



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0067307
(43) 공개일자 2011년06월22일

(51) Int. Cl.

G06F 17/30 (2006.01) G06F 17/20 (2006.01)

G06F 15/16 (2006.01)

(21) 출원번호 10-2009-0123856

(22) 출원일자 2009년12월14일

심사청구일자 2009년12월14일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가 1

(72) 발명자

정익래

서울특별시 노원구 공릉동 81번지 태강아파트
1014동 1303호

김기성

경기도 군포시 산본동 1063 백합아파트 1126동
1503호

(74) 대리인

특허법인 엠에이피에스

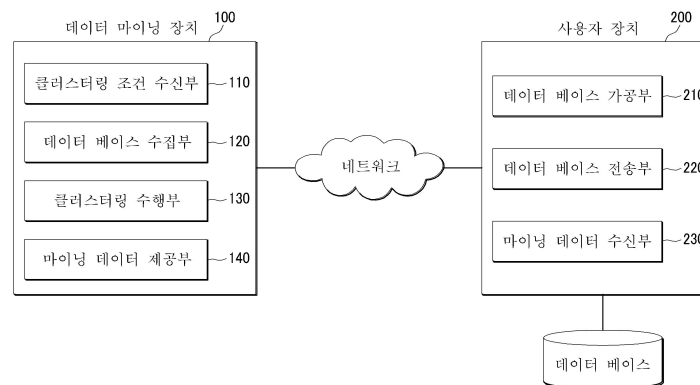
전체 청구항 수 : 총 17 항

(54) 프라이버시를 보존하는 데이터 마이닝 시스템 및 장치 그리고 그 방법

(57) 요약

데이터 마이닝 시스템은, 원본 데이터 베이스에 설정된 개수의 더미 데이터를 추가하여 변형 데이터 베이스를 생성하고, 변형 데이터 베이스를 데이터 마이닝 장치에 제공하는 사용자 장치 및 설정된 클러스터링 조건에 따라 변형 데이터 베이스를 클러스터링하는 데이터 마이닝 장치를 포함한다. 이와 같은, 데이터 마이닝 시스템은 간단한 연산만으로도 원본 데이터 베이스에 대한 프라이버시를 보존하면서도 동일한 데이터 마이닝 결과를 얻을 수 있다.

대표도 - 도1



이 발명을 지원한 국가연구개발사업

과제고유번호 2009-F-056-01

부처명 정보통신산업진흥원

연구관리전문기관

연구사업명 정보통신 산업원천기술개발

연구과제명 Car- 헬스케어 보안 기술개발

기여율

주관기관 고려대학교 산학협력단

연구기간 2009년 03월 01일 ~ 2010년 02월 28일

특허청구의 범위

청구항 1

데이터 마이닝 시스템에 있어서,

원본 데이터 베이스에 설정된 개수의 더미 데이터를 추가하여 변형 데이터 베이스를 생성하고, 변형 데이터 베이스를 데이터 마이닝 장치에 제공하는 사용자 장치; 및

설정된 클러스터링 조건에 따라 상기 변형 데이터 베이스를 클러스터링하는 데이터 마이닝 장치를 포함하는 데이터 마이닝 시스템.

청구항 2

제 1 항에 있어서,

상기 사용자 장치는,

상기 원본 데이터 베이스의 전체 데이터 개수, 데이터 속성 및 데이터 밀도에 기초하여 설정된 프라이버시 레벨에 따라 상기 더미 데이터의 개수를 결정하는 데이터 마이닝 시스템.

청구항 3

제 1 항에 있어서,

상기 클러스터링 조건은,

밀도 기반 공간 클러스터링(Density Based Spatial Clustering of Applications with Noise) 기법에 적용되는 기준 반경 및 최소 이웃 개수인 데이터 마이닝 시스템.

청구항 4

제 1 항에 있어서,

상기 데이터 마이닝 장치는,

클러스터링의 프라이버시 능력 및 정확도에 기초하여 더미 데이터 개수의 한계 비율을 설정하고, 상기 설정된 더미 데이터 개수의 한계 비율을 복수의 사용자 장치에 제공하는 데이터 마이닝 시스템.

청구항 5

제 4 항에 있어서,

상기 사용자 장치는,

상기 원본 데이터 베이스의 데이터 개수 중 상기 더미 데이터 개수의 한계 비율만큼의 개수를 산출하고, 상기 산출된 개수의 더미 데이터를 상기 원본 데이터 베이스에 삽입하는 데이터 마이닝 시스템.

청구항 6

제 1항에 있어서,

상기 사용자 장치는,

상기 원본 데이터 베이스의 전체 데이터를 정규화한 후 상기 더미 데이터를 삽입하는 데이터 마이닝 시스템.

청구항 7

데이터 마이닝 장치에 있어서,

데이터베이스를 밀도 기반 공간 클러스터링(Density Based Spatial Clustering of Applications with Noise)하기 위한 기준 반경 및 최소 이웃 개수를 수신하는 클러스터링 조건 수신부;

원본 데이터 베이스에 설정된 개수의 더미 데이터를 삽입한 변형 데이터 베이스를 수신하여 관리하는 데이터 베이스 수집부; 및

상기 변형 데이터 베이스를 상기 수신한 기준 반경 및 최소 이웃 개수를 적용하여 클러스터링하여 클러스터(cluster)를 생성하는 클러스터링 수행부를 포함하는 데이터 마이닝 장치.

청구항 8

제 7 항에 있어서,

상기 생성된 클러스터를 상기 변형 데이터 베이스 제공 측으로 전송하는 마이닝 데이터 제공부를 더 포함하는 데이터 마이닝 장치.

청구항 9

제 7 항에 있어서,

클러스터링의 프라이버시 능력 및 정확도에 기초하여 더미 데이터 개수의 한계 비율을 설정하고, 상기 더미 데이터 개수의 한계 비율을 적어도 하나의 변형 데이터 베이스 제공 측으로 전송하는 가공 조건 전송부를 더 포함하는 데이터 마이닝 장치.

청구항 10

제 7 항에 있어서,

상기 변형 데이터 베이스는,

상기 원본 데이터 베이스의 속성에 기초하여 설정된 프라이버시 레벨에 따라 결정된 개수의 더미 데이터를 삽입하여 가공되는 데이터 마이닝 장치.

청구항 11

제 7 항에 있어서,

상기 클러스터링 수행부는,

복수의 변형 데이터 베이스에 대한 클러스터링 시, 상기 복수의 변형 데이터 베이스의 합집합을 클러스터링하는 데이터 마이닝 장치.

청구항 12

데이터 마이닝 방법에 있어서,

원본 데이터 베이스의 속성에 기초하여 더미 데이터의 개수를 설정하는 단계;

상기 원본 데이터 베이스에 상기 설정된 개수의 더미 데이터를 삽입하여 변형 데이터 베이스를 생성하는 단계; 및

상기 변형 데이터 베이스를 설정된 기준 반경 및 최소 이웃 개수를 적용하여 밀도 기반 공간 클러스터링 (Density Based Spatial Clustering of Applications with Noise)하는 단계를 포함하는 데이터 마이닝 방법.

