데이터 비식별화를 이용한 빅데이터 통합

Big Data Integration using Data De-identification

저자 (Authors)
김승환, 전성해
Seungwhan Kim, Sunghae Jun

출처 (Source)
한국지능시스템학회 논문지 29(3), 2019.6, 235-241(7 pages)
Journal of Korean Institute of Intelligent Systems 29(3), 2019.6, 235-241(7 pages)

발행처 (Publisher)
한국지능시스템학회
Korean Institute of Intelligent Systems

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08743923

APA Style
김승환, 전성해 (2019). 데이터 비식별화를 이용한 빅데이터 통합. 한국지능시스템학회 논문지, 29(3), 235-241

Journal of Korean Institute of Intelligent Systems
Vol. 29, No. 3, June 2019, pp. 235–241
http://dx.doi.org/10.5391/JKIIS.2019.29.3.235

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

데이터 비식별화를 이용한 빅데이터 통합

Big Data Integration using Data De-identification

김승환*· 전성해**†
Seungwhan Kim and Sunghae Jun†

*인하대학교 IT융합기술연구소, **청주대학교 빅데이터통계학과
*Division of Software Convergence, Inha University
**Department of Big Data and Statistics, Cheongju University

요 약

여러 곳에 흩어져 있는 방대한 데이터를 통합하여 빅데이터 플랫폼을 구축하고 분석하려는 시도가 공공부문에서 민간부문에 이르기까지 활발하게 진행되고 있다. 공공 빅데이터 플랫폼은 국가발전과 국민 삶의 질을 높이기 위하여 구축되고 민간 빅데이터 플랫폼은 고객정보를 마케팅에 활용하여 기업의 이익추구와 성장을 위하여 도입되고 있다. 빅데이터 플랫폼 구축을 위하여 공공기관 및 기업이 보유한 데이터들이 서로 통합되는 과정에서 개인정보가 개인의 동의 없이 조금이라도 공개되는 것은 불법이다. 이와 같은 경우에 비식별화 처리기법을 통하여 개인정보가 나타나지 않도록 가공한 후 빅데이터 플랫폼 구축작업이 진행되지만 이 과정에서 정보 손실이 발생한다. 즉, 데이터를 제공하는 입장에서는 개인정보 보호를 위해 비식별 처리 수준을 높게 하길 원하고 데이터를 제공받는 입장에서는 예측력 높은 분석모형을 만들기 위하여 정보손실이 작은 형태의 데이터를 원한다. 이와 같은 이해관계의 상충으로 인하여 비식별 처리 데이터의 활용 자체가 불가능할 경우도 발생한다. 본 논문에서는 최적 절단값을 이용하여 빅데이터 통합 플랫폼 구축을 위한 데이터 비식별 과정에서 데이터를 제공하는 입장과 받는 입장을 동시에 만족시킬 수 있는 방법을 제안한다. 제안 방법의 성능평가를 위하여 UCI 머신러닝 저장소의 데이터를 이용한 실험을 수행한다.

키워드 : 비식별화, 빅데이터 플랫폼, 개인정보, k 익명성, 최적 절단값

Abstract

Attempts to build and analyze big data platforms by integrating vast amounts of data scattered across multiple locations have been actively conducted from the public sector to the private sector. The Public big data platform is established to improve the national development and quality of life of the people, and the private big data platform is being introduced for the pursuit and profit of the enterprise by utilizing customer information for marketing. It is illegal for personal information to be disclosed without any personal consent in the process of integrating data held by public organizations and companies to build big data platform. In such a case, the big data platform construction work is performed after the personal information is not displayed through the de-identification processing technique, but information loss occurs in this process. That is, from the viewpoint of providing data, in order to protect personal information, it is desired to increase the level of de-identification processing. In the case of receiving data, in order to make a predictive analysis model. Such conflicts of interest may result in non-use of de-identified data. In this paper, we propose a method to satisfy both the position and the receiving position in data de-identification process for constructing big data platform using optimal cutoff value. Experiments using data from UCI machine learning repository are performed to evaluate the performance of the proposed method.

