(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2019-0128963
(43) 공개일자 2019년11월19일

(51) 국제특허분류(Int. Cl.)
G06F 21/62 (2013.01) *G06F 16/00* (2019.01)
G06F 21/71 (2013.01)
(52) CPC특허분류
G06F 21/6245 (2013.01)
G06F 16/26 (2019.01)
(21) 출원번호 10-2018-0053366
(22) 출원일자 2018년05월09일
심사청구일자 2018년05월09일

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
(72) 발명자
박석
서울특별시 종로구 평창12길 16-6, 2동 202호 (평창동, 평창삼호빌라)
심규석
경기도 하남시 위례중앙로 215, 6406동 804호 (학암동, 롯데캐슬아파트)
(뒷면에 계속)
(74) 대리인
유미특허법인

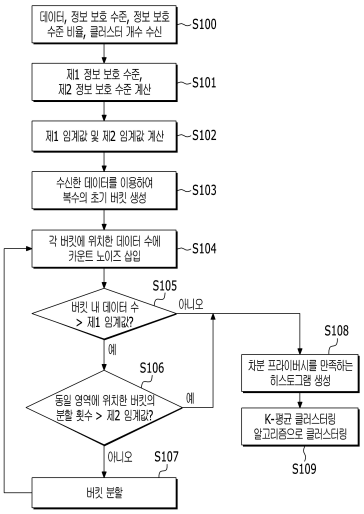
전체 청구항 수 : 총 14 항

(54) 발명의 명칭 **K-평균 클러스터링 기반의 데이터 마이닝 시스템 및 이를 이용한 K-평균 클러스터링 방법**

(57) 요 약

데이터 마이닝 시스템이 데이터를 K-평균 클러스터링하는 방법으로서, 좌표 정보로 이루어진 복수의 데이터의 분포를 미리 설정된 수만큼 분할하여 복수의 초기 버킷을 생성하면, 생성한 초기 버킷에 각각 포함된 데이터 수에 카운트 노이즈를 반영하여, 미리 계산된 제1 임계값과 제2 임계값을 토대로 복수의 초기 버킷 중 적어도 하나의 초기 버킷을 분할하여 복수의 버킷으로 생성한다. 생성된 복수의 버킷 각각에 대한 구간 정보와, 카운트 노이즈가 반영된 버킷 각각에 포함된 데이터 수를 포함하여, 차분 프라이버시를 만족하는 히스토그램을 생성하고, 차분 프라이버시를 만족하는 히스토그램을 외부로부터 수신한 클러스터 개수를 토대로 K-평균 클러스터링을 실행한다.

대 표 도 - 도2



(52) CPC특허분류
G06F 21/71 (2013.01)
G06F 2216/03 (2013.01)
(72) 발명자
구한준
서울특별시 송파구 백제고분로19길 12-14, 101호 (잠실동, 남강로얄맨션)
정우환
서울특별시 관악구 솔밭로5길 6, 204호 (봉천동)
오성웅
서울특별시 관악구 관악로25길 17, 705호 (봉천동)
권수용
서울특별시 관악구 관악로 1, 903동 614호 (신림동, 관악사)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711043688
부처명 미래창조과학부
연구관리전문기관 한국연구재단
연구사업명 차세대정보컴퓨팅기술개발사업
연구과제명 소셜 및 정보 네트워크 빅데이터 마이닝 소프트웨어 원천 기술 개발
기 여 율 10/100
주관기관 서울대학교 산학협력단
연구기간 2016.07.01 ~ 2017.06.30
이 발명을 지원한 국가연구개발사업
과제고유번호 1711074143
부처명 과학기술정보통신부
연구관리전문기관 한국연구재단
연구사업명 차세대정보컴퓨팅기술개발사업
연구과제명 웹스케일 거대 복잡형 데이터 통합 및 추론 원천기술 개발
기 여 율 35/100
주관기관 서울대학교 산학협력단
연구기간 2017.09.01 ~ 2018.05.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711070411
부처명 미래창조과학부
연구관리전문기관 정보통신기술진흥센터
연구사업명 정보통신기술인력양성(R&D
연구과제명 스마트 공간을 위한 빅데이터 기반의 스마트 라이프 케어 서비스
기 여 율 20/100
주관기관 충북대학교 산학협력단
연구기간 2018.01.01 ~ 2018.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 201839019.01
부처명 미래창조과학부
연구관리전문기관 정보통신기술진흥센터
연구사업명 정보보호핵심원천기술개발(R&D)
연구과제명 차분 프라이버시 기반 비식별화 기술 개발
기 여 율 35/100
주관기관 서강대학교 산학협력단
연구기간 2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

데이터 마이닝 시스템이 데이터를 K-평균 클러스터링하는 방법으로서,

좌표 정보로 이루어진 복수의 데이터의 분포를 미리 설정된 수만큼 분할하여 복수의 초기 버킷을 생성하는 단계,

상기 생성한 초기 버킷에 각각 포함된 데이터 수에 카운트 노이즈를 반영하고, 미리 계산된 제1 임계값과 제2 임계값을 토대로 상기 복수의 초기 버킷 중 적어도 하나의 초기 버킷을 분할하여 복수의 버킷으로 생성하는 단계,

상기 생성된 복수의 버킷 각각에 대한 구간 정보와, 상기 카운트 노이즈가 반영된 버킷 각각에 포함된 데이터 수를 포함하여, 차분 프라이버시를 만족하는 히스토그램을 생성하는 단계, 그리고

상기 차분 프라이버시를 만족하는 히스토그램을, 외부로부터 수신한 클러스터 개수를 토대로 K-평균 클러스터링을 실행하는 단계

를 포함하는 클러스터링 방법.

청구항 2

제1항에 있어서,

상기 초기 버킷을 생성하는 단계는,

상기 좌표 정보로 이루어진 복수의 데이터, 상기 데이터를 토대로 생성되는 복수의 버킷에 대한 제1 정보와 제2 정보를 보호하기 위한 정보 보호 수준, 상기 제1 정보와 제2 정보를 어느 정도의 비율로 보호할지 나타내는 정보 보호 수준 비율 그리고 상기 K-평균 클러스터링을 실행하기 위한 클러스터 개수를 포함하는 입력 정보를 수신하는 단계

를 포함하는 클러스터링 방법.

청구항 3

제2항에 있어서,

상기 초기 버킷을 생성하는 단계는,

상기 데이터를 통해 생성되는 복수의 버킷에 포함될 데이터 수를 제한하기 위한 제1 임계값과, 동일한 버킷이 분할되는 분할 횟수를 제한하기 위한 제2 임계값을 계산하는 단계

를 포함하는 클러스터링 방법.

청구항 4

제3항에 있어서,

상기 제1 임계값은 상기 데이터의 수를 토대로 계산되고, 상기 제2 임계값은 상기 데이터의 수와 상기 복수의 데이터를 이루를 좌표의 차원(dimension)을 토대로 계산되는 클러스터링 방법.

청구항 5

제4항에 있어서,

상기 복수의 버킷으로 생성하는 단계는,

상기 생성한 복수의 초기 버킷 각각에 상기 카운트 노이즈를 반영하고, 카운트 노이즈가 반영된 초기 버킷 각각

포함된 데이터 수가 상기 제1 임계값보다 많은지 확인하는 단계,

임의의 초기 버킷에 포함된 데이터 수가 상기 제1 임계값보다 많으면, 상기 임의의 초기 버킷이 분할된 횟수가 상기 제2 임계값보다 많은지 확인하는 단계, 그리고

상기 임의의 초기 버킷이 분할된 횟수가 상기 제2 임계값 보다 적다면, 상기 임의의 초기 버킷을 분할하여 복수의 새로운 버킷으로 생성하는 단계

를 포함하는 클러스터링 방법.