청구항 13

제 12 항에 있어서,

상기 더미 데이터의 개수를 설정하는 단계에서,

상기 원본 데이터 베이스의 전체 데이터 개수, 데이터 속성 값 및 데이터 밀도에 기초하여 설정된 프라이버시 레벨에 따라 상기 더미 데이터 개수를 설정하는 데이터 마이닝 방법.

청구항 14

제 12 항에 있어서,

상기 더미 데이터의 개수를 설정하는 단계는,

복수의 원본 데이터 베이스의 전체 데이터 개수, 데이터 속성 값 및 데이터 밀도에 기초하여 더미 데이터 개수 한계 비율을 설정하는 단계; 및

각 원본 데이터 베이스의 전체 데이터 개수 중 상기 더미 데이터 개수 한계 비율만큼의 개수를 각각의 더미 데이터의 개수로 결정하는 단계를 포함하는 데이터 마이닝 방법.

청구항 15

제 12 항에 있어서,

상기 밀도 기반 공간 클러스터링하는 단계 이후에,

상기 밀도 기반 공간 클러스터링하여 생성된 클러스터를 상기 변형 데이터 베이스의 생성 측으로 전송하는 단계를 더 포함하는 데이터 마이닝 방법.

청구항 16

제 12 항에 있어서,

상기 밀도 기반 공간 클러스터링하는 단계 이전에,

복수의 변형 데이터 베이스를 수신하는 단계; 및

상기 복수의 변형 데이터 베이스를 합집합하는 단계를 더 포함하는 데이터 마이닝 방법.

청구항 17

제 12 항에 있어서,

상기 변형 데이터 베이스를 생성하는 단계에서,

상기 원본 데이터 베이스를 정규화한 후 상기 더미 데이터를 삽입하는 데이터 마이닝 방법.

명세서

발명의 상세한 설명

기술 분야

[0001] 본 발명은 데이터 마이닝(DATA MINING) 시스템 및 장치와 그 방법에 관한 것으로, 더욱 상세하게는 밀도 기반 공간 클러스터링(DBSCAN: Density Based Spatial Clustering of Applications with Noise) 방식을 적용하여 원본 데이터 베이스의 프라이버시를 보존하는 데이터 마이닝 시스템 및 장치와 그 방법에 관한 것이다.

배경 기술

[0002] 최근 IT 기술의 급속한 발전으로 인해 엄청난 양의 데이터 축적이 사회 각 전반에 나타나고 있다. 예를 들어, 백화점에서는 하루에도 수만 건의 거래 내역이 자료화 되어 저장되고, 병원에서는 수많은 환자의 진료 기록이 저장된다. 대용량 자료에서 의미 있는 정보를 찾아내는 데이터 마이닝은 이러한 환경에서 그 역할이 점점 더 증대되고 있다.

[0003] 이러한 데이터 마이닝 기법 중 클러스터링(Clustering) 방식은 서로 유사한 속성을 가지는 데이터 그룹을 형성하여 그룹 사이의 유사성은 최소화하고, 그룹 내의 유사성은 최대화 시키는 기법이다.

[0004] 예를 들어, 클러스터링을 통해 백화점에서는 구매 내역의 유사성을 토대로 고객을 분류하여 관리할 수 있고, 병원에서는 특정 질환에 따라 환자들이 갖는 유사성을 찾아낼 수도 있다.

[0005] 그런데, 데이터 마이닝 기법에서는 일반적으로 프라이버시 문제가 수반된다. 예를 들어, 서로 경쟁 관계에 있는 두 백화점이 서로의 이익 증대를 위한 수단으로 거래 내역을 요청할 수도 있다. 또한, 클라우드 컴퓨팅(cloud computing) 환경에서 자료 처리 능력이 부족한 사용자가 외부 업체에 원본 자료에 대한 데이터 마이닝 작업을 요청할 수 있다. 이러한 환경에서 자료를 원본 그대로 교환하게 되면, 개인의 사생활 정보가 그대로 노출될 수 있다는 문제점이 있다.

[0006] 이러한 문제를 해결하기 위해 프라이버시를 보존할 수 있는 데이터 마이닝 방법들이 제안되고 있다. 프라이버시 보존을 위한 데이터 마이닝은 크게 2가지 방식으로 연구가 진행되어 왔다.

[0007] 첫 번째는 데이터 변형 방식으로, 데이터를 원본과 다르게 변형시켜 마이너(Miner)에게 제공하는 방식이다. 구체적으로, 데이터의 소유주가 원본 데이터에 노이즈 값을 추가하여 원본 데이터의 형태를 변화시켜 프라이버시를 보존하는 기법이다. 하지만, 이 기법은 연산량이 많고, 안전성 및 정확성 등에 문제가 지적되고 있다.

[0008] 두 번째는 다자간 환경을 기반으로 접근한 방식으로, 다수의 데이터 베이스의 소유주들이 서로의 자료에 대한 프라이버시를 보존하면서 전체적으로 만족하는 군집화 결과를 얻는 방식이다. 이 기법은 이미 안전성이 확보된 다자간 계산 방식(Secure multi-party computation)을 적용한다. 하지만, 클라우드 컴퓨팅 환경처럼 특정한 환경에서는 적용할 수 없는 제약이 있는 기법이다.

발명의 내용

해결 하고자하는 과제

[0009] 본 발명의 일 실시예는, 적은 연산으로도 원본 데이터 베이스의 프라이버시를 보존하는 데이터 마이닝 시스템 및 장치 그리고 그 방법을 제공한다.

[0010] 또한, 본 발명의 다른 실시예는, 원본 데이터 베이스의 속성에 기초하여 프라이버시 레벨을 결정하는 데이터 마이닝 시스템 및 장치 그리고 그 방법을 제공한다.

과제 해결수단

[0011] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 본 발명의 제 1 측면에 따른 데이터 마이닝 시스템은, 원본 데이터 베이스에 설정된 개수의 더미 데이터를 추가하여 변형 데이터 베이스를 생성하고, 변형 데이터 베이스를 데이터 마이닝 장치에 제공하는 사용자 장치; 및 설정된 클러스터링 조건에 따라 상기 변형 데이터 베이스를 클러스터링하는 데이터 마이닝 장치를 포함한다.

[0012] 그리고, 본 발명의 제 2 측면에 따른 데이터 마이닝 장치는, 데이터베이스를 밀도 기반 공간 클러스터링(Density Based Spatial Clustering of Applications with Noise)하기 위한 기준 반경 및 최소 이웃 개수를 수신하는 클러스터링 조건 수신부; 원본 데이터 베이스에 설정된 개수의 더미 데이터를 삽입한 변형 데이터 베이스를 수신하여 관리하는 데이터 베이스 수집부; 및 상기 변형 데이터 베이스를 상기 수신한 기준 반경 및 최소 이웃 개수를 적용하여 클러스터링하여 클러스터(cluster)를 생성하는 클러스터링 수행부를 포함한다.