Key Words : De-identification, big data plotform, privacy, k anonymity, optimal cutoff value

Received: Feb. 9, 2019
Revised: Jun. 13, 2019
Accepted: Jun. 13, 2019
†Corresponding authors
shjun@cju.ac.kr

1. 서 론

정부 및 공공기관 뿐만 아니라 민간 기업에서도 빅데이터의 활용에 대한 관심이 높아지고 있다 [1,2]. 빅데이터의 활용을 위하여 우선적으로 필요한 것은 여러 곳에 흩어져 있는 데이터 원천(sources)을 통합하여 빅데이터 플랫폼(plotform)을 구축하는 것이다 [3]. 이 과정에서 신중하게 고려해야 할 사항들 중 하나는 개인정보 보호이다. 개인정보는 빅데이터의 직접적인 활용목적이 아님에도 불구하고 개인정보 보호법에 의해 제약을 받는다 [4]. 따라서 빅데이터 플랫폼 구축과정에서 데이터를 제공하는 입장에서는 개인정보 보호를 위하여 특정 개인의 인식을 불가능하게 하기 위한 다양한 익명화(anonymity) 방법을 사용한다 [5]. 비식별 처리(de-identification)는 가장 많이 사용되는 개인정보 보호 기법이다 [6]. 비식별 처리는 데이터 내에 개인을 식별할 수 있는 정보가 있는 경우, 일부 혹은 전부를 삭제하거나 가공

처리함으로써 다른 정보와 결합하여도 특정 개인을 식별하기 어렵도록 만드는 방법이다. 현재 우리나라는 2016년 정부에서 제시한 "개인정보 비식별 조치 가이드라인: 비식별 조치 기준 및 지원, 관리체계 안내"를 통하여 개인정보 비식별 처리가 이루어지고 있다. 비식별 처리 기법은 데이터 내에 개인 식별이 가능한 속성(attributes)을 식별자(identifier), 준식별자(quasi-identifier), 일반속성, 민감속성으로 구분한다. 주민등록번호, 성명과 같이 그 자체만으로도 누구인지 식별 가능한 속성이 식별자이다. 이밖에 주소, 성별, 직업 등 개인에 대한 직접적인 식별은 불가능하지만 이들 속성이 결합되면 개인에 대한 식별 가능성이 가능해지는 속성을 준식별자라고 한다. 일반적으로 식별자는 삭제를 원칙으로 하고 준식별자는 삭제, 마스킹(masking), 범주화 등의 가공을 통하여 식별 가능성을 낮춘다. 이와 같은 비식별 처리는 개인 식별 가능성은 낮출 수 있지만 대신 구축된 빅데이터를 분석하는 입장에서는 정보 손실이 발생하여 데이터의 가치가 낮아진다. 그러므로 데이터를 제공하는 입장과 데이터를 제공받아 빅데이터를 분석하는 입장에서 개인정보 식별 가능성의 제거와 정보 손실 사이에서 타협점을 찾는 것은 빅데이터 플랫폼 구축과정에서 중요한 작업이다. 하지만 실제 빅데이터 플랫폼 구축과정에서 이와 같은 비식별화 처리를 위한 타협이 거의 이루어지고 있지 못한 실정이다. 대부분 데이터를 제공하는 입장에서 법적인 책임을 회피하기 위하여 강력한 비식별화가 이루어진다. 이와 같은 문제점을 해결하기 위하여 본 논문에서는 빅데이터 플랫폼 구축과정에서 실무적으로 비식별화 처리가 진행될 때 개인정보를 만족할만한 수준으로 유지하면서 구축된 빅데이터의 분석모형에 대한 예측성능을 높이는 방법에 대하여 연구한다. 제안방법의 타당성과 실제 적용 가능성을 보이기 위하여 UCI 머신러닝 저장소(UC Irvine machine learning repository)로부터 객관적인 데이터를 이용한 실험을 수행한다 [7]. 2절에서는 개인정보 보호와 정보손실에 대하여 알아보고, 제안방법인 최적 절단값을 이용한 데이터 비식별화 처리방법은 3절에서 설명한다. 실험 및 결과와 본 논문의 결론 및 향후 연구과제는 각각 4절과 5절에서 다룬다.

