데이터 재식별 가능성 감소를 위한 행 단위 방법론 연구

이강원, 성민경, 한주연
한국정보통신기술협회
[blong116, mksung, hanjy]@tta.or.kr

A study on the row-based method for reducing possibility of data re-identification

Lee Kangwon, Sung Min Kyoung, Han Ju Yeun
Telecommunications Technology Association

요 약

데이터3법이 개정되고 가명처리된 개인정보 활용에 대한 관심이 높아짐에 따라 가명정보의 재식별 가능성에 대한 문제가 대두되고 있다. 특히 특정 개인에 대한 식별 가능성이 매우 높은 특이정보에 대한 처리가 필요하지만 이를 처리하는 방법에 대한 연구는 많지 않으며, 통계학이나 인공지능 학습용에서 활용되는 이상치 탐지 기법 적용은 적합하지 않다. 본 논문에서는 빈도수를 기반으로 특이정보를 판단하고, 재식별 가능성을 감소시키는 방법론을 제안한다. 제안하는 방법론을 통해 전문가의 도움을 받을 수 없는 환경에서도 정량적인 지표로 특이정보를 판단하고 재식별 위험도를 낮출 수 있다.

Ⅰ. 서 론

2020년 8월 데이터3법이 개정됨에 따라 개인정보를 가명처리하여 활용할 수 있는 기반이 새롭게 마련되었으며, 각 관련 부처에서는 개인정보의 오·남용을 방지하고 안전한 가명정보 처리 및 활용을 위해 가명정보 처리 가이드라인[1-5]을 작성하여 제공하고 있다. 안전한 가명정보 처리를 위해 개인정보의 가명처리에 대한 단계별 절차를 제시하고 있으며, 위험성 검토, 안전한 관리 단계에서 특정 개인을 알아보거나 다른 정보와의 사용 및 결합을 통해 개인을 알아볼 수 있는 재식별 가능성에 대해 언급하고 있다. 가명정보에서 개인을 식별할 수 있는 식별 가능성이 높은 정보로 특이정보가 존재한다. 특이정보란 희귀 성씨, 직업 또는 특정 지역의 고액급여수급자, 고액체납금액 등 다른 정보와 확연히 구분되거나 비정상적으로 분포를 벗어난 값으로서, 특정 개인의 식별 가능성이 매우 높은 정보이다.

이러한 특이정보를 처리하는 것만으로도 가명정보에 대한 재식별 위험을 낮출 수 있지만, 특이정보의 판단은 전문가의 경험 또는 지식 등을 활용하여 주관적으로 탐지하고 처리되는 경우가 많다. 정성적으로 전문가의 판단으로 탐지되기 때문에 사람의 실수로 인해 특이정보를 일부 보지 못하는 휴먼에러 등과 같은 문제가 발생할 수 있다. 또한, 병명이나 의약품 코드 등 일반인이 알아보기 어려운 정보들이 많아 전문가의 도움이 필수적으로 요구되는 의료 분야와 같은 특정 분야에서는 전문가의 도움을 받을 수 없는 경우 데이터 재식별 위험에 그대로 노출될 가능성이 매우 높다. 가명정보 처리 가이드라인에서는 3시그마규칙[6]이나 도수분포표 등을 이용하여 특이정보를 검토할 수 있다고 언급하고 있으나, 구체적인 내용이나 절차에 대한 내용이 없다.

본 논문에서는 정형 데이터에서 칼럼별 빈도수를 기준으로 특이정보를 판단하고, 이를 기반으로 데이터 행 단위 재식별 위험도(특정가능성, 연결가능성, 추론가능성)를 낮추기 위한 처리 방법론을 제안한다.

Ⅱ. 본론

1. 특이정보 분석

일반적인 데이터 유형은 크게 범주형 데이터와 수치형 데이터로 구분된다. 범주형 데이터는 성별, 혈액형, 학력 등과 같이 카테고리로 분류되는 데이터로 주로 문자형태로 구성되며, 수치형 데이터는 나이, 몸무게, 급여 등 숫자 형태로 구성된 데이터를 의미한다.

범주형 데이터는 도수분포표, 히스토그램 등을 통해 빈도수를 분석하여 특이정보를 판단하고, 수치형 데이터는 통계 분야의 이상치 탐지 기법[7-10]을 활용하여 비정상적으로 분포를 벗어난 최소값 또는 최대값을 판단한다.