[0013] 또한, 본 발명의 제 3 측면에 따른 데이터 마이닝 방법은, 원본 데이터 베이스의 속성에 기초하여 더미 데이터의 개수를 설정하는 단계; 상기 원본 데이터 베이스에 상기 설정된 개수의 더미 데이터를 삽입하여 변형 데이터 베이스를 생성하는 단계; 및 상기 변형 데이터 베이스를 설정된 기준 반경 및 최소 이웃 개수를 적용하여 밀도 기반 공간 클러스터링(Density Based Spatial Clustering of Applications with Noise)하는 단계를 포함한다.

효 과

[0014] 전술한 본 발명의 과제 해결 수단에 의하면, 원본 데이터 베이스에 소정의 더미 데이터를 추가 삽입하여 가공한 변형 데이터 베이스를 클러스터링(clustering)함으로써, 간단한 연산만으로도 원본 데이터 베이스에 대한 프라이버시를 보존하면서도 동일한 데이터 마이닝 결과를 얻을 수 있다는 장점이 있다.

[0015] 또한, 본 발명의 과제 해결 수단에 의하면, 원본 데이터 베이스의 속성에 기초하여 설정된 개수의 더미 데이터를 원본 데이터 베이스에 추가 삽입하여 데이터 마이닝함으로써, 데이터 마이닝 정확도 및 프라이버시 보존 능력을 향상시킬 수 있다.

발명의 실시를 위한 구체적인 내용

[0016] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 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.

[0017] 도 1은 본 발명의 일 실시예에 따른 데이터 마이닝 시스템의 구조도이다.

[0018] 그리고, 도 2는 본 발명의 일 실시예에 따른 데이터 베이스 변형 방법을 설명하기 위한 순서도이다.

[0019] 또한, 도 3은 본 발명의 일 실시예에 따른 프라이버시 레벨에 따른 데이터 마이닝 결과를 설명하기 위한 리스트이다.

[0020] 도 1에서와 같이 본 발명의 일 실시예에 따른 데이터 마이닝 시스템은 클러스터링 조건 수신부(110), 데이터 베이스 수집부(120), 클러스터링 수행부(130) 및 마이닝 데이터 제공부(140)를 포함하는 데이터 마이닝 장치(100)와, 데이터 베이스 가공부(210), 데이터 베이스 전송부(220) 및 마이닝 데이터 수신부(230)를 포함하는 사용자 장치(200)로 구성된다.

[0021] 데이터 마이닝 장치(100)의 클러스터링 조건 수신부(110)는 사용자가 입력하는 클러스터링 조건들을 수신하여 클러스터링 수행부(130)로 제공한다.

[0022] 클러스터링 조건 수신부(110)는 기준 반경(Eps) 값 및 상기 Eps 내에 존재하는 최소 데이터의 개수 즉, 최소 이웃 개수(MinPts) 등을 포함하는 클러스터링 조건을 수신한다.

[0023] 참고로, 본 발명의 실시예에 따른 데이터 마이닝 시스템에서는 데이터 베이스에 대해 밀도 기반 공간 클러스터링(DBSCAN: Density Based Spatial Clustering of Applications with Noise)을 수행하기 위해 클러스터링 대상 데이터 베이스의 전체 데이터들을 d차원 벡터로 정의하고, 각 벡터들을 점(point)으로 간주한다.

[0024] 이때, 클러스터링 대상 데이터 베이스에 포함된 각 점의 종류는 핵심점, 핵심점으로부터 Eps 내에 존재하는 변두리점 및 그 외의 노이즈점으로 분류된다.

[0025] 구체적으로, 하기 수학적 1을 이용하여 점 p를 중심으로 Eps 내에 존재하는 변두리점들의 집합을 확인할 수 있다.

[0026] [수학식 1]

$$N_{Eps}(P) = \{q \in D' | dist(p, q) \leq Eps\}$$

[0028] 이때, 점 p 및 q는 변형 데이터 베이스 D'의 원소이며, 두 점 p 및 q 간의 유클리디안 거리(Euclidean distance)인 dist(p, q)는 하기 수학식 2를 이용하여 산출할 수 있다.

[0029] [수학식 2]

$$|p - q| = \sqrt{\sum_{i=1}^d (p_i - q_i)^2}$$

[0031] 이때, 점 p가 하기 수학식 3을 만족할 때 Eps 내에 MinPts 이상의 다른 점이 존재한다고 판단하여 점 p의 종류를 핵심점으로 구분한다.

[0032] [수학식 3]

$$|N_{Eps}(P)| \geq MinPts$$

[0034] 한편, 상기 클러스터링 조건은 클러스터링 대상의 속성에 기초하여 기설정된 값으로서, 클러스터링 조건 수신부(110)는 데이터 마이닝 장치(100)의 운용자가 입력하는 클러스터링 조건을 수신하거나, 클러스터링 대상 데이터 베이스를 제공하는 사용자 장치(200)로부터 클러스터링 조건을 수신하는 것도 가능하다.

[0035] 본 발명의 실시예에서는 클러스터링 대상 데이터들에 대해 시뮬레이션을 통해 가장 적절한 클러스터링 조건(즉, Eps 및 MinPts)이 설정되는 것을 예로 나타내었다. 이때, 클러스터링 대상의 속성은 데이터 베이스에 포함되는 전체 데이터의 개수 및 밀도 등을 의미한다.

[0036] 구체적으로, 클러스터링 조건을 설정하기 위한 시뮬레이션에서는 데이터 베이스 내에서 하나의 점으로부터 m번 채로 가까운 점(이하, 'm'이라 함)까지의 거리를 m-dist라고 가정한다. 이때, 데이터 베이스의 모든 점들에 있어서 동일 클러스터에 포함되는 점들의 m-dist의 값은 작게 나타난다. 그리고, 노이즈점과 같이 동일 클러스터에 포함되지 않는 점들의 m-dist는 상대적으로 크게 나타난다. 따라서, 데이터 베이스의 모든 점들에 대해 계산된 m-dist를 오름차순으로 정렬하여 연결 그래프로 나타내면, 적절한 Eps의 값에 대응하는 m-dist에서 가파른 변화가 기대된다. 즉, 해당 m-dist이 가파른 변화가 나타난 거리를 Eps의 값으로 설정하고, 그에 해당되는 m의 값 즉, m개를 MinPts 값으로 설정할 수 있다.

[0037] 데이터 베이스 수집부(120)는 외부(본 발명의 실시예에서는 '사용자 장치')로부터 복수의 데이터들을 포함하는 데이터 베이스를 수신하여 저장한다. 이때, 데이터 베이스 수집부(120)는 가공된 원본 데이터 베이스(이하, '변형 데이터 베이스'라고 함)를 수신하며, 상기 변형 데이터 베이스를 클러스터링 수행부(130)에 제공한다. 본 발명의 실시예에서, 변형 데이터 베이스는 원본 데이터 베이스에 더미 데이터(dummy data)가 추가 삽입된 데이터 베이스를 의미한다.