2. k-익명성과 정보손실

k-익명성(k-anonymity)은 특정 개인을 식별할 수 있는 가능성을 1/k 이하로 낮추는 방법으로 2016년 정부에서 제공한 개인정보 비식별 가이드라인에 따라 현재 국내에서 가장 많이 사용되는 비식별 처리 기법이다. k-익명성은 데이터 셋 내에 해당 개인의 준식별자 조합과 같은 사람의 수가 k명 이상 중복 되도록 하기 위하여 준식별자에 대한 마스킹, 범주화 등 다양한 방법을 이용하여 이 기준을 만족할 수 있도록 원래의 데이터 셋을 변형한다 [8-10]. 이와 같이 변형된 자료는 각 케이스마다 같은 준식별자를 가진 개인의 수가 k 명 이상이 존재하므로 해당 정보로 특정 개인을 식별할 수 있는 확률은 1/k 보다 작아진다. 이때 주로 사용되는 비식별 처리 기법은 표 1과 같다.

표 1. 비식별 처리 기법

Table 1. De-identification Methods

Method	Explanation	Example
Alias processing	replace individual identifiers with different values	Hong, Kisu->Hong, Kildong
Total processing	convert personal information to statistical values	height 178->170(mean of height)
Removing	delete property	social security number->remove
Categorization	categorize attributes to hide explicit values	Lee, Sunshin, age 36->Lee, 30s
Masking	replace with whitespace, '*', etc.	HanKook high school -> Han* high school

표 1에서 마스킹은 '홍길동'을 '홍**'으로 바꾸는 것이고 범주화는 '35세'를 '30대'로 바꾸는 것이다. 이 과정에서 데이터분석을 위한 정보의 손실이 발생이다. 일반적으로 정부의 개인정보 비식별 가이드라인에서는 k-익명성, l-다양성, t-근접성 변환을 사용하는데 k-익명성이 가장 기본이 되는 변환이다. 따라서 본 논문에서는 k-익명성 변환에 대한 제안방법을 적용한다. 그림 1은 비식별화를 수행하는 오픈소스 프로그램인 ARX 소프트웨어에 의해 원자료를 3-익명성으로 처리한 결과이다 [11,12].

	ZipCode	Age	Salary	Disease
1	47602	22	4	1.위염
2	47677	29	3	1.위궤양
3	47678	27	5	1.위암
4	47605	30	7	2.기관지염
5	47607	32	10	1.위암
6	47673	36	9	2.폐렴
7	47905	43	6	1.위염
8	47909	52	11	2.감기
9	47906	47	8	2.기관지염

⇩

	ZipCode	Age	Salary	Disease
1	476**	[22, 30[	4	1.위염
2	476**	[22, 30[	5	1.위암
3	476**	[22, 30[	3	1.위궤양
4	476**	[30, 40[	7	2.기관지염
5	476**	[30, 40[	10	1.위암
6	476**	[30, 40[	9	2.폐렴
7	479**	>=40	6	1.위염
8	479**	>=40	8	2.기관지염
9	479**	>=40	11	2.감기

그림 1. 3-익명성 처리 결과

Fig. 1. Result of 3-anonymity processing

그림 1에서 ZipCode와 Age는 준식별자이고, Salary와 Disease는 민감속성이다. 준식별자에 대해 3-익명성 처리를 하면 같은 준식별자를 가지는 케이스가 3개 이상이어서 해당 속성을 가진 사람을 식별할 확률은 1/3보다 작아진다. k-익명성을 만족하도록 변환을 수행할 때, k가

커질수록 정보 손실도 커진다. 일반적으로 정보손실 측정은 엔트로피(entropy) 변화로 측정한다 [5,8-10,13]. 표 2는 Age 속성에 대하여 범주화 변환을 수행한 결과이다.

표 2. 나이 범주화 처리

Table 2. Grouping of age

Original ages	Ages with categorization
20	20-39
65	60-79
55	40-59
40	40-59
65	60-79
65	60-79

표 2에서 정보 손실을 계산하기 위해 다음의 식 (1)의 엔트로피를 계산한다 [14].

$$entropy(S) = \sum_{i=1}^{c} -p_i \log(p_i) \quad (1)$$

여기서 p_i는 i번째 범주의 상대빈도(확률)를 나타낸다. 원자료의 나이 속성에 대한 엔트로피를 계산하면 다음과 같다.

$$-3\frac{1}{6}log\frac{1}{6}-\frac{3}{6}log\frac{3}{6} = 1.2424$$

같은 식을 이용하여 범주화 변환후 나이 속성의 엔트로피는 다음과 같다.

$$-\frac{1}{6}log\frac{1}{6}-\frac{2}{6}log\frac{2}{6}-\frac{3}{6}log\frac{3}{6} = 1.011$$