그러나 수치형 데이터의 이상치 탐지 방법은 일반적인 통계 또는 인공지능 학습용 데이터에 대한 이상치를 탐지하는데 적합하지만, 최소값 또는 최대값에 데이터가 밀집될 가능성이 있는 가명정보에서의 특이정보 판단에는 적합하지 않다. 또한, 2차원 테이블 형태로 표현되는 가명정보에서 위와 같은 이상치 탐지 기법을 그대로 적용하게 되면 행 삭제, 로컬 삭제, 값 대체 등 특이정보 처리 시 무분별한 데이터 가공으로 인해 데이터 분석에 어려움이 발생할 수 있다. 일반적인 인공지능 학습용 또는 통계용 자료와는 다르게 가명정보에서는 전체 데이터에 대해 특이정보 여부를 먼저 탐지하고 이를 행 단위로 판단하여 특이정보에 대해 유연하게 처리할 필요가 있다.

2. 제안 방법론

본 논문에서 제안하는 특이정보 판단 방법은 전체 데이터를 대상으로 각 칼럼에 대한 범주별 데이터 빈도수를 측정하고, 특이정보 임계값 X라는 변수와 비교하여 특이정보 여부를 판단한다. 판단한 특이정보를 2차원 테이블 형태로 구성하여 행 단위로 특이정보가 몇 개인지 확인하고 확인된 건수와 특이정보 처리 기준을 설정하기 위한 변수값 α, β, γ와 비교하여 값 대체, 로컬 삭제, 행(row) 삭제 처리 방법을 결정한다. 값 대체 방법은

데이터를 삭제하지 않고 다른 값으로 치환하는 방법을 말하며, 로컬 삭제는 한 칼럼의 정보를 빈 값(null)으로 삭제하고, 행 삭제는 행 전체를 삭제하는 방법을 말한다. 제안한 행 단위로 특이정보를 판단하는 방법은 다음과 같은 절차로 구성된다.

① 전체 데이터를 대상으로 각 칼럼에 대한 범주별 빈도수 집계
② 특이정보 임계값 X 설정
③ 범주별 빈도수와 특이정보 임계값 X 비교
- 범주별 빈도수 < 특이정보 임계값인 경우 특이정보로 판단
④ 행 단위로 특이정보로 판단된 건수 집계(예시: 하나의 행에 10개의 칼럼이 있고 그 중, 2개의 칼럼이 특이정보로 판단되었다면 특이정보 건수는 2개임)
⑤ 그림1과 같은 기준으로 특이정보 건수와 변수값 α, β, γ을 비교하여 특이정보 처리(α < β < γ , α, β, γ는 양의 정수)
- 특이정보 건수 < α : 별도의 처리를 수행하지 않음
- α ≤ 특이정보 건수 < β : 값 대체 처리방법 수행
- β ≤ 특이정보 건수 < γ : 로컬 삭제 처리방법 수행
- γ ≤ 특이정보 건수 : 행 삭제 처리방법 수행

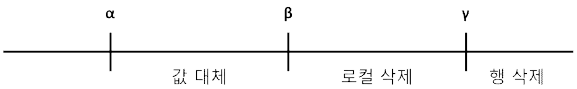

그림 1. 특이정보 처리를 위한 변수 기준

3. 실험수행

본 논문에서 제안한 특이정보 처리 방법론에 대한 실험을 위한 데이터로 15개 칼럼과 32,561건의 행으로 구성된 UCI Machine Learning Repository의 Adult 데이터[11]를 활용하였다. 범주 종류가 적거나 개인별로 부여된 값이 있는 칼럼 등을 제외한 9개 칼럼(age, workclass, education, marital_status, occupation, relationship, race, hours_per_week, native_country)을 대상으로 실험을 수행하였다. 특이정보 임계값 X는 1에서 50의 정수값으로 설정하였으며 특이정보 처리 기준에 대한 변수값 α, β, γ는 각각 1, 2, 3으로 설정하였다. 범주별 빈도수와 특이정보 임계값 X를 비교하여 특이정보가 아닌 경우와 특이정보인 경우로 구분하여 테이블 형태로 구성하고 행 단위로 특이정보의 수를 집계하였다. 집계된 건수 n과 변수값 α, β, γ와의 비교를 통해 행 단위 특이정보 건수가 α ≤ n < β 면 값 대체 처리방법을, 건수가 β ≤ n < γ 면 로컬 삭제 방법을 제시하고, γ ≤ n 이면 행 삭제 처리방법을 적용하도록 구성하였다. 실험 결과에 대한 특이정보 임계값 X별 값 대체, 로컬 삭제, 행 삭제 처리 방법이 필요한 행 수를 그림2를 통해 그래프에 표현하였다.