[0038] 클러스터링 수행부(130)는 클러스터링 조건 수신부(110)로부터 제공받은 클러스터링 조건에 따라 데이터 베이스 수집부(120)로부터 제공받은 데이터 베이스를 클러스터링한다. 그리고, 클러스터링 수행부(130)는 생성된 클러스터들을 마이닝 데이터 제공부(140)에 전달한다.

[0039] 본 발명의 실시예에 따른 클러스터링 수행부(130)는 하기 알고리즘 1을 수행하여 변형 데이터 베이스 D'에 대한 클러스터(cluster)들을 생성한다.

[0040] [알고리즘 1]

```

ClustID = label of first cluster;
for each point p in set D'
    if(p.ID = UNCLASSIFIED) then
        if ExpandClust(D', p, ClustID, Eps, Min) then
            ClustID = ClustID +1 (Next cluster name)
        endif
    endif
endfor

```

[0041]

[0042] 이때, 알고리즘 1에 포함되는 ExpandClust(D, p, ClustID, Eps, Min)함수는 하기 함수 1과 같이 표현된다.

[0043] [함수 1]

```

Set =  $N_{Eps}(p)$ 
if |Set| < Min then
    p.ID = NOISE
    return 0
else
    for each point q in Set do
        q.ID = ClustID
    end for
    delete p from Set
    while |Set| > 0 do
        currentP = first point in Set
        currentS =  $N_{Eps}(p)$ 
        if |Set| >= Min then
            for each point q in currentS do
                if q.ID = UNCLASSIFIED then
                    q.ID = ClustID
                    add q to Set
                elseif q.ID = NOISE then
                    q.ID = ClustID
                endif
            endfor
        endif
        delete currentP from Set
    endwhile
    return 1
endif

```

[0044]

[0045] 이때, 클러스터링 수행부(130)는 알고리즘 1에 변형 데이터 베이스 D'에 포함된 전체 점(point)들과 기설정된 Eps 값 및 최소 이웃 개수 Min을 입력한다. 참고로, 알고리즘 1에 의해 D'의 모든 점이 3가지 종류의 점으로 분류된다.

[0046] 구체적으로, 알고리즘 1에서는 각각의 점을 검사하여 검사한 점이 분류되지 않은 점이면, 이를 ExpandClust 함수에 넘겨준다. 그리고, ExpandClust 함수는 넘겨받은 점이 핵심점인지 여부를 판단한다. 만일, 상기 넘겨받은 점이 핵심점이 아니라면 임시적으로 노이즈점으로 처리하여 실패 처리를 수행한다. 반면, 핵심점일 경우 그 점의 N_{Eps} 에 속하는 모든 점을 현재의 클러스터 이름(ClusterID)으로 지정하고, 이 점들을 하나의 집합(Set)으로 생성한다.

[0047] 다음으로, 상기 생성한 집합 안의 점이 다시 핵심점이 되는 경우 동일한 클러스터 이름으로 지정하고 그 점에 대한 N_{Eps} 을 다시 검사하여 N_{Eps} 에 속하는 점 중 지정되지 않은 점일 경우 집합에 포함시키고, 노이즈점으로 임시 지정된 점일 경우 동일 클러스터 이름으로 변경하고 현재 핵심점에 대한 변두리점으로 지정한다. 이러한 일련의 작업들로 하나의 핵심점으로부터 하나의 클러스터가 생성된다.

[0048] 그런 다음, ExpandClust 함수에서는 정상적으로 클러스터가 형성되면 성공 값인 1을 반환하고 클러스터 이름을

증가시켜 다음 클러스터에 대한 이름을 부여한다.

[0049] 이처럼, 다른 점들에서 또 다른 핵심점이 존재하면 이전의 작업과 같은 방식으로 또 하나의 클러스터를 생성한다. 결과적으로, 핵심점들이 복수의 클러스터의 각 씨앗이 되고 변두리점으로 경계를 채우게 된다.

[0050] 마이닝 데이터 제공부(140)는 상기 생성된 클러스터들을 상기 변형 데이터 베이스를 전송한 사용자 장치(200)로 전송한다. 이때, 마이닝 데이터 제공부(140)는 인터넷망 등의 네트워크를 통해 상기 생성된 클러스터들을 해당 사용자 장치(200)로 전송하며, 이를 전송하기 위한 규격 별 인터페이스를 포함하고 있다.

[0051] 한편, 사용자 장치(200)의 데이터 베이스 가공부(210)는 데이터 마이닝을 하고자 하는 원본 데이터 베이스를 가공하여 변형 데이터 베이스를 생성한다.

[0052] 구체적으로, 도 2에서 나타낸 바와 같은 순서에 따라 데이터 베이스 가공부(210)는 원본 데이터 베이스를 가공한다.

[0053] 먼저, 데이터 베이스 가공부(210)는 원본 데이터 베이스의 각 데이터들을 정규화한다(S210). 예를 들어, 본 발명의 실시예에 따른 데이터 베이스 가공부(210)는 하기 수학적 식 4와 같은 z-score 정규화를 통해 원본 데이터 베이스를 정규화할 수 있다.

[0054] [수학적 식 4]

$$v' = \frac{v - \mu}{\delta}$$

[0055]

[0056] 수학적 식 4에서 v' 는 정규화 결과 값, v 는 정규화 대상 데이터 값, μ 는 평균, δ 는 표준편차를 의미한다. 이때, z-score 정규화를 통해서 평균은 0이 되고 표준편차는 1이 되는 정규화가 수행된다. 이와 같이, 원본 데이터 베이스를 정규화함으로써 데이터들의 속성을 통일할 수 있다. 본 발명의 실시예에서는, z-score 정규화 방식 외에도 다양한 정규화 방법이 적용될 수 있다.

[0057] 그런 다음, 데이터 베이스 가공부(210)는 설정된 프라이버시 레벨에 기초하여 상기 원본 데이터 베이스에 삽입할 더미 데이터(dummy data)의 개수를 결정한다(S220).

[0058] 참고로, 프라이버시 레벨은 원본 데이터 베이스의 프라이버시를 보존할 수 있는 성능별 등급으로서, 일정 기준 개수까지 더미 데이터의 개수를 증가시킬수록 프라이버시를 보존할 수 있는 능력도 상승한다.

[0059] 이때, 더미 데이터가 상기 일정 기준 개수를 초과하면 프라이버시 보존 능력은 상승하나 데이터 마이닝 정확도는 하강하는 반비례 관계가 성립될 수 있다. 따라서, 프라이버시 레벨은 데이터 마이닝 시 적용되는 클러스터링 조건과 원본 데이터 베이스의 전체 데이터 개수 등의 속성을 참고하여 설정된다. 참고로, 데이터 마이닝 정확도는 원본 데이터 베이스와 변형 데이터 베이스에 대해 각각 클러스터링을 수행했을 때 생성된 클러스터의 개수가 동일하다는 것을 의미한다.