따라서 정보손실이 커지면 예측 모형 정확도는 감소할 수 있다. 그러나 정보손실이 실제 통계적 모형 정확도에 어느 정도 영향을 주는지 사전에 알 수 없다. 그러므로 비식별화 처리와 정보손실의 간의 최적 관계를 찾기 이해서는 절단값(cutoff value)를 구해야 한다. 다음 절에서는 본 연구에서 제안하는 최적 절단값을 이용한 데이터 비식별 처리에 대하여 설명한다. 이를 통하여 빅데이터 플랫폼 구축을 위하여 데이터를 제공하는 입장에서의 개인정보 보호와 데이터를 제공받는 입장에서의 예측모형의 성능 향상 사이에서 최적의 균형점(절단값)을 구한다.

3. 최적 절단값을 이용한 데이터 비식별 처리

데이터에 포함된 개인정보의 보호 관점에서는 데이터를 제공할 때 개인정보를 최대한 보호하려고 한다. 개인정보를 보호하기 위하여 개인정보에 대한 익명성(anonymity)을 최대한 높이려고 한다. 반면에 데이터를 제공 받는 입장에서는 개인정보에 대하여 최대한 정확히 알아야 한다. 왜냐하면 서로 흩어져 있는 데이터를 통합하여 제대로 된빅데이터 플랫폼을 구축하기 위해서는 데이터 원천들을 연결해 줄 수 있는 키(key)를 알고 싶은데 이 값이 바로 주민등록번호 등과 같은 대표적인 개인정보이기 때문이다. 또한 개인정보에 해당되는 변수들이 예측모형에서 가장 중요한 예측 변수의 역할을 수행하기 때문이다. 그림 2는 이와 같은 문제점을 해결하기 위하여 본 연구에서 제안하는 방법의 기본적인 개념을 나타내고 있다.

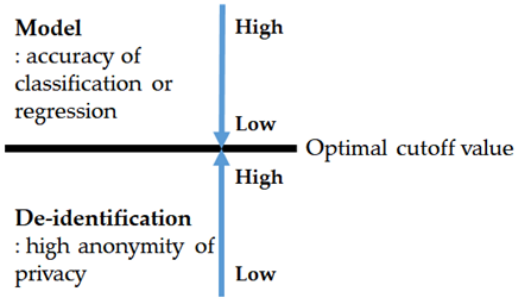


그림 2. 최적 절단값을 이용한 비식별화

Fig. 2. De-identification using optimal cutoff value

분류(classification)와 회귀(regression) 모형의 성능을 나타내는 오분류율(misclassification rate)과 평균제곱오차(mean squared error; MSE)는 모두 작을수록 좋은 성능을 나타내게 된다. 반면에 개인정보의 보호를 위한 익명성은 클수록 좋다. 그림 2에서와 같이 개인정보의 익명성과 모형의 성능은 서로 반비례 한다. 따라서 모형의 성능과 개인정보의 익명성 사이의 최적 절단값은 실제 구축되는 빅데이터 플랫폼의 사용자가 최종적으로 결정할 수 있다. 그림 3은 비식별화 과정을 통하여 구축되는 빅데이터 플랫폼을 나타낸다.

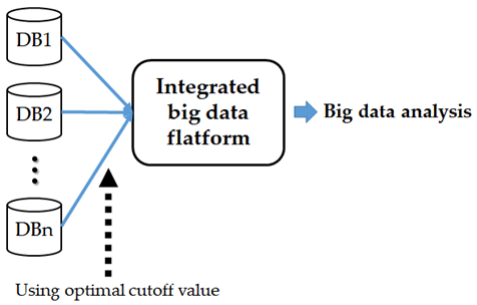


그림 3. 빅데이터 플랫폼

Fig. 3. Big data plotform

공공 또는 민간 기관에서는 효율적인 빅데이터 분석을 위하여 여러 곳에 흩어져 있는 빅데이터를 하나의 플랫폼으로 통합하는 작업을 수행한다. 이 때 최적의 절단값을 사용한다. 본 연구에서 제안하는

방법은 다음과 같은 단계를 통하여 이루어 진다.