청구항 6

제5항에 있어서,

상기 새로운 버킷으로 생성하는 단계는 각각의 버킷 내 데이터 수가 상기 제1 임계값보다 적고, 각각의 버킷이 분할되는 경우 분할된 횟수가 상기 제2 임계값보다 적을 경우 반복 실행하여 새로운 버킷을 생성하고,

상기 생성된 새로운 버킷 각각은 복수의 버킷의 구간과 상기 복수의 버킷 각각에 포함된 데이터 수가 상이한 클러스터링 방법.

청구항 7

제6항에 있어서,

상기 입력 정보를 수신하는 단계는,

상기 정보 보호 수준과 상기 정보 보호 수준 비율을 이용하여, 상기 제1 정보를 보호하기 위한 제1 정보 보호 수준을 계산하고, 상기 제2 정보를 보호하기 위한 제2 정보 보호 수준을 계산하는 단계

를 포함하고,

상기 제1 정보는 상기 복수의 버킷 각각에 대한 구간 정보를 보호하기 위하여 버킷을 나누는 기준 정보이고, 상기 제2 정보는 상기 복수의 버킷 각각에 포함되는 데이터 수를 보호하기 위하여 상기 카운트 노이즈를 생성하는 기준 정보인 클러스터링 방법.

청구항 8

제7항에 있어서,

상기 복수의 버킷 각각에 포함되어 있는 복수의 데이터는 쿼드 트리 방식으로 분포되는 클러스터링 방법.

청구항 9

데이터 마이닝 시스템으로서,

좌표 정보로 이루어진 복수의 데이터, 상기 데이터를 토대로 생성되는 복수의 버킷에 대한 제1 정보와 제2 정보를 보호하기 위한 정보 보호 수준, 상기 제1 정보와 제2 정보를 어느 정도의 비율로 보호할지 나타내는 정보 보호 수준 비율 그리고 상기 K-평균 클러스터링을 실행하기 위한 클러스터 개수를 포함하는 입력 정보를 수신하고, 상기 수신한 정보들을 이용하여 상기 복수의 데이터를 쿼드 트리를 이용하여 복수의 버킷을 생성하며, 생성한 복수의 버킷에 대한 정보를 보호하여 차분 프라이버시를 만족하는 히스토그램을 생성하는 히스토그램 생성 모듈, 그리고

상기 차분 프라이버시를 만족하는 히스토그램과 상기 복수의 데이터를 상기 클러스터 개수만큼의 클러스터에 분류하는 K-평균 클러스터링 모듈

을 포함하는 데이터 마이닝 시스템.

청구항 10

제9항에 있어서,

상기 히스토그램 생성 모듈은,

상기 데이터를 통해 생성되는 복수의 버킷에 포함될 데이터 수를 제한하기 위한 제1 임계값과, 동일한 버킷이

분할되는 분할 횟수를 제한하기 위한 제2 임계값을 계산하며,

상기 제1 임계값은 상기 데이터의 수를 토대로 계산하고, 상기 제2 임계값은 상기 데이터의 수와 상기 복수의 데이터를 이루를 좌표의 차원(dimension)을 토대로 계산하는 데이터 마이닝 시스템.

청구항 11

제10항에 있어서,

상기 히스토그램 생성 모듈은,

상기 정보 보호 수준과 상기 정보 보호 수준 비율을 이용하여, 상기 제1 정보를 보호하기 위한 제1 정보 보호 수준을 계산하고, 상기 제2 정보를 보호하기 위한 제2 정보 보호 수준을 계산하며,

상기 제1 정보는 상기 복수의 버킷 각각에 대한 구간 정보를 보호하기 위하여 분할 대상인 버킷의 구간에 대한 기준 정보이고, 상기 제2 정보는 상기 복수의 버킷 각각에 포함되는 데이터 수를 보호하기 위하여 상기 카운트 노이즈를 생성하는 기준 정보인 데이터 마이닝 시스템.

청구항 12

제11항에 있어서,

상기 히스토그램 생성 모듈은,

상기 수신한 데이터의 분포를 확인하고 확인한 데이터 분포를 미리 설정된 수로 분할하여 복수의 초기 버킷을 생성하며,

상기 복수의 초기 버킷 각각에 포함된 데이터 수에 상기 카운트 노이즈가 반영된 데이터 수가 상기 제1 임계값 보다 많은지, 상기 복수의 초기 버킷이 분할된 횟수가 상기 제2 임계값 보다 적은지 확인하여 복수의 새로운 버킷으로 생성하는 데이터 마이닝 시스템.

청구항 13

제12항에 있어서,

상기 생성한 복수의 새로운 버킷은 복수의 버킷의 구간과 상기 복수의 버킷 각각에 포함된 데이터 수가 상이한 데이터 마이닝 시스템.

청구항 14

제12항에 있어서,

상기 히스토그램 생성 모듈은,

분할되어 생성된 각각의 버킷에 포함되어 있는 카운트 노이즈가 반영된 데이터 수와 상기 분할되어 생성된 각각의 버킷에 대한 구간 정보를 포함하여, 차분 프라이버시를 만족하는 히스토그램을 생성하는 데이터 마이닝 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 K-평균 클러스터링 기반의 데이터 마이닝 시스템 및 이를 이용한 K-평균 클러스터링 방법에 관한 것이다.

배 경 기 술

[0002] 데이터를 공개하는 경우 그리고 데이터를 활용한 알고리즘의 결과를 공개할 때에도, 개인 정보 노출 방지는 필수적이다. 공개되는 데이터에서 개인 정보 노출을 방지하기 위하여 k-익명화와 l-다양성과 같은 다양한 기법들이 연구되었지만, 데이터에 대한 공격자의 사전 지식이 있을 때 최소성 공격(minimality attack) 등을 통해 개인 정보가 노출 될 수도 있다.

[0003] 반면, 차분 프라이버시를 활용한 데이터 보호 기법은 공격자의 사전 지식과 상관 없이 특정 사람의 데이터의 존재 유무를 확률적으로 숨겨 주기 때문에 기존의 방법들에 비하여 안전하다. 이러한 차분 프라이버시 기법을 적용한 데이터 공개 기법들이 다양하게 연구되었으며, 데이터 공개뿐 아니라 데이터 마이닝의 결과를 공개하는 기법들도 연구가 되었다.

[0004] 차분 프라이버시를 적용하는 방법은 일반적으로 라플라스 분포를 따르는 노이즈를 삽입하여 만족시키는데, 히스토그램의 경우 각 버킷의 도수에 Lap(1/ε)의 노이즈를 삽입하면 차분 프라이버시가 성립된다. 버킷의 숫자가 많아지면 데이터를 더 세밀하게 표현할 수 있지만, 노이즈의 삽입이 많아지게 되기 때문에 데이터의 활용도가 떨어진다.

발명의 내용

해결하려는 과제

[0005] 따라서, 본 발명은 쿼드 트리를 이용한 차분 프라이버시 K-평균 클러스터링 알고리즘을 이용하여 데이터 마이닝하는 시스템 및 이를 이용한 K-평균 클러스터링 방법을 제공한다.

