(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2014년03월27일
(11) 등록번호 10-1377260
(24) 등록일자 2014년03월17일

(51) 국제특허분류(Int. Cl.)
G06F 17/16 (2006.01)
(21) 출원번호 10-2012-0116945
(22) 출원일자 2012년10월19일
심사청구일자 2012년10월19일
(56) 선행기술조사문헌
Deflation Methods for Sparse PCA, Lester Mackey, Neural Information Processing Systems (NIPS'08), 2008.*
Power Iteration Clustering, Frank Lin et al. Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010.*
JP2009093655 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732, 국제캠퍼스 내 (서천동, 경희대학교)
(72) 발명자
이승룡
경기 성남시 분당구 구미로144번길 25, 102동 102호 (구미동, 삼환빌라)
팜 더 안
경기 용인시 기흥구 덕영대로 1732, 전자정보대학 351호 (서천동, 경희대학교)
(74) 대리인
특허법인 신지

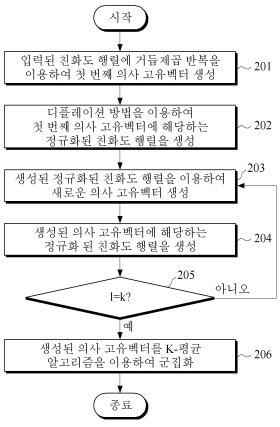
전체 청구항 수 : 총 4 항 심사관 : 양찬호

(54) 발명의 명칭 **디플레이션 기반의 거듭제곱 반복 군집화 방법**

(57) 요 약

본 발명은 디플레이션 방법을 거듭제곱 반복 군집화 방법에 적용하여 거듭제곱 반복 군집화의 빠른 연산을 유지하면서, 스펙트럼 군집화 방법과 유사한 수준의 정확도를 나타낼 수 있는 디플레이션 기반의 거듭제곱 반복 군집화 방법이다. 먼저, 입력된 정규화된 친화도 행렬에 거듭제곱 반복(Power Iteration)을 적용하여 의사 고유벡터(Pseudo Eigenvector)를 생성한다. 그리고 생성된 의사 고유벡터에 디플레이션 방법(Deflation Method)을 적용하여 새로운 정규화된 친화도 행렬을 생성한다. 다음으로 새로운 정규화된 친화도 행렬에 거듭제곱 반복을 적용하여 새로운 의사 고유벡터를 생성하고, 새로운 의사 고유벡터에 디플레이션 방법을 적용하여 또 다른 새로운 정규화된 친화도 행렬을 생성한다. 그리고 생성된 둘 이상의 의사 고유벡터를 K-평균 알고리즘(K-means Algorithm)을 이용하여 군집화한다.

대 표 도 - 도2



이 발명을 지원한 국가연구개발사업

과제고유번호	1415123336
부처명	지식경제부
연구사업명	대학 IT 연구센터 육성지원
연구과제명	동서신의학 u-라이프케어연구센터
기 여 율	1/1
주관기관	경희대학교 산학협력단
연구기간	2012.01.01 ~ 2012.12.31

특허청구의 범위

청구항 1

입력된 정규화된 친화도 행렬에 거듭제곱 반복(Power Iteration)을 적용하여 의사 고유벡터(Pseudo Eigenvector)를 생성하는 단계;

상기 생성된 의사 고유벡터에 디플레이션 방법(Deflation Method)을 적용하여 새로운 정규화된 친화도 행렬을 생성하는 단계;

상기 새로운 정규화된 친화도 행렬에 거듭제곱 반복을 적용하여 새로운 의사 고유벡터를 생성하고, 상기 새로운 의사 고유벡터에 디플레이션 방법을 적용하여 또 다른 새로운 정규화된 친화도 행렬을 생성하는 단계; 및

상기 생성된 둘 이상의 의사 고유벡터를 K-평균 알고리즘(K-means Algorithm)을 이용하여 군집화하는 단계;

를 포함하며,

상기 둘 이상의 유사 고유벡터는 상기 입력된 정규화된 친화도 행렬의 가장 큰 둘 이상의 고유벡터의 선형 결합인 것을 특징으로 하는 디플레이션 기반의 거듭제곱 반복 군집화 방법.

청구항 2

제 1항에 있어서,

상기 새로운 정규화된 친화도 행렬은,

$$W_l = W_{l-1} - \frac{W_{l-1} v_l v_l^T W_{l-1}}{v_l^T W_l v_l}$$

디플레이션 방법을 이용한 에 의해 산출되며,

상기 W_l 는 l번째 반복의 친화도 행렬이고, W_{l-1} 은 l-1번째 반복의 친화도 행렬이고, v_l 은 l번째 반복의 반복행렬이고, l은 반복 횟수인 것을 특징으로 하는 디플레이션 기반의 거듭제곱 반복 군집화 방법.

청구항 3

삭제

청구항 4

제 1항에 있어서,

상기 둘 이상의 유사 고유벡터는 상호 직교(Orthoginal)하는 것을 특징으로 하는 디플레이션 기반의 거듭제곱 반복 군집화 방법.

