고 시도할 확률에 따라 결정된다. 사용되는 배포 모델에 따라 다양한 종류의 공격자를 고려해야 하며 다양한 유형의 재식별 공격을 분석해야 한다. 예를 들어 공개 데이터 배포의 경우 누군가가 데이터 세트에 대한 데모 공격을 시도한다고 가정해야 한다. 비공개 데이터 배포의 경우 내부자 및 데이터 침해로 인한 위협을 평가해야 한다.

다른 비식별 기술. 재식별 위험 수준이 파악되고 필요한 비식별의 양이 계산된 후에서는 해당 정보를 데이터 세트에서 제거해야 한다. 이것은 마스킹, 일반화 및 삭제와 같은 기술 등의 다양한 방법으로 수행 할 수 있다.

다양한 유형의 배포. 비식별은 개인의 신원 공개와 이와 정보로의 연결을 막는다. 그러나 그들은 특정의 개인 집단과 관련된 속성을 공개하는 것은 보호하지 않는다. 비식별 데이터 세트를 배포 할 때 개인 신원 공개에 대비해야 하는 것은 물론 속성 공개에 대한 보호도 고려해야 한다. 이를 위해서는 데이터 세트에 대한 윤리적 검토를

포함하는 거버넌스 모델을 개발해야 할 수 있다.

6 비식별화, 재식별화, 데이터 공유 모델

비식별의 주된 목적은 개인 사생활을 보호하는 것이다. 데이터 세트에 개인 정보가 포함되어있는 경우 해당 데이터 세트는 비식별된 것으로 간주 될 수 없다.

동시에 비식별 데이터 세트를 공개하는 주된 이유 중 하나는 연구 목적으로 원시 데이터의 가치와 속성을 연구할 기회를 제공하는 것이다. 따라서 비식별은 개인의 사생활을 보호하면서 최대한 많은 정보를 유용하게 유지해야 한다.

이러한 비식별의 두 가지 목적은 개방형 데이터, 정보 요구에 의한 접근, 그리고 기관 간 데이터 공유 등을 포함한 다양한 컨텍스트에서 사용을 고려해야 한다.

개방형 데이터

비식별은 기관이 개인 정보를 공개 할 권한이 없는 상황에서 데이터 공유를 가능하게 하는 데 사용될 수 있다. 공개 데이터 이니셔티브는 데이터 세트를 사전에 공개하고 사용 및 재 게시를 위해 누구에게나 자유롭게 사용할 수 있게 함으로써 정부의 투명성과 책임성을 높이려고 한다. 이러한 이니셔티브가 제공하는 정보의 양과 가용성이 늘어남에 따라 기관이 개인의 사생활을 보호하는 방식으로 데이터 세트를 공개하는 것이 중요하다.

개방형 데이터 이니셔티브는 연구, 혁신 및 새로운 애플리케이션 및 서비스 개발을 촉진을 추구한다. 개방형 데이터 세트의 유용성이 높을수록 공개 데이터를 사용하고자하는 연구자, 신생 기업 및 기업가가 성공할 확률이 높아진다.

정보 요청에 대한 접근

비식별은 구조화된 데이터 또는 데이터 세트에 대한 정보 접근 요청에 응답하는 데 유용 할 수 있다. 기관은 공개 면제된 정보를 제외하고 많은 기록을 공개해야 한다. 비식별을 사용함으로써 기관은 정보의 유용성을 유지하면서 개인정보 보호 방식으로 요청에 응답 할 수 있다. 비식별은 이전에는 불가능했던 방식으로 투명성을 증진할 기회를 기관에게 제시할 수 있는 혁신적인 도구이다.

조직 내 및 조직 간 데이터 공유

정보 요청에 의한 접근 및 개방형 데이터 이니셔티브는 일반 대중에게 정보를 제공하지만, 정부 기관들 간에 벽을 허물고 더 많은 정보를 내부 및 기관 간에 공유하려는 정부 서비스에 대한 욕구 또한 커지고 있다. 이것은 여러 가지 이유로 발생할 수 있다. 예를 들면,

- ‥한 기관 또는 프로그램 지역의 정보가 다른 기관 또는 지역의 프로그램 또는 서비스 계획과 관련 될 수 있다

- ‥한 기관은 다른 기관에서 요구하는 데이터 처리 또는 소프트웨어 개발에 대한 전문 지식을 보유하고있을 수 있다.

- ‥다른 기관에서 제공한 프로그램이나 서비스에 자금을 지원한 기관은 프로그램이나 서비스의 효율성을 평가하고자 할 수 있다.

개인 정보가 포함 된 데이터 세트는 개인정보보호법에 의거하여 공개가 허용되는 경우에만 기관 내 및 기관간에 공유 될 수 있다. 공개가 허용되지 않고 기관이 여전히 데이터 세트를 공유하고자 하는 경우 (정보 요청에 대한 접근 또는 공개 데이터 배포와 유사하게) 개인 정보를 제거해야 한다.

그러나 공개가 허용 되더라도 고려해야 할 중요한 개인 정보 보호 문제가 있을 수 있다. 기관 간 정보 공유가 보다 효율적이고 효율적인 서비스를 제공하는 데 중요한 역할을 할 수 있지만 개인 정보에 대한 통제력을 감소시킴으로써 개인의 프라이버시를 침해하는 의도하지 않은 결과를 초래할 수도 있다. 따라서, 모범 사례로서, 기관들은 데이터 세트를 공유하기 전에 항상 데이터 세트를 비식별하는 것을 고려해야 한다.

