결합정보의 안전성 확보를 위한 부트스트랩 기반 특이정보 판단 방법론 연구

이강원, 성민경, 한주연

한국정보통신기술협회

[blong116, mksung, hanjy]@tta.or.kr

A Study on the Bootstrap-Based Outlier Detection Method for the Safety of Combined Information

Lee Kangwon, Sung Min Kyoung, Han Ju Yeun

Telecommunications Technology Association

요 약

데이터3법이 개정되고 다양한 이종 분야 간 데이터 결합 및 안전한 활용을 위해 관련 부처에서 가명정보의 결합 및 반출 절차를 구성하고 이를 지원하고 있다. 그러나 전문가집단을 통한 적정성 검토로 인해 전문가마다 서로다른 기준을 적용하거나 일부 정보에 대한 검토가 누락되는 경우가 발생할 수 있어 이를 정량적으로 처리하는 것이 필요하다. 본 논문에서는 칼럼 조합에 대한 교차빈도를 계산하고 이에 대한 안전성 확보를 위해 부트스트랩 기반 특이정보 판단 방법론을 제안한다.

Ⅰ. 서 론

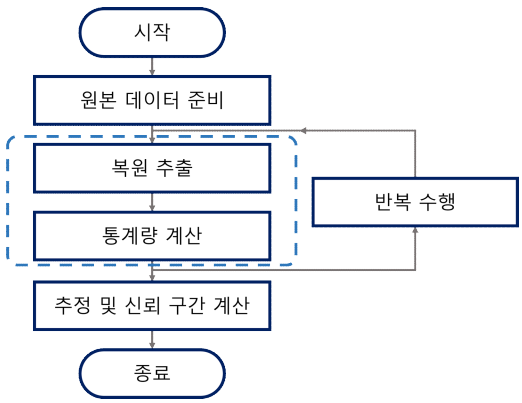
2020년 8월 데이터 3법이 개정되고 가명정보의 안전한 제공 및 활용을 위해 관련 부처에서는 지속적으로 가명정보 처리 관련 가이드라인[1-5]을 개선하고 있다. 또한, 개인정보보호위원회와 한국인터넷진흥원에서는 2021년부터 매년 「가명정보 활용 우수사례·아이디어 경진대회」 개최를 통해 다양한 이종 분야 간 데이터 결합 및 활용에 대한 관심을 증대시키고 있다.

가명정보의 안전한 결합 및 활용을 위해 「가명정보 처리 가이드라인」에서 가명정보의 결합 및 반출 절차를 마련하여 지원하고 있다. 결합된 가명정보 즉, 결합정보가 안전하게 활용될 수 있도록 조치되었는지 적정성을 검토하기 위해 결합전문기관에서 반출심사위원회를 구성하여 반출가능여부를 판단하고 있다. 그러나 전문가집단의 경험과 지식을 기반으로 정성적으로 판단하므로 전문가집단 구성원마다 서로 다른 기준이 적용되거나 일부 정보에 대한 적정성 검토가 누락되는 문제가 발생할 가능성이 존재한다. 이를 정량적으로 해소하기 위해 데이터의 특수성이나 통계적 특성을 분석하고 활용할 수 있으나, 원본 데이터의 수집 목적과 시기 등 여러 요인으로 인해 데이터의 특성이 달라질 수 있어 일반적인 상식을 적용한 판단 기준을 그대로 적용하기에는 적합하지 않다. 데이터의 특수성 및 통계적 특성으로 확인되는 특정 데이터 조합은 다른 데이터 조합 대비 상대적으로 분포 즉, 빈도가 적어 특정 개인의 식별 가능성이 높아지는데, 이러한 특성에 따라 데이터 조합에 대한 빈도를 확인하고 이를 기반으로 빈도가 현저히 적은 특이정보 판단을 통한 결합정보의 안전성 확보가 필요하다.

본 논문에서는 결합정보의 안전성을 확보하기 위해 데이터 조합의 교차빈도와 통계학에서 사용되는 부트스트랩(Bootstrap)[6]을 활용하여 빈도수에 대한 이상치 즉, 특이정보를 판단하는 방법론을 제안한다.