과제의 해결 수단

[0006] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 하나의 특징인 데이터 마이닝 시스템이 데이터를 K-평균 클러스터링하는 방법으로서,

[0007] 좌표 정보로 이루어진 복수의 데이터의 분포를 미리 설정된 수만큼 분할하여 복수의 초기 버킷을 생성하는 단계, 상기 생성한 초기 버킷에 각각 포함된 데이터 수에 카운트 노이즈를 반영하고, 미리 계산된 제1 임계값과 제2 임계값을 토대로 상기 복수의 초기 버킷 중 적어도 하나의 초기 버킷을 분할하여 복수의 버킷으로 생성하는 단계, 상기 생성된 복수의 버킷 각각에 대한 구간 정보와, 상기 카운트 노이즈가 반영된 버킷 각각에 포함된 데이터 수를 포함하여, 차분 프라이버시를 만족하는 히스토그램을 생성하는 단계, 그리고 상기 차분 프라이버시를 만족하는 히스토그램을, 외부로부터 수신한 클러스터 개수를 토대로 K-평균 클러스터링을 실행하는 단계를 포함한다.

[0008] 상기 초기 버킷을 생성하는 단계는, 상기 좌표 정보로 이루어진 복수의 데이터, 상기 데이터를 토대로 생성되는 복수의 버킷에 대한 제1 정보와 제2 정보를 보호하기 위한 정보 보호 수준, 상기 제1 정보와 제2 정보를 어느 정도의 비율로 보호할지 나타내는 정보 보호 수준 비율 그리고 상기 K-평균 클러스터링을 실행하기 위한 클러스터 개수를 포함하는 입력 정보를 수신하는 단계를 포함할 수 있다.

[0009] 상기 초기 버킷을 생성하는 단계는, 상기 데이터를 통해 생성되는 복수의 버킷에 포함될 데이터 수를 제한하기 위한 제1 임계값과, 동일한 버킷이 분할되는 분할 횟수를 제한하기 위한 제2 임계값을 계산하는 단계를 포함할 수 있다.

[0010] 상기 제1 임계값은 상기 데이터의 수를 토대로 계산되고, 상기 제2 임계값은 상기 데이터의 수와 상기 복수의 데이터를 이루를 좌표의 차원(dimension)을 토대로 계산할 수 있다.

[0011] 상기 복수의 버킷으로 생성하는 단계는, 상기 생성한 복수의 초기 버킷 각각에 상기 카운트 노이즈를 반영하고, 카운트 노이즈가 반영된 초기 버킷 각각 포함된 데이터 수가 상기 제1 임계값보다 많은지 확인하는 단계, 임의의 초기 버킷에 포함된 데이터 수가 상기 제1 임계값보다 많으면, 상기 임의의 초기 버킷이 분할된 횟수가 상기 제2 임계값보다 많은지 확인하는 단계, 그리고 상기 임의의 초기 버킷이 분할된 횟수가 상기 제2 임계값 보다 적다면, 상기 임의의 초기 버킷을 분할하여 복수의 새로운 버킷으로 생성할 수 있다.

[0012] 상기 입력 정보를 수신하는 단계는, 상기 정보 보호 수준과 상기 정보 보호 수준 비율을 이용하여, 상기 제1 정보를 보호하기 위한 제1 정보 보호 수준을 계산하고, 상기 제2 정보를 보호하기 위한 제2 정보 보호 수준을 계산하는 단계를 포함하고, 상기 제1 정보는 상기 복수의 버킷 각각에 대한 구간 정보를 보호하기 위하여 버킷을 나누는 기준 정보이고, 상기 제2 정보는 상기 복수의 버킷 각각에 포함되는 데이터 수를 보호하기 위하여 상기 카운트 노이즈를 생성하는 기준 정보이다.

[0013] 상기 본 발명의 기술적 과제를 달성하기 위한 본 발명의 또 다른 특징인 데이터 마이닝 시스템으로서,

[0014] 좌표 정보로 이루어진 복수의 데이터, 상기 데이터를 토대로 생성되는 복수의 버킷에 대한 제1 정보와 제2 정보를 보호하기 위한 정보 보호 수준, 상기 제1 정보와 제2 정보를 어느 정도의 비율로 보호할지 나타내는 정보 보

호 수준 비율 그리고 상기 K-평균 클러스터링을 실행하기 위한 클러스터 개수를 포함하는 입력 정보를 수신하고, 상기 수신한 정보들을 이용하여 상기 복수의 데이터를 쿼드 트리를 이용하여 복수의 버킷을 생성하며, 생성한 복수의 버킷에 대한 정보를 보호하여 차분 프라이버시를 만족하는 히스토그램을 생성하는 히스토그램 생성 모듈, 그리고 상기 차분 프라이버시를 만족하는 히스토그램과 상기 복수의 데이터를 상기 클러스터 개수만큼의 클러스터에 분류하는 K-평균 클러스터링 모듈을 포함한다.

[0015] 상기 히스토그램 생성 모듈은, 상기 데이터를 통해 생성되는 복수의 버킷에 포함될 데이터 수를 제한하기 위한 제1 임계값과, 동일한 버킷이 분할되는 분할 횟수를 제한하기 위한 제2 임계값을 계산하며, 상기 제1 임계값은 상기 데이터의 수를 토대로 계산하고, 상기 제2 임계값은 상기 데이터의 수와 상기 복수의 데이터를 이루를 좌표의 차원(dimension)을 토대로 계산할 수 있다.

[0016] 상기 히스토그램 생성 모듈은, 상기 정보 보호 수준과 상기 정보 보호 수준 비율을 이용하여, 상기 제1 정보를 보호하기 위한 제1 정보 보호 수준을 계산하고, 상기 제2 정보를 보호하기 위한 제2 정보 보호 수준을 계산하며, 상기 제1 정보는 상기 복수의 버킷 각각에 대한 구간 정보를 보호하기 위하여 분할 대상인 버킷의 구간에 대한 기준 정보이고, 상기 제2 정보는 상기 복수의 버킷 각각에 포함되는 데이터 수를 보호하기 위하여 상기 카운트 노이즈를 생성하는 기준 정보이다.

[0017] 상기 히스토그램 생성 모듈은, 상기 수신한 데이터의 분포를 확인하고 확인한 데이터 분포를 미리 설정된 수로 분할하여 복수의 초기 버킷을 생성하며, 상기 복수의 초기 버킷 각각에 포함된 데이터 수에 상기 카운트 노이즈가 반영된 데이터 수가 상기 제1 임계값보다 많은지, 상기 복수의 초기 버킷이 분할된 횟수가 상기 제2 임계값보다 적은지 확인하여 복수의 새로운 버킷으로 생성할 수 있다.

발명의 효과

[0018] 본 발명에 따르면 기존의 히스토그램 기반 알고리즘에 비해 적은 버킷을 만들어 데이터 분포를 표현함에도 클러스터 내 오차가 적어진다.

도면의 간단한 설명

[0019] 도 1은 본 발명의 실시예에 따른 데이터 마이닝 시스템의 구조도이다.

도 2는 본 발명의 실시예에 따른 클러스터링 방법에 대한 흐름도이다.

도 3은 본 발명의 실시예에 따른 쿼드 트리를 이용한 차분 프라이버시 L-평균 클러스터링 알고리즘의 예시도이다.

도 4는 본 발명의 실시예에 따른 차분 프라이버시를 만족하는 히스토그램의 예시도이다.

도 5는 본 발명의 실시예에 따라 K-평균 클러스터링을 수행한 예시도이다.

도 6은 본 발명의 실시예에 따라 차분 프라이버시를 만족하는 K-평균 클러스터링을 수행한 예시도이다.

도 7은 본 발명의 실시예에 따라 정보 보호 수준 비율에 따른 상대적 클러스터링의 성능 차이를 나타낸 예시도이다.