7 비식별화 기법

7.1 직접 식별자 삭제

직접 식별자는 "한 개인을 직접 식별하는 데이터" 이다. 이의 예는 이름, 사회 보장 번호 및 전자 메일 주소가 있다. 직접 식별자는 삭제되거나, 다른 범주 이름으로 대체되거나 특정 심벌로 대체되거나 난수로 대체되거나 가명으로 대체된다.

7.2 가명화

가명화는 개인을 직접 식별 이름과 기타 정보를 가명으로 대체하는 변환이다. 가명 화는 모든 직접 식별자가 체계적으로 가명 된 경우 여러 데이터 레코드 또는 정보 시스템에서 개인에 속한 정보를 연결할 수 있다.

7.3 준식별자의 비식별화

간접 식별 변수라고도하는 준식별자는 그 자체로는 특정 개인을 식별하지 못하지만 정보 주제를 식별하기 위해 다른 정보와 통합되고 "연결" 될 수 있는 식별자이다.

준식별자를 비식별하는 다음과 같은 여러 방법이 존재한다.

준식별자 삭제: 준식별자는 삭제되거나 제거 될 수 있다.

일반화: 특정 준식별자 값은 주어진 범위 또는 집합의 구성원으로 보고 될 수 있다. 예를 들어, 우편 번호 12345는 12000에서 12999 사이의 우편 번호로 일반화 될 수 있습니다. 일반화는 전체 데이터 세트 또는 특정 레코드에 적용될 수 있다.

순열(Perturbation): 특정 값은 정의된 일반화 수준 내에서 각 개인에 대해 일관된 방식으로 다른 값으로 대체 될 수 있습니다. 예를 들어, 모든 연령대는 원래의 나이에 무작위로 (-2 ... 2) 년 동안 조정되거나 날짜 또는 입원 및 퇴원이 같은 수의 (-1000 ... 1000) 일로 이동 될 수 있다.

교환 (Swapping): 준식별자 값은 정의 된 일반화 수준 내에서 레코드 간에 교환될 수 있다. 교환은 통계적 속성을 보존해야 할 경우 주의해서 처리해야 한다.

서브 샘플링: 전체 데이터 세트를 배포하는 대신, 조직은 샘플만을 배포할 수 있다. 서브 표본만 발표하면 재식별 확률은 감소한다.

8 비식별화 과정

정보에 가능한 한 많은 유용성을 유지하면서 정보주체의 프라이버시를 보호하기 위해 데이터의 공개와 관련된 재식별 위험의 수준과 종류를 체계적으로 분석하여 비식별 양과 유형을 결정되어야 한다. 데이터 세트를 비식별화하는 경우 다음 과정을 고려해야 한다.

- 배포 모델 결정
- 변수 분류
- 허용 가능한 재 식별 위험 임계치 결정
- 데이터 위험 측정
- 문맥 위험 측정
- 전반적인 위험 계산
- 데이터 비식별화 수행
- 데이터 유용성 평가
- 과정 문서화

8.1 배포 모델 결정

위에서 언급했듯이 비식별 데이터 세트는 공개적으로 배포, 준-공개적으로 배포 또는 비-공개적으로 배포 될 수 있다. 각 배포 모델은 다양한 수준의 가용성 및 정보 보호를 허용한다. 데이터 공개의 목적 또는 법적 요구사항에 따라 각 모델의 적절성은 변할 수 있다.

배포 모델은 모델에 따라 필요한 비식별의 크기가 다르므로 비식별 과정에서 중요하다. 예를 들어, 공용 데이터 배포는 큰 가용성을 제공하지만 보호 수준이 가장 낮기 때문에 개인 프라이버시를 보호하려면 상당한 양의 비식별이 필요하다. 비공개 데이터 배포는 가용성은 가장 낮지만 상대적으로 높은 보호 수준을 제공 할 수 있으므로 더 적은 양의 비식별이 필요로 하다.

비식별 데이터 세트의 접근 요청은 공용 데이터 배포인 것처럼 처리되어야 한다.

공개 데이터로 배포 할 때 비식별 데이터에 접근자와 접근 방법을 포함해 비식별 데이터에 접근을 위한 가능한 작은 제한을 두는 것이 좋다. 개인이 비식별 데이터를 공개하는 조직에 등록하고 자신을 식별하는 요구는 그러한 비식별 데이터를 접근하고 사용하는 능력에 장벽으로 간주된다. 따라서 데이터 세트를 다운로드 한 개인을 식별 할 수 없는 경우 이러한 공개는 공개적인 데이터 배포로 처리되어야 한다.

그러나 개인의 등록 및 신원 확인이 필요한 경우가 있다. 예를 들어, 정부나 대학이 후원하는 경쟁 프로그래밍 또는 "hackathon"은 비식별 데이터 세트를 일반 또는 학생 단체에 대해서만 배포하지만, 참가자에게 비식별 데이터의 이용을 특정한 방법으로만 제한할 수 있다. 이용 악관에서 비식별 데이터 세트에서 개인을 식별하고 이용 약관을 통해 제3자에게 비식별 데이터를 공개하는 것을 금지하는 것이다. 이용 약관에 참여자가 개인정보 보호 및 보안 조치를 추가로 요구하지 않거나 그러한 조치가 집행 할 수 없는 경우, 이러한 종류의 배포는 준-공개적인 배포로 처리되어야 한다.

마지막으로, 기관간에 정보를 공유 할 때 데이터 세트에 대한 접근이 수신 기관으로 제한되기 때문에 데이터 공유 계약을 통해 비식별 데이터 세트의 개인정보 보호 및 보안에 관한 요구사항을 수립하고 데이터 공유 계약을 통해 시행 할 수 있다. 이러한 경우 이러한 배포는 비공개 데이터 배포로 처리될 수 있다.