[0060] 이와 같은, 프라이버시 레벨은 원본 데이터 베이스의 속성에 기준하여 더미 데이터의 개수를 변경함으로써 구분될 수 있으며, 원본 데이터 베이스의 속성은 더미 데이터가 원본 데이터 베이스에 적용될 수 있는 실제 값의 범위 및 원본 데이터 베이스의 전체 데이터 개수 및 데이터들이 분포한 밀도 등을 의미한다. 예를 들어, 데이터 마이닝 대상 원본 데이터 베이스의 속성이 몸무게 정보일 경우 더미 데이터 또한 몸무게 정보를 나타내는 실제 값을 포함한다.

[0061] 도 3에서는 복수 개의 원본 데이터 베이스들(House, A1, A2, S1, S3)에 대해 복수의 프라이버시 레벨에 따른 더미 데이터들을 삽입한 후 데이터 마이닝한 결과 값을 예로서 나타내었다.

[0062] 이때, 도 3에서와 같이 프라이버시 보존 능력 및 데이터 마이닝 정확도는 프라이버시 레벨 2일 때 즉, 원본 데이터 베이스의 전체 데이터 개수의 50%의 더미 데이터를 각 원본 데이터 베이스에 삽입했을 때 가장 양호한 것을 확인할 수 있다.

[0063] 또한, 도 3에서는 프라이버시 레벨 3 또는 레벨 4일 때 즉, 원본 데이터 베이스의 전체 데이터 개수의 100% 또는 200%의 더미 데이터를 각 원본 데이터 베이스에 삽입했을 때 데이터 베이스 House, A1 및 S1에서는 프라이버시 보존 능력 및 데이터 마이닝 정확도가 상승되었으나, 나머지 데이터 베이스들에서는 데이터 마이닝 정확도 및 프라이버시 보존 능력이 불량한 것을 확인할 수 있다. 이때, 전체 데이터 베이스의 프라이버시 레벨은 시물

레이션 결과인 프라이버시 레벨 2로 설정될 수 있다.

[0064] 본 발명의 실시예에서는 사용자 장치(200)의 데이터 베이스 가공부(210) 자체에서 사용자의 입력에 따라 프라이버시 레벨을 설정하는 것을 나타내었다. 그런데, 데이터 마이닝 장치(100)가 원본 데이터 베이스의 속성에 기준하여 더미 데이터의 개수를 변화시키는 시뮬레이션을 통해 적절한 프라이버시 레벨을 설정하여 사용자 장치(200)로 전송하는 것도 가능하다.

[0065] 다시 도 2를 참조하면, 단계 S220 수행 후 데이터 베이스 가공부(210)는 설정된 개수의 더미 데이터를 원본 데이터 베이스에 추가 삽입한다(S230).

[0066] 예를 들어, 데이터 베이스 가공부(210)는 프라이버시 레벨 2로 설정된 경우 원본 데이터 베이스의 데이터 개수의 50%의 더미 데이터가 추가 삽입하여 변형 데이터 베이스를 생성한다.

[0067] 이때, 데이터 베이스 가공부(210)는 하기 알고리즘 2를 이용하여 변형 데이터 베이스를 생성한다.

[0068] [알고리즘 2]

INPUT : D(원본 데이터베이스), k(삽입될 데이터의 개수)

OUTPUT : D'(변형된 데이터베이스)

Begin

For i = 1 to k

$r_i \leftarrow \text{unni} \text{formly random}(D \text{의 범위})$

$D' = D \cup r_i$;

End for

End

[0069]

[0070] 데이터 베이스 가공부(210)는 상기 알고리즘 2에 원본 데이터 베이스(D)와 삽입하고자 하는 더미 데이터의 개수(k)를 입력한다. 그리고, D의 범위에서 원하는 개수만큼의 더미 데이터를 균등하게 임의로 선택하여 원본 데이터 베이스에 추가 삽입한다. 이때, D의 범위는 각각의 데이터들이 속할 수 있는 범위 즉, 속성을 의미한다. 예를 들어, 체중이라는 속성의 값일 경우 더미 데이터들의 값은 0에서 200 사이의 값이 데이터의 범위로 설정될 수 있다.

[0071] 상기 알고리즘 2에서 더미 데이터들은 균등 랜덤하게 선택되므로 D의 밀도는 전체적으로 상승하지만, 밀도의 차이에는 변화는 생기지 않는다. 따라서, D의 모든 부분이 평균적으로 동일하게 밀도가 변하게 되기 때문에 밀도 차이의 기반으로 실행되는 클러스터링에 영향을 주지 않는다.

[0072] 예를 들어, 도 7에서와 같이 본 발명의 실시예에 따른 데이터 마이닝 방법에서 더미 데이터의 개수를 변경하여 클러스터링을 수행하여도 핵심점과 변두리점으로 구성되는 각 클러스터의 개수가 동일한 결과를 확인할 수 있다.

[0073] 도 7은 본 발명의 실시예에 따른 데이터 마이닝의 결과를 설명하기 위한 도면이다.

[0074] 이때, 도 7의 (a)는 원본 데이터 베이스를 클러스터링한 결과를 나타내고, (b) 내지 (d)는 각각 원본 데이터 베이스의 데이터 개수의 30%, 50%, 100% 개수의 더미 데이터를 삽입하여 생성한 변형 데이터 베이스를 클러스터링한 결과를 나타낸다. 도 7의 (a)에 비교하여 같은 클러스터에 포함된 점들은 (b) 내지 (d)에서도 같은 클러스터에 속한다는 것을 확인할 수 있다.

[0075] 다시 도 2를 참조하면, 단계 S230 수행 후 데이터 베이스 가공부(210)는 상기 변형 데이터 베이스를 데이터 베이스 전송부(230)로 전송한다(S240). 이때, 데이터 베이스 가공부(210)는 사용자 장치(200)의 자체 데이터 베이스에 저장된 원본 데이터 베이스를 가공하고, 상기 원본 데이터 베이스에 상응하는 변형 데이터 베이스를 상기 자체 데이터 베이스에 매칭하여 저장할 수 있다.

[0076] 한편, 데이터 베이스 전송부(220)는 수신되는 변형 데이터 베이스를 네트워크를 통해 데이터 마이닝 장치(100)로 전송하여 데이터 마이닝 요청을 한다.

[0077] 그리고, 마이닝 데이터 수신부(230)는 네트워크를 통해 데이터 마이닝 장치(100)로부터 마이닝 데이터 즉, 클러스터들로 구성된 데이터를 수신한다. 이때, 마이닝 데이터 수신부(230)는 상기 수신된 마이닝 데이터를 상기 원본 데이터 베이스에 매칭하여 저장할 수 있다.