Ⅱ. 부트스트랩

부트스트랩은 통계학에서 사용되는 재표본화(Resampling) 기법으로, 주어진 데이터에서 여러 번의 복원 추출을 통해 새로운 표본을 생성하여 통계적 추정을 수행하는 방법이다. 부트스트랩은 그림 1과 같은 절차로 수행된다.



<그림 1. 부트스트랩 수행 절차>

먼저 분석에 활용할 n개의 데이터로 구성된 원본 데이터셋을 준비하고, 준비된 데이터셋에서 중복을 허용하여 크기가 n인 새로운 표본을 무작위로 추출한다. 이 때, 각 표본은 원본 데이터셋과 같은 크기인 n개를 가지지만, 특정 데이터가 중복되어 포함될 수 있다. 각 표본에서 평균, 표준편차 등과 같은 통계량을 계산한다. 복원 추출 과정과 통계량 계산 과정에 대해 1,000번 이상 반복 수행을 하게 되며, 반복 과정에서 계산된 통계량을 이용해 전체 분포를 추정하고 이를 기반으로 신뢰 구간이나 표본의 불확실성을 추정한다.

부트스트랩 신뢰구간은 데이터를 반복적으로 샘플링하여 다양한 통계량 분포를 반영하기 때문에 데이터의 변동성과 불확실성 포착에 활용될 수 있고, 이를 통해 신뢰구간 밖에 위치한 데이터는 원본 데이터셋의 전반적인 경향에서 벗어난 것으로 간주될 수 있어 이상치 판단에 유효한 기준이 될 수 있으며, 특히 데이터가 비정상적으로 분포되거나 비대칭적인 경우에 유용하다.

Ⅲ. 본론

1. 제안 방법론

부트스트랩 기법은 신뢰구간을 활용하여 기존 이상치 판단 기법[7-11]과 유사하게 특정 범위를 벗어나는 정보들을 이상치로 판단할 수 있어 결합정보의 안전성을 확보할 수 있으나, 빈도가 현저히 적은 값을 가진 이상치 즉, 특이정보를 판단하기 위해 해당 기법을 그대로 적용하는 것은 적합하지 않다.

본 논문에서 제안하는 부트스트랩 기반 특이정보 판단 방법론은 대상 칼럼 조합의 교차빈도에 대한 중앙값(Median), 중앙 절대 편차값(Median Absolute Deviation)을 통해 샘플링 대상 빈도를 제한하여 부트스트랩 기법을 수행하고 이를 통해 교차빈도가 적은 칼럼 조합을 이상치 즉, 특이정보로 판단한다. 제안하는 부트스트랩 기반 특이정보 판단 방법론은 다음과 같은 절차로 구성된다.

① 대상 칼럼 선택 및 교차빈도 산출

② 부트스트랩 샘플링 대상을 제한하기 위한 기준값(Threshold) 설정

(1). 데이터 분포에 대한 영향을 최소화하기 위한 교차빈도 중복제거

(2). 중복 제거된 교차빈도를 통해 중앙값(Median)과 중앙 절대 편차값(MAD, Median Absolute Deviation) 계산

(3). 중앙값과 중앙 절대 편차값을 활용하여 기준값 산출

- 산식 : $Threshold = |Median - 1.5 \times MAD|$

③ 부트스트랩 절차 수행

(1). '②' 과정에서 산출된 기준값보다 작은 교차빈도를 대상으로 복원 추출을 수행하여 재표본(resample) 생성

(2). 생성된 재표본에 대한 통계량(평균) 계산

(3). (1)~(2) 과정을 1,000번 반복 수행 후 전체 통계량의 분포에서 특이정보 판단 기준점 설정

(하위 5%를 특이정보 판단 기준점으로 설정)

④ 특이정보 판단 기준점보다 작은 교차빈도를 이상치(특이정보)로 판단

2. 실험수행 및 결과

본 논문에서 제안하는 부트스트랩 기반 특이정보 판단 방법론에 대한 실험을 위해 32,561건의 행으로 구성된 UCI Machine Learning Repository의 'Adult' 데이터셋[12]을 활용하였다. 대상 데이터셋에서 특정 칼럼 조합에 대해 교차빈도를 구하고, 이를 기반으로 제안하는 부트스트랩 기반 특이정보 판단 방법론 실험 결과를 <표 1>에 나타내었다.