데이터 공개가 비공개의 경우, 두 당사자간에 데이터 공유 계약이 체결되어야 한다. 데이터 공유 계약은 이러한 배포에서 위험 완화 전략의 중요한 부분이 된다.

8.2 변수 유형 분류

만약 데이터 세트가 개인에 관한 것이면 파일의 각 행은 개인을 나타내고 각 열은 개인에 대해 수집된 정보의 변수를 나타낸다. 데이터 유형에 따라 일부 변수는 직접

또는 간접적으로 개인을 식별하는 데 사용될 수 있지만, 다른 변수는 식별하는데 이용될 수 있다. 비식별은 개인을 식별하는 데 사용될 수 있는 변수에만 관련된다. 위에서 언급했듯이 직접 식별자와 간접 또는 간접 식별자의 두 가지 변수가 존재한다.

직접 식별자

직접 식별자는 그 자체로 또는 다른 쉽게 구할 수 있는 정보 소스와 함께 하나의 개인을 식별하는 데 사용할 수 있는 하나 이상의 변수로 구성된다. 예를 들면 이름, 주소, 전자 메일 주소, 전화 번호, 팩스 번호, 신용 카드 사회 번호, 의료 기록 번호, 의료 기록 번호, 장치 식별자, 생체 인식 식별자, 인터넷 프로토콜 (IP) 주소 번호 및 웹 범용 리소스 로케이터 (URL)와 같은 정보를 포함 할 수 있다.

일반적으로 직접 식별자는 데이터 분석의 목적으로는 유용하지 않다. 예를 들어, 개인의 이메일 주소는 직장 통근 연구와 관련이 없을 것이다. 그러나 직접 식별자의 값이 관련된 경우 이를 간접 식별자로 분류하고 이 변수에 대해 비식별을 해야 한다. 그러나 변수가 데이터 분석에 유용하지 않은 경우에는 변수를 직접 식별자로 분류하고 특성에 관계없이 가명으로 제거하거나 대체하도록 플래그를 지정해야 한다.

준식별자

준식별자는 두 가지 중요한 특징을 갖는 변수이다. 1) 적이 준식별자에 대한 배경 지식을 가지고 있다고 가정하고, (2) 데이터 세트에서 개인을 재식별하기 위해 개별적으로 또는 조합으로 사용할 수 있다. 적이 변수에 대한 배경 지식이 있는 경우에만 변수가 간접 식별자가 될 수 있다. 준식별자를 분류하는 문제는 배경 지식의 가능한 출처를 개대하는 데 있습니다. 적들은 다음과 같은 다양한 방법으로 데이터 세트의 한 명 이상의 개인에 대한 배경 지식을 얻을 수 있다.

- 공개된 저장소 (예, 유권자 목록이나 법원 기록)에서, 매체(예, 사망 기사)에서, 전문 조직 (예, 회원 목록) 또는 고용주 (예, 직원 명부 또는 약력) 입장에서, 개인에 관한 정보가 가용할 수 있다.
- 적이 하나 이상의 개인 (예, 이웃, 동료 또는 전 배우자)을 알 수 있다.
- 하나 이상의 개인이 유명 인사 일 수 있고 공개적으로 사용할 수 있는 정보가 존재한다.
- 적이 개인에 관한 추가 정보 소스 (예, 다른 연구 프로젝트의 데이터 세트) 에 접근할 수 있다.
- 개인은 온라인 (예, 소셜 네트워킹 사이트 또는 개인 블로그)에 자신에 관한 정보를 게시 할 수 있다.

간접 식별자의 예로는 성별, 생년월일 또는 나이, 행사 날짜 (사망, 입원, 절차, 퇴원,

방문), 위치 (예, 우편 번호, 건물 이름, 지역), 출신 민족, 출생 국가, 원주민 지위, 눈에 띄는 소수 민족 지위, 직업, 결혼 상태, 교육 수준, 학교 전체 연수, 범죄 경력, 총 수입 및 종교 교단 등이 포함된다. .

간접 식별자의 값은 데이터 세트에서 상관 관계를 공유하는 하나 이상의 변수로부터 예측 될 수 있다. 예를 들어, 개인의 나이는 졸업 날짜 또는 연도로부터 예측 될 수 있다. 이러한 변수는 간접 식별자의 값을 나타낼 수 있으므로 간접 식별자로 분류해야 한다.

8.3 허용 가능한 재식별 위험의 임계치 결정

비식별은 개인을 식별하는 정보를 제거하거나, 개인을 식별하기 위해 단독 또는 다른 정보와 함께 사용할 수 있다는 합리적 기대치가 존재하는 개인의 프라이버시를 보호한다. 개인 프라이버시를 보호하기 위해 적용해야하는 비식별화 양은 데이터 세트 배포와 관련된 재식별의 위험 수준에 비례한다. 데이터 배포의 재식별 위험이 높을수록 필요한 요구되는 비식별화 양은 커진다.

데이터 세트에 대한 재식별 위험 (또는 임계 값)의 수용 가능한 수준을 결정하려면 데이터 셋의 배포가 정보주체의 프라이버시를 침범 할 정도가 평가되어야 한다. 평가 결과는 전형적으로 "낮음", "중간", 또는 "높음"의 범위에 있는 정성적 값이어야 한다.