- [0078] 도 4는 본 발명의 실시예에 따른 클러스터링 방법을 설명하기 위한 순서도이다.
- [0079] 먼저, 클러스터링 조건을 수신한다(S410).
- [0080] 이때, 수신되는 클러스터링 조건은 클러스터링 대상 데이터 베이스의 각 점(point)을 밀도 기반으로 클러스터링 하기 위한 Eps 및 MinPts를 포함한다. 참고로, 클러스터링 조건은 클러스터링 대상 데이터 베이스를 제공하는 사용자 장치로부터 수신되거나, 데이터 마이닝 장치의 운용자로부터 입력될 수 있다.
- [0081] 상기 클러스터링 조건이 설정된 상태에서, 더미 데이터가 삽입된 변형 데이터 베이스를 수신한다(S420).
- [0082] 이때, 상기 더미 데이터의 개수는 클러스터링 대상 데이터 베이스의 원본 데이터 베이스의 속성에 기초하여 기 설정된 프라이버시 레벨에 따라 결정된다.
- [0083] 다음으로, 상기 변형 데이터 베이스의 각 데이터의 종류를 분류한다(S430).
- [0084] 이때, 본 발명의 실시예에서 상기 각 데이터는 벡터 점으로 정의되며, 각 점은 핵심점, 상기 핵심점 주변의 Eps 내에 존재하는 변두리점, 및 상기 Eps 외부에 존재하는 노이즈점으로 분류된다.
- [0085] 그런 후, 상기 클러스터링 대상 데이터 베이스의 전체 데이터에 대해 상기 각 데이터의 종류를 분류하는 일련의 과정을 수행하여 적어도 하나의 클러스터를 생성한다(S440). 이와 같이, 클러스터링 대상 데이터 베이스의 각 점에 대해 핵심점, 변두리점, 노이즈점을 분류하고 클러스터를 생성하는 과정을 클러스터링이라 한다.
- [0086] 다음으로, 상기 생성된 클러스터 즉, 마이닝 데이터를 데이터 마이닝을 요청 측으로 전송한다(S450). 이때, 데이터 마이닝 요청 측이 외부일 경우 네트워크를 통해 상기 마이닝 데이터를 전송할 수 있으며, 데이터 마이닝 요청 측과 변형 데이터 베이스 제공 측이 동일할 경우 상기 마이닝 데이터를 저장한 후, 사용자의 요청에 따라 출력하는 것도 가능하다.
- [0087] 한편, 본 발명의 실시예에 따른 데이터 마이닝 시스템은 다자간 환경에서도 적용 가능하다. 이하, 도 5를 참조하여 다자간 환경에서의 데이터 마이닝 시스템을 설명하도록 한다.
- [0088] 도 5는 본 발명의 다른 실시예에 따른 다자간 환경에서의 데이터 마이닝 시스템을 나타내는 도면이다.
- [0089] 참고로, 도 5에서 나타낸 본 발명의 다른 실시예에 따른 데이터 마이닝 시스템의 구성 및 동작에 대해서 설명하기에 앞서 하기 알고리즘 3을 통해 다자간 환경에서 데이터 마이닝을 수행하는 방법에 대해서 계략적으로 설명하도록 한다.
- [0090] [알고리즘 3]

INPUT : $D_1, D_2, \dots, D_n$

OUTPUT : $DBSCAN(D_1 \cup D_2 \dots \cup D_n, Eps, Min)$

Begin

n명의 참가자에게 정규화 규격 및 원본 데이터 개수에 비례하는 더미 데이터 개수 한계 비율 k(%)를 전송

For i=1 to n

원본 데이터 베이스 개수의 k% 이하인 개수 k'을 선택

$D_i' =$ 더미 데이터 삽입 알고리즘(D_i, k')

D_i 전송

End for

$D_1' \cup D_2' \dots \cup D_n'$ 에 대해 DBSCAN을 실행하여 결과 값을 n명의 참가자에게 전송

End

- [0091]
- [0092] 상기 알고리즘 3은 n명의 데이터 베이스를 가진 참여자(즉, 복수의 사용자 장치)와 실제 데이터 마이닝을 실시하는 마이너(즉, 데이터 마이닝 장치)를 통해 수행되며, 기본적으로 서로 신뢰할 수 없는 상태라고 가정한다.
- [0093] 구체적으로, 알고리즘 3에 따르면, 마이너가 전송하는 원본 데이터 베이스(D)의 가공 조건 정보에 따라 각각의 참여자가 더미 데이터 삽입 알고리즘을 수행하여 변형 데이터 베이스(D')를 생성하고, 생성된 변형 데이터 베이스를 마이너로 전송한다. 그리고, 각 참여자로부터 변형 데이터 베이스를 수신한 마이너는 각 변형 데이터 베이스에 대한 합집합을 생성하여 클러스터링을 수행한다. 그리고, 마이너는 클러스터링 결과 생성된 클러스터들을 n명의 참가자에게 동일하게 전송한다. 참고로, 상기 더미 데이터 삽입 알고리즘은 상기 알고리즘 2를 이용

할 수 있다.

