쿼드트리와 균등 샘플링를 이용한 효과적 차분 프라이버시 K-평균 클러스터링 알고리즘

A Differentially Private K-Means Clustering using Quadtree and Uniform Sampling

홍 대 영, 구 한 준, 심 규 석
서울대학교

Daeyoung Hong, Hanjun Goo, Kyuseok Shim
Seoul National University

요약

최근 데이터를 공개할 때 프라이버시를 보호하기 위한 방법들이 연구되고 있다. 그 중 차분 프라이버시(differential privacy)는 최소성 공격 등에 대해서도 안전함이 증명된 익명화 기법이다. 본 논문에서는 기존 차분 프라이버시 -평균 클러스터링 알고리즘의 성능을 개선하고 실생활 데이터를 이용한 실험을 통해 이를 검증한다.

Ⅰ. 서론

최근 소셜 미디어, 모바일 기기 등으로 방대한 데이터가 축적되고 있으나, 그대로 공개할 경우 개인 정보가 침해될 수 있으므로 다양한 익명화 기법들이 연구되었다. 그 중 차분 프라이버시는 최소성 공격 등에 대해서도 안전함이 증명되어[1] 요약된 데이터에 대한 정보 보호의 강력한 기준이 되고 있다.

차분 프라이버시 기법을 적용한 다양한 공개 기법들이 연구되었는데, k-평균 클러스터링의 결과를 공개하는 기존 알고리즘[3]은 일반적인 데이터 특성과 상이한 가정으로 인해 정확도가 저해되는 단점이 있다. 이에 대해 본 논문은 균등 샘플링을 이용하여 성능을 향상시키고 실생활 데이터를 이용한 실험을 통해 검증한다.

Ⅱ. 관련 연구

프라이버시를 보호하면서 k-평균 클러스터링 알고리즘의 결과를 공개하기 위한 다양한 연구가 진행되었는데, [2]에서는 데이터를 등간격 히스토그램으로 나타내고 각 버킷에 노이즈를 삽입한 뒤, k-평균 클러스터링을 수행하는 알고리즘이 제안되었고 우수한 성능을 보였다. 하지만 데이터 분포와 무관하게 등간격 히스토그램을 만들기 때문에 데이터가 분포하지 않는 영역에도 불필요한 노이즈를 삽입하게 되는 단점이 있다. 이를 보완하기 위해 [3]에서 데이터의 분포를 더 적은 버킷으로 나타낼 수 있는 쿼드 트리를 이용하여 히스토그램을 만든 뒤 k-평균 클러스터링을 수행하는 알고리즘을 제안하였다. 그러나 k-평균 클러스터링을 수행할 때 데이터가 히스토그램의 각 버킷 영역의 중앙에만 위치한다고 가정하여 대부분의 데이터 특성과 상이하다는 단점이 있다. 따라서 본 논문은 이를 보완하기 위한 균등 샘플링 기반의 알고리즘을 제안한다.

[4]에서는 차분 프라이버시를 만족시키도록 쿼드트리로 데이터를 나눈 후, 밀도 기반 클러스터링을 수행하여 그 결과를 공개하는 알고리즘을 제안하였는데, [3]에서 쿼드트리를 생성하는 방법은 [4]를 참고하였다.

Ⅲ. 배경 이론

1. 차분 프라이버시 (Differential Privacy)

어떤 알고리즘 $A : D \rightarrow R$이 임의의 두 인접한 데이터 $d, d' \in D$와 $S \subseteq R$에 대해 다음의 조건을 만족하면 알고리즘 A는 ϵ-차분 프라이버시를 만족한다.

$$\Pr[A(d_1) \in S] \leq \Pr[A(d_2) \in S] \quad (1)$$

2. 순차 구성 (Sequential Composition)

알고리즘 A_1, A_2가 각각 ϵ_1, ϵ_2-차분 프라이버시를 만족하면, A_1, A_2를 순차적으로 적용하는 알고리즘 A는 $(\epsilon_1 + \epsilon_2)$-차분 프라이버시를 만족한다.

3. 쿼드트리 (Quadtree)

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2018-2013-0-00881). 또한 이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A1A02937186). 또한 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보·컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No. NRF-2017M3C4A7063570).

d-차원 데이터 $D=\{x_1,x_2,...,x_{|D|}\}$에 대하여 각 리프 노드에 속하는 점의 개수가 최대 점 개수 σ보다 작거나 같아질 때까지 혹은 최대 깊이 δ에 도달할 때까지 해당 리프 노드의 공간을 2^d개의 공간으로 분할하는 과정을 재귀적으로 수행하여 얻어지는 구조이다.