정보주체의 잠재적 프라이버시 침범의 수준을 평가할 때, 데이터 세트의 데이터가 식별 가능하고 비식별이 발생하지 않았다고 가정한다. 이 가정 하에서, 프라이버시 침해 수준은 다음을 포함하는 여러 요인의 함수가 된다.

- 데이터의 민감도
- 데이터의 범위와 세부 수준
- 정보주체의 수
- 위반 또는 부적절한 사용으로 인해 정보주체에게 발생할 수 있는 잠재적 피해나 손해
- 데이터 공개가 정보주체 동의없이 관련 법령에 근거해 허용되는지 여부
- 프라이버시에 대한 기대가 거의 없거나 전혀 없는 상태에서, 데이터가 정보주체에 의해 자유롭게 주어졌는지 여부
- 정보주체가 이차적 목적을 위해 데이터가 비식별 형태로 배포 되는 데 명시적으로 동의했는지 여부 또는 이 데이터를 수집 할 때 적절하게 통지되었는지 여부

프라이버시 평가의 침해의 결과는 정량적 수치이다. 데이터 세트에 적용되어야 하는

비식별의 양은 수치로 정량화된다. 프라이버시 침해 가치를 평가 한 후에 그 결과를 수치로 변환해야 한다. 이 수치 값은 해당 위험 수준에 비례하는 비식별의 양을 나타낸다. 이 "재식별 위험 임계치"는 일반적으로 개인정보가 더 이상 식별되지 않는 것으로 간주되기 위해 데이터 세트에 적용되어야 하는 비식별의 최소 양을 나타낸다. 따라서 이 값은 향후 진행될 비식별에 관한 계산값과 비교되는 기준선을 형성한다.

프라이버시 침해의 정량적 값과 재식별 위험 임계치 값 사이를 변환 할 때, 비식별의 핵심적인 측면을 고려해야 한다. 즉, 비식별은 재식별 가능성이 없는 데이터 집합을 생성하지 않는다. -신분증. 오히려 배포와 관련된 재 식별 위험 수준을 고려할 때 재 식별 확률이 매우 낮은 데이터 세트를 생성합니다. 프라이버시 가치의 침해에 비례하는 탈 식별의 양은 그 위험 수준을 감안할 때 매우 낮은 재확인 확률과 같아야 한다.

다음 표는 사생활 가치의 침해가 다른 데이터 세트에 대한 재발견 가능성에 대해 매우 낮은 가치로 간주 될 수 있는 것을 결정할 때 가이드 라인으로 사용될 수 있다.

<표 8-1> 비식별화 등급 예

프라이버시 침해 가능성	재식별화 확률	등가 셀 크기
낮음	0.1	10
보통	0.075	15
높음	0.05	20

예를 들어, 재식별 확률이 0.1 인 데이터 세트는 데이터 세트의 각 행이 일반적으로 9 개의 다른 행과 같은 준식별자에 대해 동일한 값을 갖게 됨을 의미합니다. 즉, "셀 크기"가 10임을 의미한다.

8.4 데이터 위험 측정

허용 가능한 재식별 위험 임계치를 결정한 후, 다음 단계는 데이터 세트 자체의 재식별 위험의 양을 측정하는 것이다. 데이터 위험은 배포와 관련된 재식별 위험 수준을 결정하는 데 사용된다.

데이터 세트에서 재 식별 위험의 양을 측정하는 것은 두 단계의 과정으로 구성된다. (1) 각 행의 재식별 확률을 계산하고 (2) 사용된 배포 모델을 기반으로 적절한 위험 측정 방법을 적용해야 한다.

각 행의 재식별 확률 계산

개인에 관한 데이터 세트의 각 행에는 한 개인에 대한 정보가 들어 있다. 따라서, 각 행은 재식별 가능성을 갖는다. 주어진 행에 대해, 재식별 확률은 데이터 세트의 다른 행의 개수가 간접 식별자인 변수에 대해 동일한 값을 갖는지에 따라 결정된다.

간접 식별자인 변수에 대해 동일한 값을 갖는 데이터 세트의 모든 행은 "동등 클래스"를 형성한다. 예를 들어, 성별, 연령, 그리고 교육 수준이 포함된 데이터 세트에서, 대학 교육을 받은 35세 이상의 남자에 해당하는 모든 행은 등가 클래스를 형성한다. 등가 클래스의 크기는 간접 식별자 값이 같은 행의 개수와 같다.

각 행에 대해 재식별 확률은 1을 등가 클래스의 크기로 나눈 값과 동일하다. 예를 들어, 크기 5의 등가 클래스의 각 행은 0.2의 재 식별 확률을 가진다.

주어진 행에 대해 재식별 확률 = 1/등가 클래스의 크기

더 많은 등가 클래스를 가진 행은 데이터 세트의 더 많은 행과 더 많은 개인이 간접 식별자에 대해 동일한 값을 가지기 때문에 재식별 확률이 낮다. 더 작은 등가 클래스를 가진 행은 더 적은 행 (개인 수가 적음)이 간접 식별자에 대해 동일한 값을 갖기 때문에 재 식별 가능성이 높다.

적절한 위험 평가 방법의 적용

각 행의 재 식별 확률은 1을 등가 클래스의 크기로 나눈 값과 같다. 사용된 배포 모델에 따라 데이터 세트의 재식별 위험 량을 측정하기 위해 여러 가지 방법이 있다.

공개 데이터: 최대 위험

공용 데이터 배포의 경우 누군가가 홍보를 위해 데모 공격을 시도한다고 가정해야 한다. 이러한 종류의 공격은 데이터 집합에서 가장 취약한 행을 대상으로 한다. 이러한 행은 등가 클래스가 가장 낮고 재식별 가능성이 가장 크다. 이 때문에 모든 행에 대해 재식별의 최대 확률을 사용하여 재 식별 위험의 양을 측정해야 한다.