- [0094] 이때, 알고리즘 3에 따르면, 마이너는 복수의 변형 데이터 베이스들의 합집합하기 위한 데이터 정규화 규격을 설정하고, 더미 데이터의 개수에 따른 정확도를 고려하여 더미 데이터 개수 한계 비율을 설정한다. 참고로, 상기 가공 조건 정보는 상기 정규화 규격 및 더미 데이터 개수 한계 비율을 포함하며, 마이너는 상기 가공 조건 정보를 각 참여자들에게 전송한다.
- [0095] 이와 같은, 알고리즘 3을 이용함으로써 참가자 n명 사이에는 데이터 송수신이 발생하지 않으며, 마이너 또한 변형 데이터 베이스를 획득하여 데이터 마이닝함으로써 다자간 환경에서도 프라이버시를 보존할 수 있다.
- [0096] 다시 도 5를 참조하면, 본 발명의 다른 실시예에 따른 데이터 마이닝 시스템은 다자간 환경에서 데이터 마이닝 장치(100)와 복수의 사용자 장치(200)가 네트워크를 통해 연결된다.
- [0097] 참고로, 다자간 환경이란 데이터 마이닝을 요청하는 사용자 장치(200)가 복수 개로써, 각 사용자 장치(200)가 데이터 마이닝 장치(100)에 각각 변형 데이터 베이스를 제공하고, 데이터 마이닝 장치(100)가 전체 변형 데이터 베이스에 대해 클러스터링한 결과를 복수의 사용자 장치(200)에 동일하게 전송하는 환경을 의미한다.
- [0098] 본 발명의 다른 실시예에서는, 하나의 데이터 마이닝 장치(100)와 복수의 사용자 장치(200)가 연결된 것을 나타내었으나, 임의의 사용자 장치(200)에 데이터 마이닝 장치(100)가 자체적으로 포함되고, 상기 임의의 사용자 장치(200)가 데이터 마이닝을 수행한 결과를 나머지 사용자 장치(200)로 전송하는 것도 가능하다.
- [0099] 한편, 도 5에서 나타난 본 발명의 다른 실시예에 따른 데이터 마이닝 장치(100)와 사용자 장치(200)의 구성은 도 1에서 설명한 본 발명의 일 실시예에 따른 데이터 마이닝 장치 및 사용자 장치의 각 구성과 동일하다.
- [0100] 다만, 본 발명이 다른 실시예에 따른 데이터 마이닝 장치(100)는 복수의 사용자 장치(200)에서 각각 원본 데이터 베이스를 변형 데이터 베이스로 가공하는데 필요한 가공 조건 정보를 생성하여 각 사용자 장치(200)로 전송하는 가공 조건 전송부(미도시)를 더 포함한다. 이때, 상기 가공 조건 정보는 복수의 사용자 장치(200)에서 더미 데이터의 개수를 결정하는데 필요한 더미 데이터 개수 한계 비율을 의미한다.
- [0101] 본 발명의 다른 실시예에서는, 데이터 마이닝 장치(100)가 상기 가공 조건 정보를 생성하여 각 사용자 장치(200)로 전송하는 것을 나타내었다. 그런데, 각 사용자 장치(200) 자체적으로 사용자가 동일한 가공 조건을 입력할 수 있으며, 각 사용자 장치(200)의 데이터 베이스 가공부(210)에 상기 가공 조건 정보가 기설정되어 있는 것도 가능하다.
- [0102] 본 발명의 다른 실시예에 따른 데이터 마이닝 장치(100)의 데이터 베이스 수집부(130)는 복수의 사용자 장치(200)로부터 각각 수신되는 변형 데이터 베이스를 수신하여 저장한다.
- [0103] 그리고, 본 발명의 다른 실시예에 따른 데이터 마이닝 장치(100)의 클러스터링 수행부(130)는 각 사용자 장치(200)로부터 수신된 복수의 변형 데이터 베이스들의 합집합에 대해 클러스터링을 수행하여 클러스터를 생성한다.
- [0104] 또한, 본 발명의 다른 실시예에 따른 데이터 마이닝 장치(100)의 마이닝 데이터 제공부(140)는 상기 생성된 클러스터를 복수의 사용자 장치(200)에 각각 전송한다.
- [0105] 한편, 본 발명의 다른 실시예에 따른 사용자 장치(200)의 데이터 베이스 가공부(210)는 수신하거나 기설정된 가공 조건 정보에 따라 원본 데이터 베이스의 전체 데이터 개수 중 더미 데이터 개수 한계 비율만큼의 개수를 더미 데이터의 개수로 설정한다.
- [0106] 도 6은 본 발명의 다른 실시예에 따른 데이터 마이닝 방법을 설명하기 위한 순서도이다.
- [0107] 도 6에서는 본 발명의 다른 실시예에 따른 다자간 환경에서의 데이터 마이닝 방법을 나타내었다.
- [0108] 먼저, 기설정된 프라이버시 레벨에 기준하여 원본 데이터 베이스에 삽입할 더미 데이터 개수를 설정한다(S610).
- [0109] 이때, 상기 프라이버시 레벨은 원본 데이터 베이스의 속성에 기초하여 설정될 수 있으며, 다자간 환경에서는 기설정된 더미 데이터 개수 한계 비율과 동일하게 설정될 수 있다. 참고로, 더미 데이터 개수 한계 비율은 복수의 사용자 장치의 각 원본 데이터들을 사전에 데이터 마이닝 시뮬레이션을 하여 원본 데이터 베이스들의 전체 데이터 개수에 따라 프라이버시 보존 능력 및 데이터 마이닝 정확도가 가장 양호하도록 하는 비율이 설정될 수 있다.
- [0110] 다음으로, 원본 데이터 베이스에 상기 설정된 개수의 더미 데이터를 추가 삽입하여 변형 데이터 베이스를 생성

한다(S620).

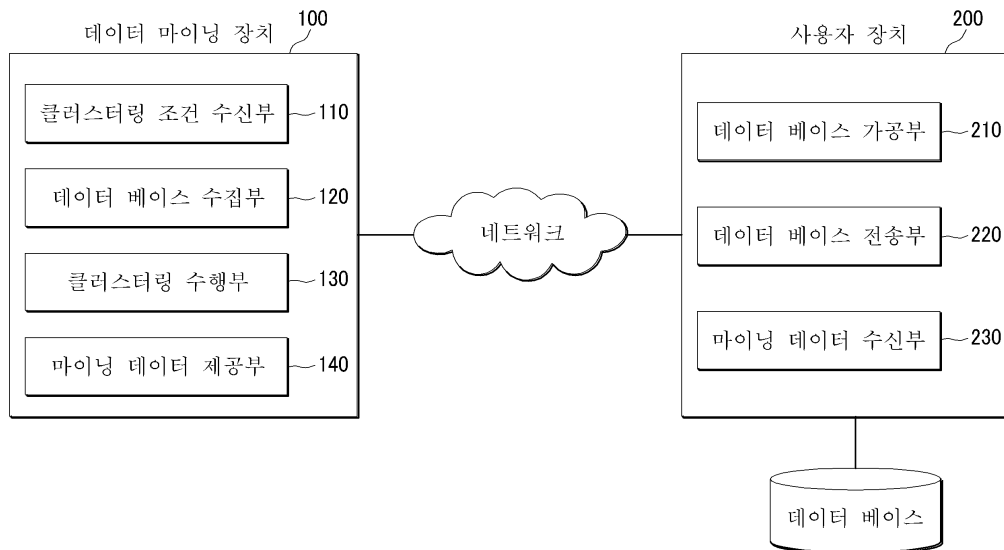
- [0111] 이때, 다자간 환경에서는 상기 설정된 더미 데이터 개수 한계 비율을 원본 데이터 베이스를 제공하는 복수의 사용자 장치에 각각 제공하고, 각 사용자 장치에서 원본 데이터 베이스의 데이터 개수 중 상기 더미 데이터 한계 비율만큼의 개수로 설정된 더미 데이터를 삽입한다.
- [0112] 그런 후, 상기 생성된 변형 데이터 베이스에 대한 클러스터링을 수행한다(S630).
- [0113] 구체적으로, 복수의 사용자 장치로부터 각각의 변형 데이터 베이스를 수집하고, 상기 수집된 전체 변형 데이터 베이스를 기설정된 클러스터링 조건 즉, Eps 및 MinPts를 적용하여 클러스터링한다.
- [0114] 그런 다음, 상기 클러스터링의 결과로 생성된 클러스터 즉, 마이닝 데이터를 각 사용자 장치로 제공한다(S640).
- [0115] 이상에서, 본 발명의 장치 및 방법은 특정 실시예와 관련하여 설명되었지만, 그것들의 구성 요소 또는 동작의 일부 또는 전부는 범용 하드웨어 아키텍처를 갖는 컴퓨터 시스템을 사용하여 구현될 수 있다.
- [0116] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.
- [0117] 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면의 간단한 설명