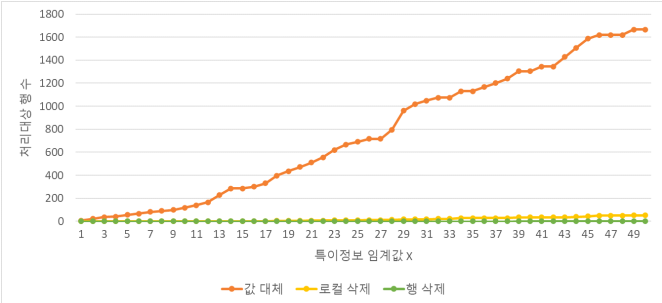

그림 2. 특이정보 임계값 X 변화에 따른 '값 대체', '로컬 삭제', '행 삭제' 변화 실험 결과

특이정보 임계값이 증가함에 따라 값 대체, 로컬 삭제, 행 삭제로 처리가 필요한 행 수가 증가하는 것을 확인하였으며, 실험에 사용된 Adult 데이터의 경우 특이정보 임계값 X가 39 이상인 경우 행 삭제 처리가 필요한 행이 확인되었다.

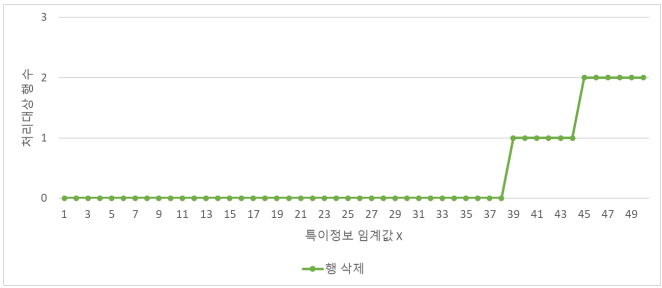

그림 3. 특이정보 임계값 X 변화에 따른 '행 삭제' 변화 실험 결과

Ⅲ. 결론

본 논문에서는 정형 데이터에서 칼럼별 빈도수를 통해 판단한 특이정보를 행 단위로 판단하고 처리하는 방법론을 제안하였다. 제안한 방법론을 통해 전문가의 도움을 받을 수 없는 경우에도 정량적인 지표를 활용하여 행 단위 재식별 위험도를 낮출 수 있다. 향후 가명정보가 증가되고 이를 활용하는 기관 및 분야가 많아짐에 따라 가명정보에 대한 재식별 위험도에 관한 연구가 활발히 이루어질 것이며, 데이터 손실을 최소화한 특이정보 처리 방법에 관한 연구가 필요할 것으로 보인다.

ACKNOWLEDGMENT

이 논문은 2021년 과학기술정보통신부의 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," 2022.
[2] 교육부, "교육분야 가명·익명정보 처리 가이드라인," 2022.
[3] 보건복지부, "보건의료데이터활용가이드라인개정(안)," 2022.
[4] 금융감독원, "금융분야 가명·익명처리 안내서," 2021.
[5] 행정안전부, "공공분야 가명정보 제공 실무안내서," 2021.
[6] Pukelsheim, F. "The three sigma rule," The American Statistician, Vol. 48, No. 2, pp 88-91, 1994.
[7] Grubbs, F. E. "Sample criteria for testing outlying observations," The Annals of Mathematical Statistics, Vol. 21, No. 1, pp. 27-58, 1950.
[8] Yang, J., Rahardja, S., & Fränti, P. "Outlier detection: how to threshold outlier scores?," Proceedings of the international conference on artificial intelligence, information processing and cloud computing, No. 37, pp 1-6, Dec. 2019.
[9] S. Walfish, "A review of statistical outlier methods", Pharmaceutical Technol., Vol. 30, No. 11, pp. 1-5, 2006.
[10] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure", Technometrics, Vol. 25, No. 2, pp. 165-172, May. 1983.
[11] UCI Machine Learning Repository,https://archive.ics.uci.edu/