비공개 데이터: 엄격한 평균 위험

비공개 데이터 배포의 경우, 데이터 세트에 대한 액세스가 지정된 수의 수신자로 제한되기 때문에 어떤 행도 재식별 공격에 대해 다른 행보다 더 취약하지 않다고 가정해야 한다. 여기서 모든 행에 걸친 재식별의 평균 확률을 사용하여 데이터 세트의

재 식별 위험의 양을 측정해야 한다. 그러나 재식별의 위험이 높은 특정 행 또는 등가 클래스를 보호하기 위해 어떤 행도 특정 값보다 큰 재식별 확률을 가지지 않게 해야 하는 평균이 "엄격한" 평균이여야 한다. 종종 0.33 임계치로 제안된다. 즉, 데이터 집합에서 등가 클래스의 최소 크기는 3이여야 한다. 그러나 실제 0.5의 재식별 확률이 최대로 사용될 수 있다. 이 경우, 엄격한 평균은 재식별 확률이 매우 작은 고유한 행이 없고 평균 위험이 허용 가능하게 작다는 것을 보장한다.

세미-공개 데이터 배포: 최대 위험

세미 공개 데이터는 누구나 다운로드 할 수 있기 때문에 가장 취약한 행이 다른 공격에 비해 공격 위험이 높다고 가정해야 한다. 따라서 공용 데이터 배포와 마찬가지로 재식별 위험의 량을 측정하기 위하여 모든 행에 대해 최대 재식별 확률이 되도록 해야 한다.

8.5 컨텍스트 위험 평가

데이터 세트의 위험은 데이터 세트 배포와 관련된 재식별 위험 수준을 결정하는 데 중요한 역할을 하지만 고려해야 할 유일한 위험 요인은 아니다. 재식별 위험은 이 주어진 데이터 세트에서 가능한 재식별 공격의 사용된 배포 모델에 따라 달라진다. 가능한 공격의 관점에서 재식별 위험을 더 분석하면 상황에 따른 위험이 초래된다. 데이터 위험과 함께 이 값은 데이터 집합 배포와 관련된 재식별의 전체 위험을 계산하는 데 사용된다.

컨텍스트 위험은 데이터 세트에 대해 하나 이상의 재식별 공격이 시작될 확률이다. 일단 데이터 세트가 배포되고 되면 재식별 공격이 비식별 데이터에 대해 시작될 수 있지만 사용되는 배포 모델에 따라 공격자와 공격의 종류가 다르다.

공공 데이터 배포

공개 데이터 배포 상황에서 컨텍스트 위험을 측정하는 데 이용되는 계산은 간단하다. 데이터 세트는 아무 조건없이 다운로드하거나 사용할 수 있도록 만들어 졌기 때문에 누군가가 홍보를 위해 데모 공격을 시도한다고 가정해야 한다. 따라서 공격자가 데이터 세트에 대한 재식별 공격을 시작할 확률은 1이다.

비공개 데이터 배포

이에 반해, 비공개 데이터 배포에 대한 컨택스트 위험을 측정하기 위한 계산은 더욱 복잡하며 전문 지식이나 기술이 필요할 수 있다. 소개에서 언급했듯이 이 계산을 수행할 수 있는 자신감이 없다면 기술자나 현장 전문가에게 조언을 구할 수 있다.

기술 자문이나 전문가의 조언이 없는 경우, 다른 선택은 공개 데이터를 배포에서처럼 위 (훨씬 간단한) 방법을 사용하여 컨텍스트 위험을 측정 할 수 있다. 이로 인해 활용도가

낮은 데이터 세트가 될 수 있지만 재식별 공격에 대한 보호의 양은 비공식 데이터 배포와 등가일 수 있다.

비공개 데이터 배포의 경우 세 가지 다른 재식별 공격 또는 위협의 확률을 결정해야 한다.

- 의도적 내부자 공격
- 지인에 의한 데이터 세트 내에서 개인의 의도치 않은 인식
- 데이터 유출

컨텍스트 위험을 측정 할 때 이 확률 중 가장 높은 값을 사용해야 한다.

공격 1: 의도적 내부자 공격

데이터 세트에서 하나 이상의 개인을 재식별하려고 시도하는 비공개 데이터 배포 수신자의 확률은 두 가지 요소에 기반한다.

- 데이터의 개인정보 보호 및 보안에 관한 데이터 공유 계약에 명시된 통제 범위
- 재식별 공격 수행과 관련하여 수신자의 동기와 능력

이 두 요소는 모두 질적 평가를 수반하며 결과적으로 일반적으로 "낮음", "중간", 또는 "높음" 범위의 값을 갖는다.

개인정보 보호 및 보안 통제

비공개 데이터 배포에 대한 데이터 공유 계약에 명시된 개인정보 보호 및 보안 통제에 따라 수신인이 재식별 공격을 시도할 확률이 다를 수 있다. 개인정보 보호 및 보안 통제 수준이 높을수록 재식별 공격을 시작할 확률이 낮다. 보다 완벽한 통제 목록을 사용할 수 있지만 데이터 공유 계약에서 고려할 수 있는 개인정보 보호 및 보안 통제에는 다음을 포함해야 한다.