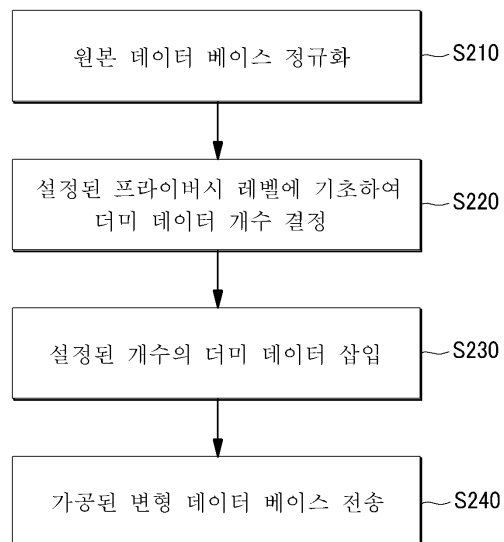
- [0118] 도 1은 본 발명의 일 실시예에 따른 데이터 마이닝 시스템의 구조도이다.
- [0119] 도 2는 본 발명의 일 실시예에 따른 데이터 베이스 변형 방법을 설명하기 위한 순서도이다.
- [0120] 도 3은 본 발명의 일 실시예에 따른 프라이버시 레벨에 따른 데이터 마이닝 결과를 설명하기 위한 리스트이다.
- [0121] 도 4는 본 발명의 실시예에 따른 클러스터링 방법을 설명하기 위한 순서도이다.
- [0122] 도 5는 본 발명의 다른 실시예에 따른 다자간 환경에서의 데이터 마이닝 시스템을 나타내는 도면이다.
- [0123] 도 6은 본 발명의 다른 실시예에 따른 데이터 마이닝 방법을 설명하기 위한 순서도이다.
- [0124] 도 7은 본 발명의 실시예에 따른 데이터 마이닝의 결과를 설명하기 위한 도면이다.

도면

도면1



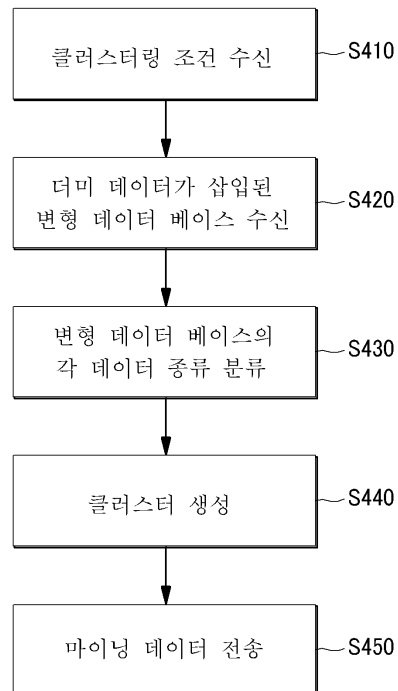
도면2



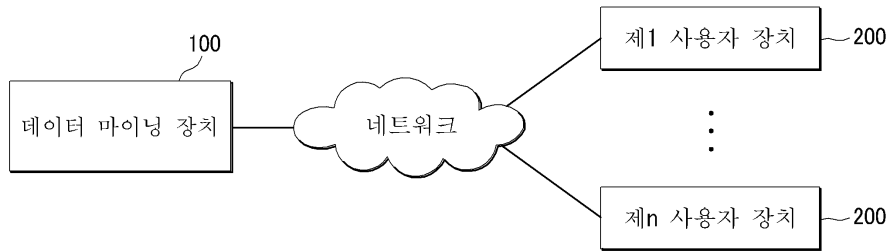
도면3

데이터베이스	House	A1	A2	S1	S3
데이터 개수	1857	3000	5250	5000	5000
차원	3	2	2	2	2
군집 개수	1	20	31	15	15
프라이버시 레벨 1	전체 데이터 개수의 30%만큼 더미 데이터 삽입				
정확도	O	O	O	O	O
프라이버시	0.08	0.02	0.07	0.09	0.05
프라이버시 레벨 2	전체 데이터 개수의 50%만큼 더미 데이터 삽입				
정확도	O	O	O	O	O
프라이버시	0.12	0.05	0.13	0.18	0.12
프라이버시 레벨 3	전체 데이터 개수의 100%만큼 더미 데이터 삽입				
정확도	O	O	X	O	X
프라이버시	0.23	0.13	-	0.34	-
프라이버시 레벨 4	전체 데이터 개수의 200%만큼 더미 데이터 삽입				
정확도	O	X	-	X	-
프라이버시	0.31	-	-	-	-

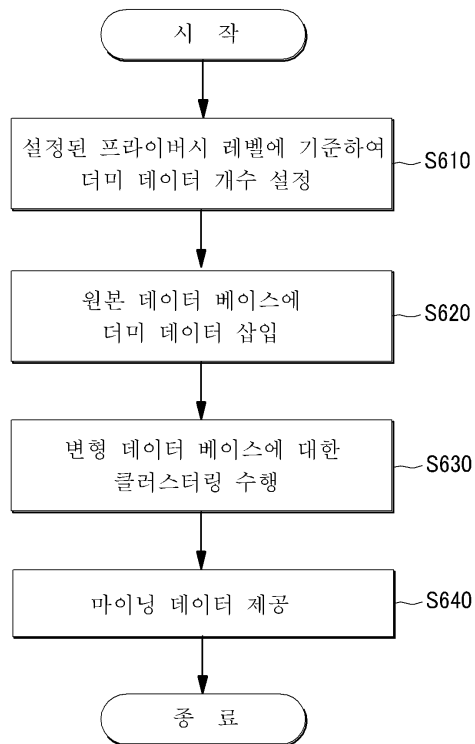
도면4



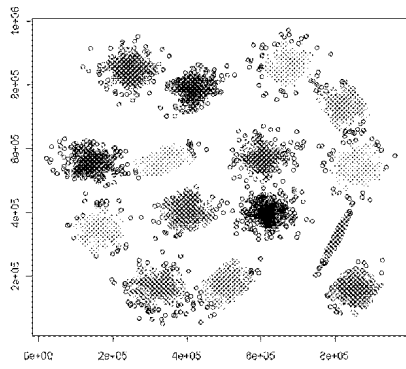
도면5



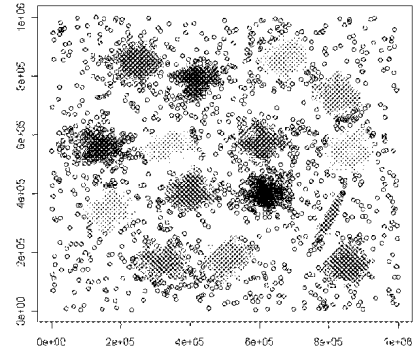
도면6



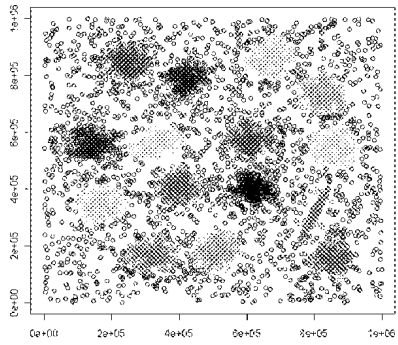
도면7



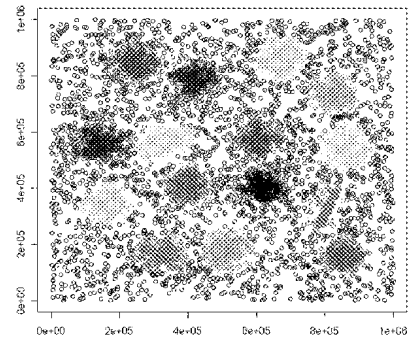
(a)



(b)



(c)



(d)