Step 1. Determining target performance of model

(1.1) Misclassification rate, α

(1.2) Regression accuracy, MSE

Step 2. Pre-integrating big data sources using $\alpha = \alpha_0$, or MSE=m

(2.1) Performance evaluation of anonymized books with de-identification satisfying α_0 or m through repeated experiments

(2.2) Adjusting k anonymity value for the final satisfactory de-identification

Step 3. Constructing big data plotform

(3.1) Analyzing big data

(3.2) Applying analytical results to practical problem

4. 실험 및 결과

제안 방법의 성능평가를 위하여 본 논문에서는 UCI 기계학습 저장소 (machine learning repository)에서 제공하는 미국 성인 소득 데이터(adult data set)를 사용하였다. [7]. 이 데이터의 반응변수는 미국 성인의 급여액을 50K 이하와 50K 초과로 구분한 범주형 변수(categorical variable)이고 이를 예측하기 위하여 사용되는 설명변수는 성별(sex), 나이(age), 결혼상태(marital status), 교육기간(education years), 직업 구분(work class) 등 여러 변수들로 이루어진다. 표 3에 실험에 사용되는 데이터의 주요 독립변수에 대한 설명과 식별속성을 정리하였다.

표 3. 미국 성인 소득 데이터셋 주요변수

Table 3. US Adult Income Dataset variables

Independent variable	Explanation	Identification
Sex	gender	semi-identifier
Age	age	semi-identifier
Work class	job	semi-identifier
Marital status	married	semi-identifier
Occupation	rank	general information
Education	period of training	general information
Capital gain	capital gain	general information
Hours per weeks	working hours per week	general information
Capital loss	capital loss	general information

그림 4는 미국 성인소득 원자료의 준식별자 조합에 대한 빈도수를 나타낸 것이다. 준식별자를 조합했을 때, 조합 별 빈도가 1인 유일한 케이스가 대부분이기 때문에 식별가능성이 높은 데이터 셋이 된다.

원자료를 성별, 나이(5세 단위), 결혼상태, 직업구분에 의해 적절한 변환을 수행하고 빈도분석을 수행한 후에 빈도수가 5 미만이 되어 5

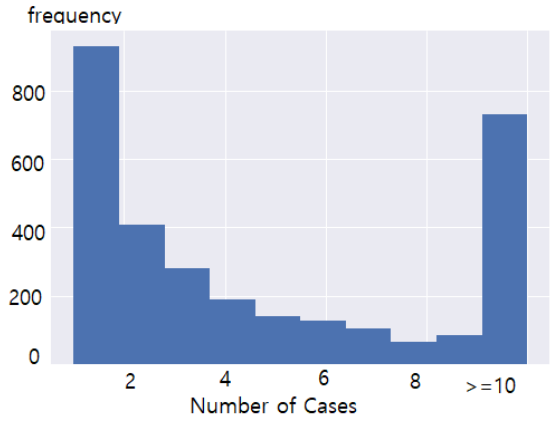


그림 4. 준식별자 조합에 의한 빈도수: 원자료

Fig. 4. Frequency of quasi-identifier combination: original data

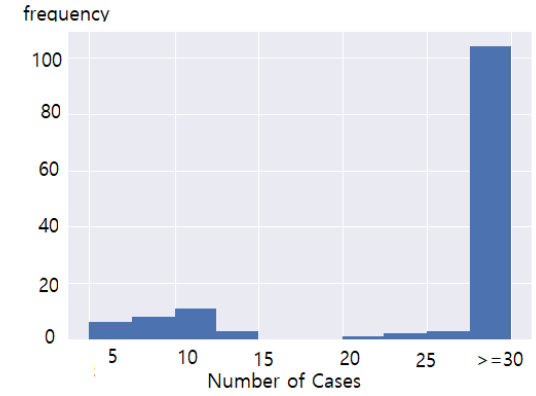


그림 5. 준식별자 조합에 의한 빈도수: 5–익명성 처리 후 데이터

Fig. 5. Frequency of quasi-identifier combination: data after 5-anonymity transformation

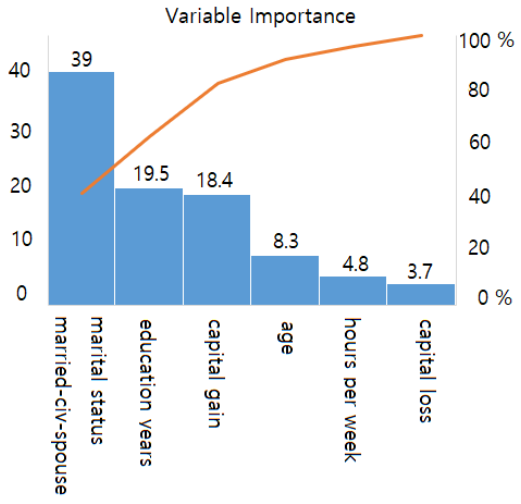


그림 6. 변수 중요도

Fig. 6. Variable Importance

익명성을 만족하지 못하는 케이스(case)는 그림 5처럼 데이터에서

제거하여야 한다.