도 8은 본 발명의 실시예에 따라 정보 보호 수준에 따른 상대적 클러스터링 성능 차이를 나타낸 예시도이다.

도 9는 본 발명의 실시예에 따라 정보 보호 수준에 따른 버킷 숫자의 예시도이다.

발명을 실시하기 위한 구체적인 내용

[0020] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0021] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0022] 기존 차분 프라이버시 알고리즘으로 차분 프라이버시 Lloyd 클러스터링과, 확장된 등간격 그리드를 활용한 차분

프라이버시 k-평균 클러스터링이 있다.

[0023] k-평균 알고리즘은 각 클러스터를 구할 때, 이전 스텝에서 구한 클러스터들 중 가장 가까운 클러스터에 데이터가 속한다고 가정하고, 새로운 클러스터를 구한다. 이 과정에서, 좌표 값의 합과 클러스터에 속한 점의 개수를 카운트하는데, 여기에 각각 미리 계산된 노이즈(Lap(2*maxIter/ε))를 섞어주면서 알고리즘을 미리 설정한 총 반복 횟수(maxIter)만큼 반복하여 진행하면 ε-차분 프라이버시를 만족시킨다.

[0024] 이 방법의 경우 클러스터를 찾기 전에 총 몇 번 반복할 것인지(maxIter)를 정한 뒤 시작하고, 정보 보호 수준을 반복 횟수만큼 나누어 쓴다. 따라서 여러 번 반복할수록 삽입되는 노이즈의 양이 증가하게 된다. 한편 반복 횟수를 줄이면 클러스터의 중심을 찾기 전에 멈출 수 있다.

[0025] 확장된 등간격 그리드를 활용한 차분 프라이버시 k-평균 클러스터링은, 여러 번 평균을 구하더라도 삽입되는 노이즈가 증가하지 않도록 차분 프라이버시를 만족하는 데이터 분포를 찾고, 그 데이터 분포를 이용하여 차분 프라이버시 알고리즘이 아닌 일반 k-평균 알고리즘을 사용한다. 따라서 노이즈는 클러스터링 알고리즘의 반복 횟수와 상관없이 1번 삽입된다.

[0026] 하지만, 차분 프라이버시를 만족시키는 데이터의 분포를 찾을 때, 데이터를 등간격으로 나누어 히스토그램을 만드는데, 이 경우 데이터의 양이 커지거나, 정보 보호 수준이 증가하면 데이터를 나누어 생성된 구간인 버킷의 개수가 증가한다. 버킷의 숫자가 증가하면 데이터의 분포를 세밀하게 표현할 수 있지만, 노이즈가 버킷 수만큼 삽입되므로 삽입되는 노이즈의 양이 커진다.

[0027] 이러한 문제를 해결하기 위하여 버킷의 숫자를 줄이면서도 데이터의 분포를 표현하기 위해 본 발명의 실시예에서는 쿼드 트리를 이용하여 차분 프라이버시 k-평균 클러스터링 알고리즘을 사용한다. 이하 도면을 참조로 하여 본 발명의 실시예에 대해 상세히 설명한다.

[0028] 도 1은 본 발명의 실시예에 따른 데이터 마이닝 시스템의 구조도이다.

[0029] 도 1에 도시된 바와 같이, 적어도 하나 이상의 프로세서에 의해 동작하는 데이터 마이닝 시스템(100)은 히스토그램 생성 모듈(110), 그리고 K-평균 클러스터링 모듈(120)을 포함한다.

[0030] 히스토그램 생성 모듈(110)은 외부로부터 데이터, 정보 보호 수준(ε), 정보 보호 수준 비율(R), 그리고 클러스터 개수를 수신한다.

[0031] 데이터는 좌표로 표시될 수 있는 점들의 집합이다. 즉, n차원(여기서, n은 정수)의 좌표 형태로 표현 가능한 점들의 좌표 값을 데이터라고 지칭한다.

[0032] 히스토그램 생성 모듈(110)은 정보 보호 수준과 정보 보호 수준 비율을 이용하여, 제1 정보 보호 수준(ε_1)과 제2 정보 보호 수준(ε_2)을 계산한다. 제1 정보 보호 수준과 제2 정보 보호 수준을 계산할 때 정보 보호 수준 비율과 정보 보호 수준을 곱하여 생성하는 것은, 버킷 구간에 삽입할 노이즈와 버킷 데이터 내 데이터 수에 반영할 노이즈를 구하기 위함이다.

[0033] 본 발명의 실시예에서는 제1 정보 보호 수준(ε_1)은 하나의 버킷 구간을 나눌 때 $R*\varepsilon$이 되도록 나누고, 제2 정보 보호 수준(ε_2)은 하나의 버킷에 포함된 데이터 수가 임계값을 넘어가는 경우, 버킷에 포함된 데이터 수가 $(1-R)*\varepsilon$이 되도록 나누는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.

[0034] 그리고 히스토그램 생성 모듈(110)은 제1 임계값과 제2 임계값을 계산한다. 쿼드 트리 방식으로 데이터의 분포를 적절하게 표현하기 위하여 버킷을 생성할 때, 하나의 버킷 내에 포함되는 데이터의 수가 제1 임계값 보다 적게 포함되도록 하고, 동일한 버킷을 제2 임계값 보다 많이 분할하지 않게 하기 위한 기준 정보이다.

[0035] 히스토그램 생성 모듈(110)은 외부로부터 수신한 데이터를 미리 설정된 수로 나누어, 복수의 초기 버킷을 생성한다. 여기서 초기 버킷은 데이터 범위의 간격이 일정한 간격을 갖도록 나누어 생성된다.

[0036] 히스토그램 생성 모듈(110)은 등간격 그리드로 나뉘어 생성된 초기 버킷을 쿼드 트리를 이용하여 히스토그램을 생성한다. 즉, 히스토그램 생성 모듈(110)은 초기 버킷 각각에 포함된 데이터 수가, 계산한 제1 임계값보다 많은 초기 버킷이 있는지 확인한다.

[0037] 그리고 제1 임계값보다 많은 데이터를 가지는 초기 버킷이 있다면, 해당 초기 버킷이 제2 임계값보다 많이 나뉘었는지 확인한다. 만약 제2 임계값보다 적은 수로 초기 버킷이 나뉘었다면, 초기 버킷의 구간을 나누어 새로운

복수의 버킷을 생성한다. 그러나, 버킷이 제2 임계값 보다 많은 수로 나뉘는 것으로 판단되면, 히스토그램 생성 모듈(110)은 버킷의 분할을 중단한다.

[0038] 여기서, 히스토그램 생성 모듈(110)이 제1 임계값 보다 많은 데이터를 가지는 초기 버킷이 있는지 확인할 때, 먼저 버킷 내에 정확한 데이터 수를 카운트한다. 그리고 카운트한 수에 카운트 노이즈를 삽입한 값을 제1 임계값과 비교한다. 이를 위해, 히스토그램 생성 모듈(110)은 제2 정보 보호 수준을 고려하여 카운트 노이즈를 계산한다. 이는 각 버킷에 포함된 데이터 수가 외부에 정확하게 노출되지 않도록 하기 위함이며, 이를 프라이버시를 만족한다고 지칭한다.

[0039] 예를 들어, 제1 임계값이 10이고, 임의의 초기 버킷에 위치한 데이터 수가 9개 이며, 카운트 노이즈가 2.1로 계산되었다고 가정한다. 그러면, 히스토그램 생성 모듈(110)은 임의의 초기 버킷은 실제 데이터 수 9개와 노이즈 2.1이 더해진 11.1개의 데이터가 위치하는 것으로 계산한다.