- 수신자는 "승인 된 직원” 만이 "알 필요가 있는"근거에 따라 데이터에 접근하고 사용할 수 있도록 허용한다 (임무를 수행해야 하는 경우에만)
- 외부 협력 업체 및 하청 업체를 포함한 모든 직원이 비공개 또는 기밀 유지 계약 (기밀 유지 서약)을 체결한다.
- 데이터는 지정된 보관 기간 후에 폐기된다.
- 데이터는 적절한 통제 또는 사전 승인없이 제3자에게 공개되거나 공유되지

않는다.

- 개인정보 보호 및 보안 정책 및 절차가 마련되고, 모니터링되며, 시행된다.
- 외부 협력 또는 외주 사이트를 포함하여 모든 개인 및/또는 팀 구성원에 대해 필수적이고 지속적인 개인정보 보호, 기밀 유지 및 보안 교육을 수행한다.
- 즉각적인 데이터 보관 담당자에게 서면 통보되는 것을 포함해 개인정보 보호 정책 위반 프로토콜이 시행된다.
- 바이러스 검사 및/또는 악성코드 방지 프로그램이 구현된다.
- 데이터에 접근하는 사람, 시간 및 성격을 문서화하기 위해 감사 추적을 위한 상세한 모니터링 시스템이 시행되다,
- 데이터의 전자 전송이 필요한 경우, 암호화 된 프로토콜이 사용된다.
- 공개된 정보를 보유한 컴퓨터 및 파일은 조합 잠금 도어 또는 스마트 카드 도어의 출입으로 보호되는 객실에 보안 설정된 방법으로 보관되며, 잠긴 스토리지 캐비닛에 저장된 종이 파일과 함께 보관된다.

동기와 능력

수신자가 재식별 공격을 시도 할 확률을 결정할 때 고려해야 할 추가 요소는 동기 및 능력이다. 데이터 세트에서 하나 이상의 개인을 재식별하는 것과 관련해 수신자가 더 동기가 부여되고 능력이 높을수록 재식별 공격을 시작할 확률이 높아진다. 동기와 능력을 평가할 때 다음을 고려해야 한다.

- 수신자가 사고없이 과거에 기관과 일했는지 여부.
- 수신자가 하나 이상의 개인을 재식별 하려고 시도할 이유가 금전적으로 또는 다른 방법으로 존재하는지 여부
- 수신자가 임의의 재식별을 시도 할 기술적 전문 지식 및/또는 재정 자원을 보유하고 있는지 여부
- 수신자가 하나 이상의 개인을 다시 식별하기 위해 다른 개인 데이터베이스 또는 데이터에 연결될 수 있는 데이터 세트에 접근할 수 있는지 여부

재 식별 공격 확률

데이터 공유 계약의 개인정보 보호 및 보안 통제 수준과 받는 사람의 동기와 능력에 따라 내부자가 고의적으로 재식별 공격을 시도할 확률을 추정 할 수 있다. 다음 표는 비공개 데이터 세트에 대해 재식별 공격이 시작될 확률에 대한 허용 가능한 추정치를

결정할 때 가이드 라인으로 사용될 수 있다.

<표 8-2> 재식별 공격 확률

개인정보 및 보안 통제	동기 및 능력	재식별 공격 확률
높음	낮음	0.05
	보통	0.1
	높음	0.2
보통	낮음	0.2
	보통	0.3
	높음	0.4
낮음	낮음	0.4
	보통	0.5
	높음	0.6

공격 2: 지인에 의한 데이터 세트 내에서 개인의 의도치 않은 인식

의도적으로 재 식별 공격을 시도하는 것 외에도 비공개 데이터 배포 수신자는 실수로 하나 이상의 개인을 다시 식별 할 수 있다. 이는 데이터를 분석하는 동안, 친구, 동료, 가족 또는 지인을 인식하는 경우 발생할 수 있다. 그러한 "공격"이 발생할 확률은 임의의 수신자가 데이터 집합에 있는 누군가를 알고있을 확률과 같다. 이를 계산하기 위해 다음 방정식을 사용할 수 있다.

$$1 - (1 - p)^m$$

.

이 방정식에서 p는 데이터 세트에서 논의된 특정 조건 또는 특성을 가진 개인의 모집단에서 백분율이며, m은 개인이 알고 있는 평균적으로 사람들 수이다. 예를 들어, 차량 공유를 이용해 일하러 가는 개인에 대한 데이터세트를 선택하자. 방정식은 p와 m의 값에 기초하여 임의의 개인이 차량 공유를 이용해 일하러 가는 사람을 알게 될 확률을 나타낸다.

p의 값은 최근 인구 통계에 의해 결정되어야 한다. 반면에 m의 값은 데이터 세트에서 논의된 상태 또는 특성에 관한 지식이 요구되는 개인과의 관계의 종류에 따라 달라질 수 있다. 친구의 경우 일반적으로 평균값 150과 190 사이에서 m 값을 사용해야 한다.

공격 3: 데이터 유출

비공개 데이터 배포의 경우 고려해야 할 세 번째 공격은 수신자 측에서 데이터 유출 공격이다. 데이터 유출이 수신자의 시설에서 발생하면 외부 공격자가 재식별 공격을 시도한다고 가정해야 한다. 따라서 그러한 공격이 발생할 확률은 수신자의 시설에서 위반이 발생할 확률과 같다. 이 값을 계산하려면 수신자의 해당 산업에서 데이터 유출의 유행에 대해 공개적으로 사용 가능한 데이터를 사용해야 한다.