<표 1. 제안 방법론 실험결과>

칼럼 조합	샘플링 대상 제한 기준값 (Threshold)	특이정보 판단 기준점	특이정보 건수
age, marital_status	20.0 (Median : 121.0) (MAD : 94.0)	5.9899	215 (비율 : 0.660%)
hours_per_week, relationship	7.0 (Median : 65.0) (MAD : 48.0)	2.6122	147 (비율 : 0.451%)
education, occupation	12.5 (Median : 85.0) (MAD : 65.0)	4.3181	81 (비율 : 0.249%)
age, workclass	7.5 (Median : 66.0) (MAD : 49.0)	2.6870	75 (비율 : 0.230%)
race, native_country	8.5 (Median : 45.5) (MAD : 36.0)	2.1897	52 (비율 : 0.160%)

제안하는 부트스트랩 기반 특이정보 판단 방법론을 통해 칼럼 조합에 대한 95% 신뢰구간을 계산하고 이보다 작은 빈도를 가진 데이터 조합을 특이정보로 판단한 결과, {age, marital_status} 칼럼 조합에서 215건의 행이 특이정보로 판단되었으며, 95% 신뢰구간인 5.9899보다 작은 교차빈도는 <표 2>와 같이 분포되어 있음을 확인하였고 칼럼 조합에 따라 교차빈도가 현저히 작은 경우가 발생할 수 있음이 확인되었다.

<표 2. {age, marital_status} 칼럼 조합의 특이정보 빈도 분포>

교차빈도 (a)	데이터 조합 건수 (b)	특이정보 건수 ($a \times b$)
1	38	38
2	27	54
3	13	39
4	11	44
5	8	40
합계		215

Ⅳ. 결론

본 논문에서는 결합정보의 안정성 확보를 위해 적용 가능한 부트스트랩 기반 특이정보 판단 방법론을 제안한다. 칼럼 조합에 대한 교차빈도를 구하고 중앙값과 중앙 절대 편차값을 활용하여 기준값을 설정하고 부트스트랩 기법을 활용하여 특이정보의 판단을 최소화하였으며, 교차빈도가 1인 데이터 조합이 적지 않음을 알 수 있었다. 이를 통해 공개된 데이터셋이라 할지라도 공격자가 가지고 있는 정보에 따라 특정 개인의 식별 위험성이 존재함을 알 수 있었으며, 향후 연구를 통해 데이터의 통계적 특성을 보존할 수 있도록 판단된 특이정보를 처리하는 방법론에 대한 연구를 진행할 예정이다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00634 '대용량 정형 데이터 대상 개인정보 가명.익명처리 자동화 및 안정성 검증 기술개발')

참 고 문 헌

[1] 개인정보보호위원회, "가명정보 처리 가이드라인," 2024.

[2] 교육부, "교육분야 가명·익명정보 처리 가이드라인," 2024.

[3] 보건복지부, "보건의료데이터 활용 가이드라인," 2024.

[4] 금융감독원, "금융분야 가명·익명처리 안내서," 2022.

[5] 행정안전부, "공공분야 가명정보 제공 실무안내서," 2024.

[6] Efron, B. "Bootstrap methods: another look at the jackknife," Breakthoughts in statistics: Methodology and distribution. New York, NY: Springer New York, pp. 569-693, 1992.

[7] R. E. Shiffler, "Maximum Z Scores and outliers," The American Statistician, 42(1), pp. 79-80, 1988.

[8] B. Iglewicz, and D. Hoaglin, HOW TO DETECT AND HANDLE OUTLIERS, Vol. 16, Amerian Society for Quality Control: Statistics Division, 1993.

[9] Grubbs, F. E. "Sample criteria for testing outlying observations," The Annals of Mathematical Statistics, Vol. 21, No. 1, pp. 27-58, 1950.

[10] V. Chandola, A. Banerjee, and V. Kumar "Anomaly detection: A survey," ACM computing surveys, 41(3), pp. 1-58, 2009

[11] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure", Technometrics, Vol. 25, No. 2, pp. 165-172, May. 1983.

[12] UCI Machine Learning Repository,https://archive.ics.uci.edu/