[0040] 따라서, 히스토그램 생성 모듈(110)은 임의의 초기 버킷의 데이터 수가 11.1개로 제1 임계값인 10개를 넘는 것으로 확인한다. 이에 따라 히스토그램 생성 모듈(110)은 해당 버킷을 나누게 된다. 히스토그램 생성 모듈(110)이 나눈 해당 버킷의 구간은 차분 프라이버시를 만족하는 구간이 된다.

[0041] 이때, 초기 버킷 또는 초기 버킷이 나뉘어 새로 생성된 버킷을 다시 나누는 방법이나 초기 버킷으로부터 생성된 복수의 버킷의 수를 어느 하나의 수로 한정하지 않는다. 그리고 동일한 버킷이 제2 임계값 보다 많이 나뉘었는지 판단하는 방법 또한 다양한 방법으로 수행할 수 있으므로, 본 발명의 실시예에서는 어느 하나의 방법으로 한정하지 않는다.

[0042] 히스토그램 생성 모듈(110)은 각 버킷의 리프 노드 수에 카운트 노이즈를 삽입하여 생성한 카운트 정보와, 분할되어 생성된 최종 버킷들의 구간 정보를 포함하여 히스토그램을 생성한다. 이렇게 생성된 히스토그램은 차분 프라이버시를 만족하는 히스토그램이 된다.

[0043] K-평균 클러스터링 모듈(120)은 히스토그램 생성 모듈(110)이 수신한 클러스터 개수 정보를 이용하여, 클러스터의 중심과 분포되어 있는 개별 데이터 사이의 거리를 계산하여 클러스터에 할당한 후 클러스터 수만큼의 클러스터 중심 정보를 출력한다. 즉, K-평균 클러스터링 알고리즘을 이용하여 클러스터링을 수행한다. 이때, 데이터는 차분 프라이버시를 만족하는 히스토그램에 따라 복수의 버킷의 중심에 분포되어 있다.

[0044] 그리고 K-평균 클러스터링 모듈(120)은 데이터를 클러스터링한 결과를 출력한다. 여기서 클러스터링한 결과는 어느 하나의 형태로 한정하지 않으며, K-평균 클러스터링 알고리즘은 이미 알려진 것으로 본 발명의 실시예에서는 데이터를 K-평균 클러스터링하는 상세한 설명을 생략한다.

[0045] 이상에서 설명한 데이터 마이닝 시스템(100)을 이용하여 클러스터링을 수행하는 방법에 대해 도 2를 참조로 설명한다.

[0046] 도 2는 본 발명의 실시예에 따른 클러스터링 방법에 대한 흐름도이다.

[0047] 도 2에 도시된 바와 같이, 데이터 마이닝 시스템(100)은 좌표들로 이루어진 복수의 데이터, 정보 보호 수준, 정보 보호 수준 비율, 그리고 클러스터 개수를 수신한다(S100). 그리고, 정보 보호 수준과 정보 보호 수준 비율을 토대로, 제1 정보 보호 수준과 제2 정보 보호 수준을 계산한다(S101). 제1 정보 보호 수준은 정보 보호 비율에 정보 보호 수준을 곱한 값(R*ε)으로 계산하고, 제2 정보 보호 수준은 1에서 정보 보호 비율을 뺀 값에 정보 보호 수준을 곱한 값((1-R)*ε)으로 계산한다.

[0048] 데이터 마이닝 시스템(100)은 제1 임계값과 제2 임계값을 계산한다(S102). 여기서 제1 임계값은 쿼드 트리 방식으로 데이터의 분포를 적절하게 표현하기 위하여 버킷을 생성할 때, 하나의 버킷 내에 포함되는 데이터의 수가 제1 임계값 보다 적게 포함되도록 하기 위한 기준 정보이다. 그리고 제2 임계값은 동일한 버킷을 제2 임계값 보다 많이 자르지 않게 하기 위한 기준 정보이다. 제1 임계값과 제2 임계값은 다음 수학식 1을 이용하여 계산하는 것을 예로 하여 설명하나, 반드시 이와 같이 한정되는 것은 아니다.

[0049] [수학식 1]

[0050] 제1 임계값 = 1/1000*n

[0051] 제2 임계값 = $\frac{1}{d}\log n - 1$

[0052] 여기서, 여기서 n은 점의 개수인 정수이고, d는 원본 데이터의 차원(dimension)을 의미한다.

[0053] 상기의 절차를 통해 제1 정보 보호 수준, 제2 정보 보호 수준, 제1 임계값 그리고 제2 임계값이 계산되면, 데이터 마이닝 시스템(100)은 S100 단계에서 수신한 데이터로부터 복수의 초기 버킷을 생성한다(S103). 이때, 초기 버킷은 미리 설정된 수만큼 생성되며, 각각의 초기 버킷의 구간 즉, 데이터 범위는 모든 초기 버킷이 동일하게 생성된다.

[0054] 복수의 초기 버킷이 생성되면, 데이터 마이닝 시스템(100)은 각각의 초기 버킷에 포함된 데이터 수에 카운트 노이즈를 삽입한다(S104). 그리고 카운트 노이즈가 삽입된 버킷 각각에 대해, 각각의 초기 버킷에 포함되어 있는 데이터 수가 제1 임계값보다는 많은지 확인한다(S105). 이는 종래의 기술을 토대로 데이터를 등간격으로 나눌 경우, 데이터가 없는 영역에도 버킷이 생성되어 노이즈 삽입 횟수가 늘어나고, 데이터가 많은 곳에도 등간격으로 버킷이 나뉘기 때문에 데이터 분포를 적절하게 표현할 수 없는 단점을 해결하기 위함이다.

[0055] 만약 제1 임계값 보다 많은 데이터를 갖는 버킷이 있다면, 데이터 마이닝 시스템(100)은 해당 버킷의 분할 횟수가 제2 임계값보다 많은지 확인한다(S106). 여기서, 데이터 마이닝 시스템(100)이 동일한 버킷이 얼마나 분할되었는지 확인하는 방법은 다양하게 수행할 수 있으므로, 본 발명의 실시예에서는 상세한 설명을 생략한다.

[0056] 만약 제2 임계값 보다 적은 분할 횟수를 가진다면, 데이터 마이닝 시스템(100)은 해당 버킷을 분할하여 새로운 버킷들을 생성한다(S107). 그리고 S104 단계 이후의 절차를 반복한다.

[0057] 그러나, 버킷의 분할 횟수가 제2 임계값을 초과하거나, 버킷 내 데이터 수가 제1 임계값 보다 많은 버킷이 없다면 데이터 마이닝 시스템(100)은 분할되어 생성된 최종 버킷들 각각의 구간 정보와, 해당 버킷에 포함되어 있는 카운트 노이즈가 삽입된 데이터 수를 포함하는 히스토그램을 생성한다(S108). 이렇게 생성된 히스토그램은 차분 프라이버시를 만족하게 된다.

[0058] 데이터 마이닝 시스템(100)은 S108 단계에서 히스토그램이 생성되면, 생성된 히스토그램을 참고하여 S100 단계에서 수신한 데이터를 K-평균 클러스터링 알고리즘으로 클러스터링을 수행한다(S109).

[0059] S109 단계에서 클러스터링을 수행할 때, 각 버킷 내 데이터들이 버킷의 중심에 카운트 노이즈가 삽입된 데이터 수만큼 존재한다고 가정하고 S100 단계에서 수신한 클러스터 개수만큼 클러스터링을 수행한다. K-평균 클러스터링 알고리즘이 히스토그램을 클러스터링하는 방법은 일반적인 클러스터링 방법과 동일하므로, 본 발명의 실시예에서는 상세한 설명을 생략한다.