세미 - 공개 데이터 배포

세미 공개 데이터 배포에 대한 가능한 재식별 공격은 비공개 데이터 배포의 재식별 공격과 동일하게 간주 될 수 있습니다. 따라서 세미 공개 데이터 배포의 상황에 대한 위험을 측정하려면 비공개 데이터 배포와 동일한 방법 및 방정식을 한 가지 조정을 통해 사용해야 한다. "공격 1: 의도적 내부자 위협"과 관련하여 수신자는 동기 및 능력이 뛰어나며 낮은 프라이버시 및 보안 통제를 가정해야 한다. 이는 세미-공개 데이터 배포는 누구나 다운로드 할 수 있으며 제공 할 수 있는 보호 수준 면에서 제한적이기 때문이다.
이용 조건 약관을 개발할 때 최소한 다음 사항을 수신인이 금지하는 다음 조항을 포함해야 한다.

- 데이터 세트에서 개인을 재식별하려는 시도.
- 외부 데이터 세트 또는 정보에 연결
- 허가없이 데이터 세트 공유

8.6 전체 위험 계산

데이터 위험 및 컨택스트 위험이 측정되면 재식별의 전체 위험을 계산할 수 있다. 전반적인 위험은 데이터 위험에 컨텍스트 위험을 곱한 것과 같다.

전체 위험 = 데이터 위험 x 컨택스트 위험

전체 위험은 공격이 시작된 경우 하나 이상의 행이 재식별 될 확률과 같다. 예를 들어 데이터 집합의 위험가 0.2이고 컨텍스트 위험도가 0.5 인 경우 데이터 집합의 전체 위험은 0.1이다.

8.7 데이터 비식별화

데이터 세트에 대해, 식별 가능한 모든 정보를 제거해야 한다. 데이터 집합의 값은 개인을 식별하는 정보를 제거하거나 정보를 단독으로 또는 다른 정보와 함께 사용하여 개인

을 식별 할 수 있다고 합리적으로 기대할 수 있는 다양한 방법으로 변형 되어야 한다. 식별자의 유형 및 특성에 따라 다른 기법이 적용될 수 있습니다. 식별 가능한 정보를 제거하려면 다음을 수행해야 한다.

- 직접 식별자 마스크
- 등가 클래스 크기 수정
- 전체 위험도가 재식별 위험 임계치 보다 작거나 같음을 보장.

직접 식별자 마스크

직접 식별자로 분류 된 변수는 위에서 언급했듯이 일반적으로 연구 목적으로 유용하지 않기 때문에 데이터 분석에 사용되지 않는다. 이 때문에 가장 간단하고 가장 개인 정보를 보호 할 수 있는 방법은 직접 식별 변수의 열을 제거하여 데이터 집합에서 값을 표시하지 않는 것이다.
그러나 연구의 성격에 따라 관련 개인에게 연락하여 결과를 알려줄 필요가 있을 수 있다. 이러한 경우 직접 식별 변수는 다음과 같은 다른 마스킹 기술을 사용하여 변형되어야 한다.

- 값을 가명으로 대체하고 연결 데이터베이스를 안전한 위치에 유지
- 값을 암호화하고 암호화 키를 안전한 장소에 저장

개인을 식별하기 위해 직접적으로 변수를 식별하는 것이 이용 될 수 있기 때문에 그러한 변환을 수행 할 때에는 최대한 주의를 기울여야 한다. 직접 식별 변수가 부적절하게 또는 안전하지 않은 방식으로 변환되는 경우, 공격자는 많은 수의 개인을 재 식별 할 수 있다.
예를 들어 가명을 만드는 일반적인 방법은 일방향 해시 함수를 사용하여 직접 식별 할 수 있는 변수의 값을 되돌릴 수 없는 코드로 변환하는 것이다. 그러나 이 기술은 변수의 가능한 총 개수가 적으면 적당한 공격자가 합리적인 시간 내에 가능한 모든 변수 값의 해시 값을 계산할 수 있는 만큼 충분히 작으면 무차별 공격에 취약 할 수 있다. 해시 된 값과 원래 값의 역방향 조회 테이블을 만든다. 이러한 공격으로부터 보호하려면 일방향 해시 함수의 입력에 난수 데이터를 항상 추가하고 안전한 위치에 연결 데이터베이스와 함께이 "salt" 또는 "키" 값을 유지해야 한다.

등가 클래스의 크기 수정

데이터 세트가 식별 해제 된 것으로 간주되기 위해서는 재식별의 전체 위험이 재식별 위험 임계 값보다 작거나 같아야 한다. 전체 위험이 재식별 위험 임계 값보다 큰 경우, 데이터 위험을 줄이기 위해 데이터 세트의 등가 클래스의 크기를 수정해야 한다.
해당 유사식별자의 값에 따라 데이터 집합에 서로 다른 크기의 등가 클래스를 가질 수 있다. 비식별은 데이터 집합의 등가 클래스의 크기를 수정하기 위해 여러 가지 방법으로 간접 식별자의 값을 변환하는 것과 관련된다. 이 작업을 수행하는 두 가지 기법은 일반

화 및 억제이다.

일반화

일반화는 보다 일반적인 값을 산출하기 위해 값에서 정밀도를 제거하는 프로세스이다. 그것은 증가하는 양으로 적용될 수 있다. 예를 들어, 전체 날짜는 월 및 일로 일반화 될 수 있으며, 다시 년으로 일반화 될 수 있고, 5년 간격, 10년 간격 등으로 일반화 될 수 있다.

일반화를 사용할 때는 변수의 모든 행에 적용해야 한다. 또한 변수 내에서 사용되는 일반화 집합이 일정하고 겹치지 않도록 해야 한다. 예를 들어, 5 세 연령 간격의 일정한 세트는 10-14, 15-19, 20-24, 25-29, 30-34 등이 될 수 있다.