그림 6의 결과로 소득을 예측하는데 가장 중요변수는 결혼상태(시민권배우자), 교육년수, 자본이득, 나이, 근무시간, 자본손실 등의 순서임을 알 수 있다. 그림 6에 표시되지 않은 변수는 중요도가 1% 미만인 경우다. 따라서 이에 해당하는 성별을 비식별 처리해도 모형성능에는 작은 영향을 미칠 뿐이다. 반면, 결혼 상태, 나이는 비식별 처리할 경우, 예측 정확도에 어느 정도 영향을 주게 된다. 이 데이터 셋을 비식별 처리하고 이에 따른 예측력의 변화를 확인하기 위하여 준식별자로 성별, 연령, 직업구분, 결혼상태를 사용하였다. 준식별자에 대해 표 4와 같이 비식별 처리를 수행한 후, 예측 정확도가 어떻게 변하는지 알아보았다.

표 4. 비식별 변환

Table 4. data de-identification

Semi-identifier	Original data	Transformation
Sex	M, F	*
Age	original age original age	5 years old 10 years old
Work class	private, self-emp-not-inc, self-emp-inc, federal-gov, local-gov, state-gov, without-pay, never-worked	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Marital status 1	divorced, married-AF-spouse, married-civ-spouse, married-spouse-absent, never-married, separated, widowed	Married-civ-spouse,Married-AF-spouse,Married-spouse-absent, Widowed ,Divorced, Never-married, Separated
Marital status 2	divorced, married-AF-spouse, married-civ-spouse, married-spouse-absent, never-married, separated, widowed	Married-civ-spouse,Married-AF-spouse,Married-spouse-absent ,Divorced, Widowed Never-married, Separated

{}는 같은 그룹을 의미함

표 4에서 성별 처리는 성별 변수를 모형에서 제거함을 의미하고 나이 처리는 모형에 원래 나이 대신 5세 단위 혹은 10세 단위 데이터로 변환하는 것을 의미한다. 또한, work class와 marital status1,2는 그룹화를 통해 익명성을 강화하는 변환이다. 표 5는 준식별자에 대한 변환을 수행하고 변환된 데이터를 이용하여 예측모형을 만든 결과다.

성별을 변환했을 때, 정확도는 변화하지 않았다. 이는 예측모형에서 성별이 중요하지 않은 변수이기 때문이다. 다음으로 나이 변환에서는 의미 있는 결과가 나타났다. 5세 단위 변환에서는 오히려 정확도가 소폭 증가하였고, 10세 단위 변환에서는 정확도 감소가 일어났다. 이와 같은

표 5. 변환에 의한 예측 정확도 감소량

Table 5. Reduced accuracy by anonymization

Transformation	Prediction accuracy	Reduced accuracy
Sex	85.96%	0.00%
Age(5 years old) Age(10 years old)	85.99% 85.84%	-0.03% 0.12%
Work class	85.94%	0.02%
Marital status 1 Marital status 2	85.96% 85.64%	0.00% 0.31%
Convert all (5 years old, (marital status 1) 5-anonymity(×)	86.19%	-0.23%
Convert all (10 years old, marital status 2)	85.55%	0.41%
Convert all (5 years old, (marital status 1) 5-anonymity(O)	93.04%	-7.08%