[0060] 이상에서 설명한 방법을 프로그램으로 나타내면 다음 표 1과 같다.

표 1

[0061]

Input: 데이터 $D = \{x_1, \dots, x_n\}$, $x_i \in \mathbb{R}^d$ 정보 보호 수준 ϵ 정보 보호 수준 비율 R 클러스터 개수 k Output: $H = \{(r_1, f_1'), (r_2, f_2'), \dots, (r_B, f_B')\}$
1. $\epsilon_1 \leftarrow R * \epsilon$, $\epsilon_2 \leftarrow (1-R) * \epsilon$ 2. maxDepth $\leftarrow \frac{1}{d}\log n - 1$, Threshold $\leftarrow \frac{1}{1000}n$ 3. T $\leftarrow$ build Quadtree(D, ϵ_1, maxDepth, Threshold) 4. Leaves $\leftarrow$ find leaf nodes in Quad Tree T 5. For each leaf in Leaves 6. $f_i' = f_i + Lap(1/\epsilon_2)$ 7. $r_i \leftarrow$ center of leaf node's rectangle 8. $D' \leftarrow D' + \{(r_i, f_i')\}$ 9. C $\leftarrow$ Kmeans(D', k) 10. Return $C = \{c_1, \dots, c_k\}$

[0062] 표 1에 나타낸 바와 같이, 데이터 마이닝 시스템(100)은 원본 데이터 D, 쿼드 트리에 적용하는 정보 보호 수준인 ε, 정보 보호 수준 비율 R, 클러스터 개수 k개를 입력으로 받는다.

[0063] 정보 보호 수준과 원본 데이터를 수신한 데이터 마이닝 시스템(100)은 정보 보호 수준을 제1 정보 보호 수준(ε_1)과 제2 정보 보호 수준(ε_2)으로 나눈다. 제1 정보 보호 수준(ε_1)은 R*ε이 되도록 나누고, 제2 정보 보호 수

준(ε_2)은 (1-R)*ε이 되도록 나눈다. 제1 정보 보호 수준과 제2 정보 보호 수준을 계산할 때 정보 보호 수준 비율과 정보 보호 수준을 곱하여 생성하는 것은, 버킷 구간에 삽입할 노이즈와 버킷 데이터 내 데이터 수에 반영할 노이즈를 구하기 위함이다.

[0064] 그리고, 데이터 마이닝 시스템(100)은 제1 임계값(Threshold)과 제2 임계값(maxDepth)을 계산한다. 제1 임계값은 버킷 내에 포함된 데이터의 수의 상한 임계값에 해당하는 것으로 1/1000*n으로 계산된다. 그리고, 제2 임계값은 임의의 구역에 형성된 버킷이 maxDepth번 이상 나뉘지 않도록 하는 임계값으로, $\frac{1}{d}\log n-1$ 로 계산된다. 여기서 n은 점의 개수인 정수이고, d는 원본 데이터의 차원(dimension)을 의미한다.

[0065] 데이터 마이닝 시스템(100)은 정보 보호 수준에서 제1 정보 보호 수준을 계산하고 제1 임계값과 제2 임계값을 계산하면, 원본 데이터, 제1 정보 보호 수준, 제1 임계값 그리고 제2 임계값을 이용하여 쿼드 트리(T)를 생성한다. 쿼드 트리를 생성하는 것은, 히스토그램의 버킷 범위를 구하는 것으로, 제1 정보 보호 수준과 관련된 노이즈를 기본 범위에 더해 생성하므로, 하나의 버킷 범위가 정확하게 외부에 노출되지 않도록 즉, 버킷 범위에 대한 정보를 보호하며 노출된다.

[0066] 이에 대해 도 3을 참조로 먼저 설명한다.

[0067] 도 3은 본 발명의 실시예에 따른 차분 프라이버시를 만족하는 히스토그램의 예시도이다.

[0068] 도 3에 도시된 바와 같이, 원으로 표현된 데이터들이 다양한 위치에 분포되어 있고, 각각의 데이터들의 좌표가 2차원 좌표로 이루어졌다고 가정한다.

[0069] 기존에는 버킷을 나눌 때 등간격 그리드를 활용하여 각각의 버킷의 구간이 일정한 간격을 가지도록 나누어 생성하였다. 그러나, 본 발명의 실시예에서는 제1 정보 보호 수준, 제1 임계값, 그리고 제2 임계값을 이용하여 쿼드 트리를 생성할 때, 전체 버킷의 수를 줄이면서도 각각의 버킷의 구간이 일정 간격을 갖지 않도록 나누어, 버킷 각각에 포함되어 있는 데이터 수와 버킷 각각의 간격 정보로 구성된 히스토그램의 정보가 보호되도록 한다.

[0070] 한편, 상기 도 3을 이어 설명하면, 히스토그램의 정보가 보호된 버킷 범위를 구한 후, 데이터 마이닝 시스템(100)은 히스토그램의 버킷 안에 몇 개의 데이터가 있는지(f)를 구한다. 여기서 하나의 버킷 안에 포함된 데이터 수를 구할 때, 제2 정보 보호 수준과 관련된 노이즈를 버킷 내 데이터 수에 반영하여 구한다. 이에 대해 도 4 및 도 5를 참조로 먼저 설명한다.

[0071] 도 4는 본 발명의 실시예에 따라 K-평균 클러스터링을 수행한 예시도이고, 도 5는 본 발명의 실시예에 따라 차분 프라이버시를 만족하는 K-평균 클러스터링을 수행한 예시도이다.

[0072] 도 4에 도시된 수치들은 상기 도 4에서 생성한 버킷의 각각의 내부에 위치한 데이터의 수이고, 도 5에 도시된 수치들은 노이즈를 버킷 내 데이터 수에 반영하여 생성한 데이터의 수이다. 도 4에서 계산된 각각의 버킷 내부에 포함된 데이터 수에 제2 정보 보호 수준으로 계산된 노이즈를 반영하여 도 5와 같이 계산할 경우, 각 버킷 내 데이터 수치가 외부에 정확하게 노출되지 않으므로 정보를 보호할 수 있다.

[0073] 한편, 상기 도 3을 이어 설명하면, 버킷 내 데이터 수에 노이즈를 반영하여 구한 후, 데이터 마이닝 시스템(100)은 버킷의 중심(ri)에 모든 데이터들이 위치해 있다고 가정하고, K-평균 클러스터링을 수행한다. 이때, 입력으로 받은 클러스터 개수만큼 클러스터링을 수행하여, 클러스터 개수만큼의 클러스터 중심을 출력한다. K-평균 클러스터링 알고리즘은 다음 표 2와 같다.

표 2

[0074]

Input: 데이터 범위 $R = \{(m_1, M_1), \ldots, (m_d, M_d)\}$
클러스터 개수 k
Output: 클러스터 중심 k개 $C = \{c_1, \ldots, c_k\}$

```
1. length ← (∑_i (M_i − m_i)^2)^0.5
2. While(1)
3.     for (i=0; i<100; ++i)
4.         C ← random sample k centers in domain R
5.         flag ← true
6.         for i, j in |C|
7.             if |c_i-c_j| < length
8.                 flag ← false
9.         if flag = true
10.            return C = {c_1, ..., c_k}
11.    length ← length * 0.9
```

[0075] 데이터들이 분포하고 있고 클러스터링을 시작할 때 클러스터들의 중심 좌표를 정해주어야 한다. 초기 클러스터들의 중심들이 한 곳에 모여 있으면 클러스터링이 제대로 되지 않으므로, 표 2에 나타낸 알고리즘은 떨어진 점들을 찾아 이를 클러스터의 중심으로 찾기 위한 알고리즘이다.