알고리즘 1. UniQuadkDP

Input: 데이터 $D=\{x_1,x_2,...,x_n\}$, $x_i \in \mathbb{R}d$, 정보 보호 수준 ϵ, 정보 보호 비율 γ, 클러스터 개수 k, 샘플링 회수 m

1. $\epsilon_1 \leftarrow \gamma\epsilon$, $\epsilon_2 \leftarrow (1-\gamma)\epsilon$
2. $T \leftarrow$ buildQuadTreeHistogram$(D, \epsilon_1, \epsilon_2)$
3. $S \leftarrow \varnothing$ // the set of sampled points
4. **for** each leaf node t in T
5. **while** iterate until m times
6. $R_t \leftarrow$ the range of t
7. sample $s.x$ from uniform distribution $U(R_t)$
8. $s.w \leftarrow t.w / m$
9. $S \leftarrow S \cup s$
10. $\{c_1,c_2,...,c_k\} \leftarrow \mathrm{Kmeans}(S,k)$
11. return Cluster centroids $\{c_1,c_2,...,c_k\}$

Ⅳ. 제안하는 알고리즘 (UniQuadkDP)

[3]에서 제안한 알고리즘 QuadkDP를 개선시킨 알고리즘으로 의사코드는 알고리즘 1과 같다. buildQuadTreeHistogram$(D, \epsilon_1, \epsilon_2)$으로 쿼드트리 히스토그램을 생성시키는 과정(line 1 ~ 2)까지는 QuadkDP[3]와 동일하게 진행된다. 제안하는 알고리즘에서는 각 리프 노드 t의 영역 R_t에 대해 샘플링할 점 s 점의 좌표 $s.x$를 균등 샘플링으로 추출하고(line 7), 가중치 $s.w$를 리프 노드 t의 도수 $t.w$를 샘플링 회수 m으로 나누어서 할당한다(line 8). 이 과정을 각 리프 노드에 대해 m번 반복하여 샘플링된 점들의 집합 S를 만들고, 이에 대해 k-평균 클러스터링 알고리즘을 수행하여 중심점들을 출력한다.

Ⅴ. 실험

1. 데이터

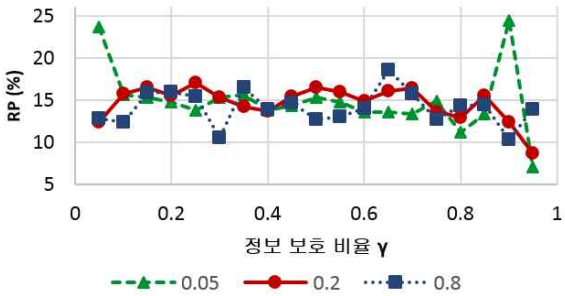
소셜 네트워크 서비스 Gowalla에서 2009 ~ 2010년도에 수집한 체크인 위치 정보 데이터[5] 중 64,428개의 좌표를 랜덤하게 추출하였다.

2. 실험 방법

평가 척도로는 다음과 같은 NICV (Normalized Intracluster Variance)로 성능을 평가하였다. 값이 낮을수록 좋은 클러스터의 중심점을 찾은 것이다.

$$\mathrm{NICV}=\frac{1}{|D|}\sum_{j=1}^{k}\sum_{x_i \in D}\| x_i - c_j \|^2 \qquad (2)$$

성능 비교를 위해 같은 클러스터 중심점들로 초기화시킨 후, 다음과 같이 상대적 성능 RP (Relative Performance)을 계산하였다.



▶▶ 그림 1. 다양한 ϵ에 따른 상대적 성능 RP

$$\mathrm{RP}=\frac{\mathrm{NICV}_{\mathrm{QuadkDP}}-\mathrm{NICV}_{\mathrm{UniQuadkDP}}}{\mathrm{NICV}_{\mathrm{noPrivacy}}} \qquad (3)$$

여기서 $\mathrm{NICV}_{\mathrm{noPrivacy}}$는 프라이버시 보호를 적용하지 않은 원래 데이터로 수행한 NICV이다. RP가 높을수록 제안하는 알고리즘의 성능이 좋음을 의미한다.

3. 실험 결과

그림 1은 $m=30$에서 다양한 정보 보호 수준 ϵ에 대해 γ를 변화시켜가며 상대적 성능을 100번 반복 측정한 것의 평균값을 나타낸 것이다. 모든 상황에서 성능이 향상됨을 확인할 수 있다.

Ⅵ. 결론

본 논문에서는 쿼드 트리와 균등 샘플링를 이용하여 효과적으로 차분 프라이버시를 지키면서 k-평균 클러스터링을 수행하는 알고리즘을 제안하였다. 실생활 데이터를 이용한 실험을 통해 기존의 중심점 기반 k-평균 클러스터링 알고리즘 QuadkDP보다 제안하는 UniQuadkDP 알고리즘이 더 좋은 성능을 보임을 확인하였다.

■ 참 고 문 헌 ■

[1] R. Wong, A. Fu, K. Wang, J. Pei, "Minimality attack in privacy preserving data publishing," Proc. of the 33rd VLDB, pp. 543-554, 2007.

[2] Su, D., Cao, J., Li, N., Bertino, E., & Jin, H., "Differentially private k-means clustering," Proc. of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 26-37, 2016.

[3] 구한준, 정우환, 오성웅, 권수용, 심규석 (2018). "쿼드트리를 이용한 동적 공간 분할 기반 차분 프라이버시 k-평균 클러스터링 알고리즘", 정보과학회논문지, 제45권, 제3호, pp. 288-293, 2018.

[4] Ho, S. S. and S. Ruan, "Differential privacy for location pattern mining," Proc. of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, pp. 17-24, 2011.

[5] https://snap.stanford.edu/data/loc-gowalla.html