결과가 나타난 이유는 연봉 예측에서는 1세 단위 보다는 5세 단위가 더 정확할 수 있기 때문이다. 10세 단위 변환은 너무 나이 변환이 크게 되어 연봉에 대한 설명력을 훼손했기 때문이다. 이는 적절한 범주화는 정확도를 증가시킬 수 있다는 의미이다. 즉, 자나친 범주화는 정확도를 감소시키게 된다. marital status 변환에서 변환 방법 1, 2에 따라 정확도 차이가 나타났다. 이는 marital status가 중요한 변수이기 때문에 변환방법에 따라 정확도의 차이가 나타난다. 모든 데이터에 대하여 나이 5세 단위, 결혼상태 변환 1로 처리 했을 때 정확도가 증가하였고, 다시 모든 데이터에 대하여 나이 10세 단위, 결혼상태 변환 2로 처리 했을 때 정확도가 감소하였다. 이를 통하여 비식별 처리가 일반적으로 정확도를 감소시키는 것은 맞지만, 정확도를 크게 감소시키지 않거나 오히려 정확도를 증가시키는 변환도 가능함을 확인할 수 있었다. 이와 같은 결과로 알 수 있는 것은 예측모형에 중요하지 않은 변수를 비식별 처리할 경우, 정확도에 영향을 주지 않는다는 것과 중요한 변수를 비식별 처리할 경우, 적절한 그룹화는 오히려 모형 예측력을 증가시킬 수 있다는 것이다. 또한 비식별 변환 후 k-익명성을 만족하지 못하는 케이스는 강제 삭제하는데 변환 후, k-익명성을 만족하지 못하는 케이스를 삭제한 후, 예측 결과는 93.04%로 큰 폭으로 예측력이 상승하였다. 이는 변환 후, k-익명성을 만족하지 못하는 케이스들은 예측모형에서 극단값에 해당하는데 이들을 제거하니 모형의 예측력이 증가한 결과로 해석된다. 일반적으로 극단값을 제거할 경우, 모형 예측력이 상승할 수도 있지만, 반대로 감소할 수도 있다.

미국 성인소득 예에서 비식별 처리 방법에 따라 예측 정확도가 달라질 수 있음을 알 수 있었고 이러한 문제를 극복하기 위해서 아래와 같은 3가지 방법을 도출할 수 있었다.

첫째 비식별 처리 전 단계에서 원자료의 준식별자를 구분하고 준식별자의 식별 중요도를 설정한다. 성별과 직장명은 모두 준식별자이고 성별 보다는 직장명의 식별성이 더 크기 때문에 식별성의 크기는 속성 집합 크기로 구하면 된다. 예를 들어 성별 속성의 집합크기는 남자 집합 크기와 여자 집합크기 중 최소값이고 직장명 속성의 집합 크기는 여러 직장명 집합크기 중 최소값이다. 집합 크기가 작은 속성일수록 식별성이 큼을 알 수 있다. 다음으로 의사결정나무 등 예측모형을 만들어 독립변수의 중요도를 구한다. 식별성이 크고 예측 중요도가 낮은 변수가 우선 비식별 처리 대상이 된다. 이 기준으로 비식별 처리를 진행하고 k-익명성 등 비식별 기준을 만족하면 더 이상의 변환 없이 데이터를 활용하면 된다. 다음으로 준식별자 중 예측모형에 중요한 변수이고, 이에 대한 적절한 비식별 처리가 필요하다면, 적절한 범주화를 통해 비식별 처리를 해야 한다. 적절한 범주화는 그룹 내 변동을 최소화하고 그룹 간 변동을 최대화하는 범주화가 되어야 한다. 본 논문의 성인 소득 데이터에서 나이에 대한 범주화는 5세 단위가 적절한 범주화임을 알 수 있다. 5세 단위는 그룹 내 변동이 작고, 그룹 간 변동이 큰 범주화이고 결혼 상태에서 'married-civ-spouse'는 매우 중요한 범주이므로 독립적인 범주로 만들고 나머지 범주는 비슷한 범주끼리 묶어 그룹 내 변동을 최소화하는 그룹화를 수행해야 한다. 마지막으로 k-익명성을 만족시키는 변환은 준식별자를 변환하는 것도 가능하지만, k-익명성을 만족하지 않은 케이스를 제거하는 것도 하나의 방법이 된다. 준식별자 변환과 k-익명성을 만족하지 않는 케이스를 제거할 것인지 선택에 따라 모형의 정확도가 달라질 수 있기 때문에 케이스의 수가 충분하다면 일반적으로 케이스 제거를 적극 활용하는 것이 효율적 방법이 된다.