[0076] 입력으로 받은 데이터의 범위(R)는 점들이 존재할 수 있는 최소값, 최대값들을 나타낸 것이다. (키, 몸무게)로 예를 들어 설명하면 (0cm , 200cm), (0kg, 100kg) 등과 같이 나타낸다.

[0077] 데이터 범위가 결정되면, 해당 범위가 형성하는 공간에서 떨어질 수 있는 가장 먼 길이 즉, 대각선을 찾는다. 대각선 길이만큼 떨어진 점들이 k개가 있다면, 이를 클러스터들의 중심으로 쓰고자 한다.

[0078] 맨 처음에는 가장 긴 길이 값(length)이 찾아졌기에, k개의 점들이 length만큼 떨어져있지 않는다. 따라서, 임의의 점들을 샘플링하여 길이가 length만큼 떨어져있는지 확인하고, 떨어져있다면 임의의 점들을 클러스터들의 중심으로 사용한다. 그런데, 샘플링을 최대 반복 횟수만큼 반복해도 길이가 length 만큼 떨어진 점들이 찾아지지 않는다면, length 길이를 줄여서 점들이 조금만 떨어져있어도 선택 되게끔 조건을 완화시켜 준다. 이러한 과정을 반복하여 클러스터 중심으로 선택할 점 k개를 찾는다.

[0079] 이상에서 설명한 알고리즘에서 사용된 정보 보호 수준 비율(R)은 제1 정보 보호 수준과 제2 정보 보호 수준을 얼만큼씩 나눌지 결정하는 비율이다. 이에 대해 도 6을 참조로 설명한다.

[0080] 도 6은 본 발명의 실시예에 따라 정보 보호 수준 비율에 따른 상대적 클러스터링의 성능 차이를 나타낸 예시도이다.

[0081] 도 6에 도시된 바와 같이 정보 보호 수준 비율(R)에 따른 오차를 확인할 때, 정보 보호 수준 비율(R)이 작으면 쿼드 트리를 구할 때 노이즈를 많이 더하게 되어(Lap(1/Rε)) 데이터의 분포가 세밀하게 찾아지지 않을 수 있다. 그러나 각 버킷의 데이터 수에는 노이즈가 적게 더해져(Lap(1/(1-R)ε)) 상대적으로 정확한 데이터 수를 갖게 된다.

[0082] ε = 0.05 일 때, R = 0.15에서 상대적인 클러스터링의 성능이 좋게 나오는 것을 확인할 수 있다. 여기서. 본 발명의 실시예에서 사용한 정보 보호 수준 비율(R) 0.15는 반드시 이와 같이 한정되는 것은 아니다.

[0083] 도 7은 본 발명의 실시예에 따라 정보 보호 수준에 따른 상대적 클러스터링 성능 차이를 나타낸 예시도이고, 도 8은 본 발명의 실시예에 따라 정보 보호 수준에 따른 버킷 숫자의 예시도이다.

[0084] 도 7에는 정보 보호 수준(ε)에 따라 상대적 클러스터링 성능의 차이를 나타내었다. 정보 보호 수준(ε)에 따라 성능이 다르지만, 최대 25% 가량의 성능 향상이 있었다. 이는 등간격으로 나누어 히스토그램을 만드는 것보다 쿼드 트리가 k-평균 클러스터링을 위한 더 적합한 데이터 분포를 찾아 내는 것을 뜻한다.

[0085] 그리고 본 발명의 실시예에서는 정보 보호 수준(ε)에 따라 버킷의 개수를 다르게 나눌 수 있다. 정보 보호 수

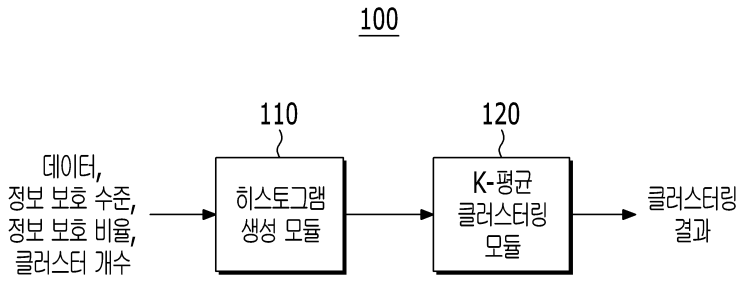
준(ε)에 따라 만드는 버킷의 숫자는 도 8에 도시된 바와 같다.

[0086] EUGkDP는 정보 보호 수준(ε)이 증가함에 따라 버킷의 숫자가 증가하는 반면, 본 발명의 실시예에서는 정보 보호 수준(ε)에 무관하게 적은 리프 노트 수를 유지한다. 이는 노이즈 삽입 횟수가 적을 뿐 아니라 대량의 데이터에 대해 클러스터링을 더 빠르게 수행할 수 있음을 의미한다.

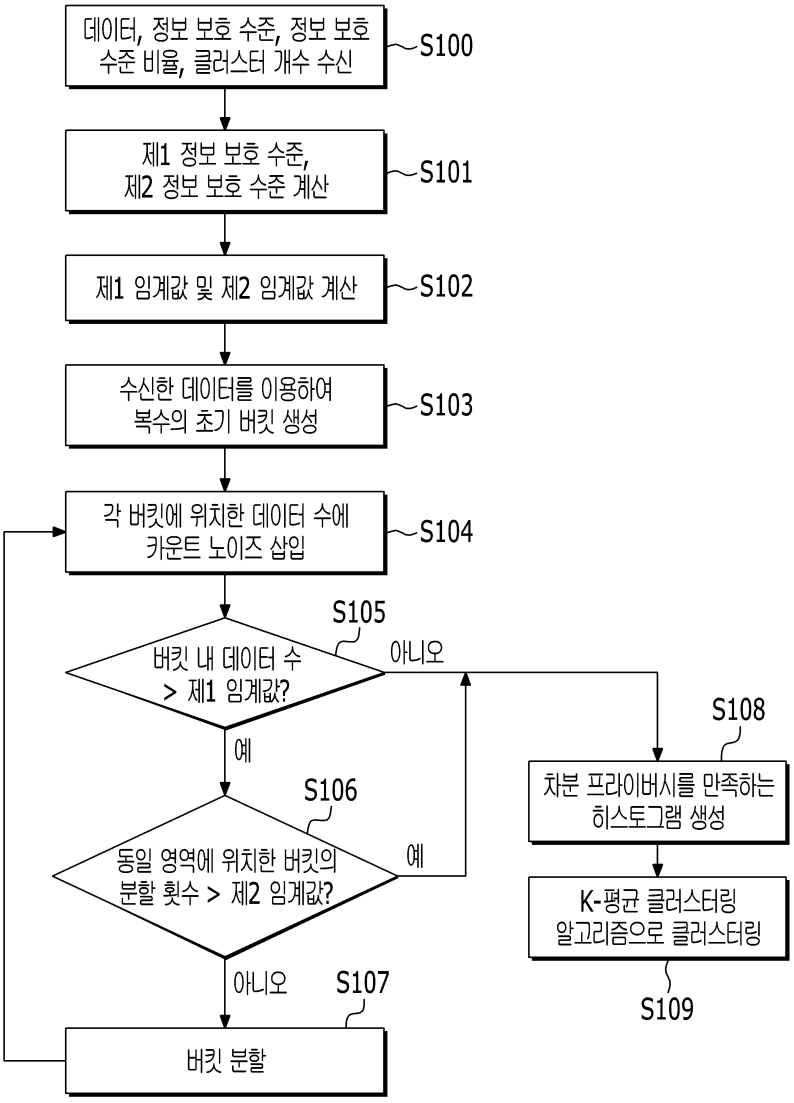
[0087] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