여기에는 예외가 하나 있다. 연속 변수의 경우 상단 또는 하단 범위 값에 절단 점을 도입하여 특이 치를 위한 범주를 만들 수 있다. 예를 들어, 개인의 나이는 90 세 이상의 개인에 대해 "90+"의 모든 카테고리를 만들어 년을 일반화 할 수 있다. 이 일반화 기술은 컷 포인트가 만들어지는 위치에 따라 상단 또는 하단 코드라고 한다.

삭제

삭제는 데이터 세트에서 값을 제거하는 프로세스이다. 간접 식별자의 모든 행에 적용되는 일반화와 달리 억제는 단일 행에만 영향을 미친다. 간접 식별자 값의 억제는 다른 레벨에서 발생할 수 있다. 예를 들어 전체 행, 행의 간접 식별자 집합 또는 개별 셀만 제거하는 작업이 포함될 수 있다. 데이터에서 제거된 정보가 적을수록 데이터 세트 활용도가 더 커질 수 있지만, 등가 클래스는 적절한 크기가 되도록 간접 식별자의 값을 억제할 때는 행에 있는 전체 행 또는 간접 식별자 세트를 제거해야 한다.

전체 위험도가 재발견 위험도보다 낮거나 같음 확인

데이터 세트의 등가 클래스의 크기가 수정 된 경우 재식별의 전체 위험을 다시 계산하여 재식별 위험 임계 값과 비교해야 한다. 데이터 세트가 비식별 된 것으로 간주되기 위해서는 전체 위험이 재식별 위험 임계 값보다 작거나 같도록 데이터 위험을 충분히 낮춰야 한다.

8.8 데이터 유용성 평가

데이터 세트에 적용된 비식별 양과 결과 정보의 유용성 사이에는 트레이드오프가 존재한다. 간접 식별자로 분류되는 변수가 일반화 및 삭제와 같은 기법을 사용하여 식별되지 않는 경우 데이터 집합의 효용에 상응하는 손실 가능성이 높아진다.

일반화 및 억제가 데이터 세트에 적용되어 재식별의 전체 위험이 재식별 위험 임계치보다 작거나 같음을 보장하지만, 이러한 비식별 기술은 이를 달성하기 위해 다양한 방식 및 조합으로 적용될 수 있다. 예를 들어, 한 접근법은 일반화에 더 의존하여 등가 클래스의 크기를 늘리기 위해 범주의 정밀도를 낮출 수 있다. 또 다른 접근법은 너무 작은 등가 클래스를 증가하기 위해 변수의 행 또는 셀을 제거하는 데 더 많이 의존 할 수 있다. 데이터 세트의 특성에 따라, 상이한 애플리케이션 및/또는 일반화 및 삭제의 조합은

개인의 프라이버시를 보호하면서 정보에 더 많은 유용성을 보존 할 수 있다.
일반적으로 데이터 세트의 행 중 5 % 이상이 이미 삭제 형식을 가지고 있지 않으면 일반화 적용 전에 억제를 고려해야 한다 삭제는 정밀도를 떨어뜨리는 일반화와 달리 정보를 단일 행에서 제거하므로 데이터 세트의 삭제를 비식별화의 시작으로 고려할 수도 있다.
일반화 및 억제 기술을 새로운 방식으로 적용 및/또는 결합하는 것은 재 식별의 전반적인 위험이 위험 기준보다 작거나 같은지 확인하면서 더 높은 유용성의 데이터 세트를 생성 할 수 있다.

8.9 프로세스 문서화

개인 정보가 포함된 데이터 세트를 비식별하는 각 시도는 동일한 단계를 따라야 하고, 동일한 문제를 평가해야 한다. 그러나 비식별화 량과 종류를 결정하는 변수 및 값, 분석은 데이터 배포마다 다를 수 있다. 개인 정보 비식별과 관련된 복잡성 및 문제점을 해결하는 데 도움이 되도록 프로세스 및 결과를 문서화한 보고서를 작성하는 것이 좋다. 이 모범 사례에서 다음과 같은 여러 이점이 존재한다.

- 개인정보 유출에 시에 중요한 불평이 있는 경우 중요한 주의 의무와 준수 증거를 입증하는 능력
- 모범 사례가 준수되고 있다는 확신 (개인, 다른 기관, 파트너 및 자신의 경영진)
- 조직의 정보 관리 관행에 대한 투명성, 인식, 이해 및 신뢰의 증대

부 록 Ⅰ-1

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

지식재산권 확약서 정보

해당 사항 없음

※ 상기 기재된 지식재산권 확약서 이외에도 본 표준이 발간된 후 접수된 확약서가 있을수 있으니, TTA 웹사이트에서 확인하시기 바랍니다.

부 록 Ⅰ-2

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

시험인증 관련 사항

해당 사항 없음

부 록 Ⅰ-3

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

본 표준의 연계(family) 표준

해당 사항 없음

부 록 Ⅰ-4

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

참고 문헌

[1] ISO/IEC 29100:2011, Information technology - Security techniques - Privacy framework

부 록 Ⅰ-5

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

영문표준 해설서

해당 사항 없음

부 록 Ⅰ-6

(본 부록은 표준을 보충하기 위한 내용으로 표준의 일부는 아님)

표준의 이력

판수	채택일	표준번호	내용	담당 위원회
제 1 판	2019.12.31	제정 TTAx.xx-xx.xxxx		개인정보보호/ID관리 및 블록체인보안 (PG502)