5. 결론 및 향후 연구

비식별 처리에 관한 상당히 많은 논문이나 문헌들이 정보보호 관점에서 기술되어 있다. 이러한 관점은 정보 보호 목적은 달성할 수 있지만 정작 활용 관점에서 어떤 가치가 있는지에 대한 고려가 약할 수 있다. 많은 기업에서 비식별 조치를 수행할 때, 이러한 어려움을 겪고 있다. k 익명성 등 가이드라인 권고 사항을 맞출 수 있는 쉬운 방법은 데이터 삭제, 그룹화 등을 만족할 때까지 처리하는 것이다. 하지만, 이러한 비식별 처리는 데이터의 품질을 많이 훼손하므로 좋은 방법이 아닐 수 있다.

본 논문에서는 비식별 처리를 진행할 때, 예측모형의 정확도 손실을 최소로 하기 위한 몇 가지 방법을 제안하였다. 이를 통해 비식별 조치가 익명성을 보장하고 동시에 데이터 가치도 유지할 수 있는 방안이 되길 기대한다.

References

[1] K. I. Kim, J. S. Jeong, G. K. Park, "Assessment of External Force Acting on Ship Using Big Data in Maritime Traffic," *Journal of The Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 379-384, 2013.

[2] K. I. Kim, K. M. Lee, "Big Data Analysis for External Forces Acting on Ship with MapReduce Processing," *Journal of The Korean Institute of Intelligent Systems*, Vol. 28, No. 2, pp. 146-151, 2018.

[3] S. Jun, "A Big Data Preprocessing using Statistical Text Mining," *Journal of The Korean Institute of Intelligent Systems*, Vol. 25, No. 5, pp. 470-476, 2015.

[4] Y. N. Shin, M. G. Chun, "Personal Information Protection for Biometric Verification based TeleHealth Services," *Journal of The Korean Institute of Intelligent Systems*, Vol. 20, No. 5, pp. 659-664, 2010.

[5] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 10, No. 5, pp. 557-570, 2002.

[6] S. Garfinkel, "De-Identification of Personal Information," *NISTIR*, 8053, 2015.

[7] UCI ML Repository, UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml, 2018.

[8] C. Caballero-Gil, J. Molina-Gil, J. Hernández-Serrano, O. León, M. Soriano-Ibañez, "Providing k-anonymity and revocation in ubiquitous VANETs," *Ad Hoc Networks*, Vol. 36, Part 2, pp. 482-494, 2016.

[9] A. Gionis, T. Tassa, "k-Anonymization with Minimal Loss of Information," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, Iss. 2, pp. 206-219, 2009.

[10] N. Man, X. Li, K. Wang, "A Privacy Protection Model Based On K-Anonymity," *Advances in Engineering Research*, Vol. 153, pp. 15-19, 2018.

[11] ARX, Data Anonymization Tool, https://arx.deidentifier.org, 2018.

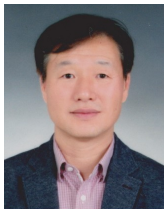
[12] F. Prasser F. Kohlmayer, "Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool," *Medical Data Privacy Handbook*, Springer, pp 111-148, 2015.

[13] H. R. Kim, "De-identification and Privacy protection for Statistical Purpose," *Journal of The Korean Official Statistics*, Special Issue, pp. 35-51, 2016.

[14] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Third Edition, Essex, UK: Pearson, 2014.

[15] Python, https://www.python.org, 2018.
[16] Anaconda, https://www.anaconda.com/distribution, 2018.

저 자 소 개



김승환(Seungwhoun Kim)

1989년 : 충북대학교 토목공학 공학사
1991년 : 인하대학교 통계학과 이학석사
1997년 : 인하대학교 통계학과 이학박사
2014년 : SK 에너지, SK M&C 재직
2015년~현재 : 인하대학교 소프트웨어 융합공학연계전공 연구교수

관심분야 : Big Data, Statistical Algorithm, Machine Learning, Software Convergence
Phone : +82-32-860-8423
E-mail : swkim4610@inha.ac.kr

전성해(Sunghae Jun)

1993년 : 인하대학교 통계학과 이학사
1996년 : 인하대학교 통계학과 이학석사
2001년 : 인하대학교 통계학과 이학박사
2007년 : 서강대학교 컴퓨터공학과 공학박사
2013년 : 고려대학교 정보경영공학과 공학박사
2003년~현재 : 청주대학교 소프트웨어융합학부 빅데이터통계학전공 교수

관심분야 : Artificial Intelligence, Bayesian Statistics, Data Science
Phone : +82-43-229-8205
E-mail : shjun@cju.ac.kr