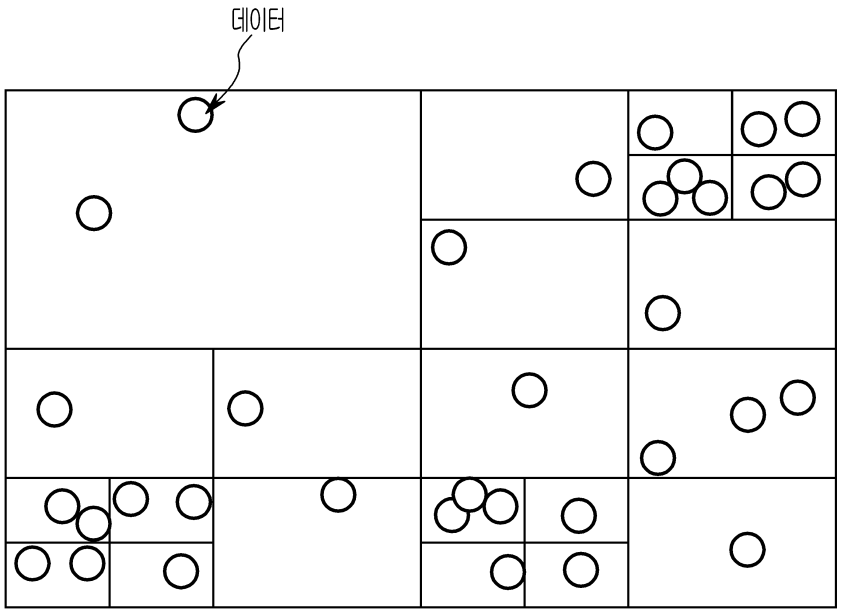
도면1



도면2


데이터, 정보 보호 수준, 정보 보호 수준 비율, 클러스터 개수 수신
S100
제1 정보 보호 수준, 제2 정보 보호 수준 계산
S101
제1 임계값 및 제2 임계값 계산
S102
수신한 데이터를 이용하여 복수의 초기 버킷 생성
S103
각 버킷에 위치한 데이터 수에 카운트 노이즈 삽입
S104
S105
버킷 내 데이터 수 > 제1 임계값?
아니오
예
S106
동일 영역에 위치한 버킷의 분할 횟수 > 제2 임계값?
예
아니오
S107
버킷 분할
S108
차분 프라이버시를 만족하는 히스토그램 생성
K-평균 클러스터링 알고리즘으로 클러스터링
S109

도면3



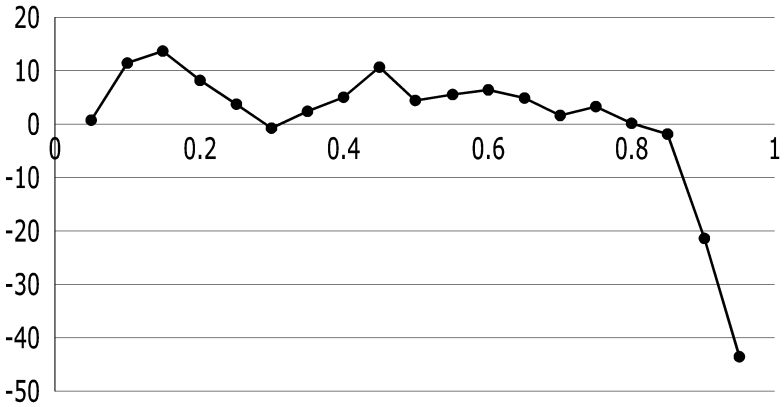
도면4

<table>
<tr><td colspan="2" rowspan="2">1400</td><td colspan="2" rowspan="2">105</td><td>1101</td><td>870</td></tr>
<tr><td>1570</td><td>1750</td></tr>
<tr><td colspan="2"></td><td colspan="2">207</td><td colspan="2">507</td></tr>
<tr><td colspan="2">1831</td><td>480</td><td colspan="2">1700</td><td colspan="2">2103</td></tr>
<tr><td>610</td><td>2300</td><td rowspan="2">570</td><td>205</td><td>1608</td><td rowspan="2">1320</td></tr>
<tr><td>2510</td><td>1003</td><td>1306</td><td>2110</td></tr>
</table>

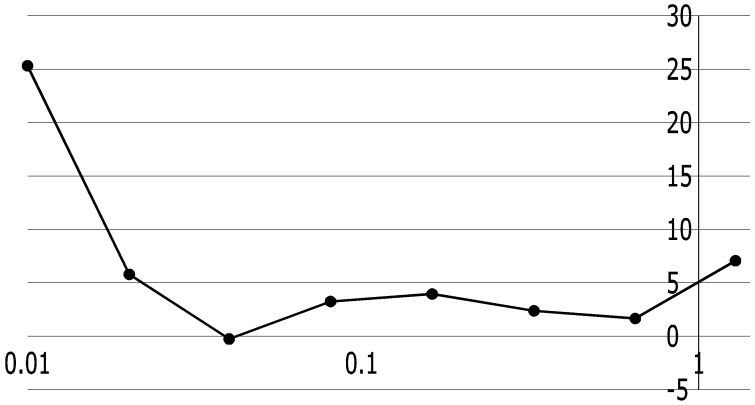
도면5

<table>
<tr><td colspan="2" rowspan="2">1421.1</td><td rowspan="2">50.3</td><td>1100.1</td><td>827.2</td></tr>
<tr><td>1510.2</td><td>1719.9</td></tr>
<tr><td colspan="2"></td><td>300.6</td><td colspan="2">471.3</td></tr>
<tr><td colspan="2">1871.2</td><td>402.5</td><td colspan="2">1503.2</td><td colspan="2">2121.2</td></tr>
<tr><td>471.3</td><td>2271.2</td><td rowspan="2">501.2</td><td>200.2</td><td>1602.2</td><td rowspan="2">1321.1</td></tr>
<tr><td>2201.1</td><td>1003.1</td><td>1311.1</td><td>2052.3</td></tr>
</table>

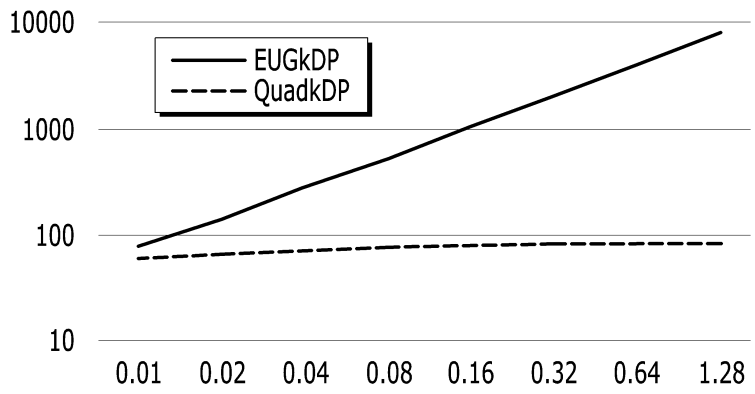
도면6



도면7



도면8


10000
1000
100
10
EUGkDP
QuadkDP
0.01
0.02
0.04
0.08
0.16
0.32
0.64
1.28