청구항 5

제 1항에 있어서,

상기 새로운 정규화된 친화도 행렬에 거듭제곱 반복을 적용하여 새로운 의사 고유벡터를 생성하고, 상기 새로운 의사 고유벡터에 디플레이션 방법을 적용하여 또 다른 새로운 정규화된 친화도 행렬을 생성하는 단계는 미리 설정된 임계값(Threshold)에 해당할 때까지 반복하는 것을 특징으로 하는 디플레이션 기반의 거듭제곱 반복 군집화 방법.

명 세 서

기 술 분 야

[0001] 본 발명은 데이터 군집화 장치 및 그 방법에 관한 것으로, 보다 상세하게는 문서, 필기체 숫자 및 얼굴 등에 대한 데이터세트를 위한 군집화 장치 및 그 동작 방법에 관한 것이다.

배 경 기 술

[0002] 스펙트럼 군집화(Spectral Clustering) 기법은 종래에 사용되던 K-평균(K-means) 알고리즘이나 대표값을 이용한 군집화(Clustering Using REpresentatives, CURE) 알고리즘보다 탁월한 장점들이 있기 때문에, 현재 가장 인기 있는 군집화(Clustering) 기법들 중에 하나이다. 그러나 스펙트럼 군집화는 복잡한 고유벡터(Eigenvector)를 계산해야 하기 때문에, 많은 연산량을 필요로 한다. 따라서 결과를 도출하는데 많은 시간과 노력을 기울여야 한다.

[0003] 이러한 스펙트럼 군집화의 문제점을 극복하기 위하여 스펙트럼 군집화를 보다 단순화하고 빠르게 계산할 수 있는 거듭제곱 반복 군집화(Power Iteration Clustering, PIC)가 제안되었다. 거듭제곱 반복 군집화는 스펙트럼 군집화와 달리 고유벡터를 계산하지 않고, 고유벡터의 선형 결합인 유일한 하나의 의사 고유벡터(Pseudo Eigenvector)만을 계산한다. 따라서 거듭제곱 반복 군집화는 스펙트럼 군집화보다 더 적은 연산으로 결과를 도출할 수 있다.

[0004] 하지만 거듭제곱 반복 군집화는 하나의 의사 고유벡터만을 사용하기 때문에, 정확성에 문제가 생길 수 있으며, 어떤 위중한 상황에서 하나의 의사 고유벡터만을 사용하는 것은 클래스 간 충돌(Inter-class Collision) 문제가 발생할 수 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명이 해결하고자 하는 과제는 디플레이션 방법을 거듭제곱 반복 군집화 방법에 적용하여 거듭제곱 반복 군집화의 빠른 연산을 유지하면서, 스펙트럼 군집화 방법과 유사한 수준의 정확도를 나타낼 수 있는 디플레이션 기반의 거듭제곱 반복 군집화 방법을 제공한다.

과제의 해결 수단

[0006] 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 먼저, 입력된 정규화된 친화도 행렬에 거듭제곱 반복(Power Iteration)을 적용하여 의사 고유벡터(Pseudo Eigenvector)를 생성한다. 그리고 생성된 의사 고유벡터에 디플레이션 방법(Deflation Method)을 적용하여 새로운 정규화된 친화도 행렬을 생성한다. 다음으로 새로운 정규화된 친화도 행렬에 거듭제곱 반복을 적용하여 새로운 의사 고유벡터를 생성하고, 새로운 의사 고유벡터에 디플레이션 방법을 적용하여 또 다른 새로운 정규화된 친화도 행렬을 생성한다. 그리고 생성된 둘 이상의 의사 고유벡터를 K-평균 알고리즘(K-means Algorithm)을 이용하여 군집화한다.

[0007] 둘 이상의 유사 고유벡터는 상기 입력된 정규화된 친화도 행렬의 가장 큰 둘 이상의 고유벡터의 선형 결합이다. 그리고 둘 이상의 유사 고유벡터는 상호 직교(Orthoginal)하는 것을 특징으로 한다. 새로운 의사 고유벡터를 생성하고, 상기 새로운 의사 고유벡터에 디플레이션 방법을 적용하여 또 다른 새로운 정규화된 친화도 행렬을 반복하는 단계는 미리 설정된 임계값(Threshold)에 해당할 때까지 반복한다.

발명의 효과

[0008] 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법을 통해 스펙트럼 군집화보다 더 빠른 연산이 가능하면서도 유사한 수준의 정확도를 나타낼 수 있기 때문에 다양한 문서, 얼굴 데이터 및 필기체 숫자 데이터 등을 군집화 하는데 효과적으로 적용될 수 있다.

도면의 간단한 설명

[0009] 도 1은 본 발명에 따른 거듭제곱 반복 알고리즘을 나타내는 도면이다.

도 2는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법을 나타내는 흐름도이다.

도 3은 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 알고리즘을 나타내는 도면이다.

도 4a는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법과 스펙트럼 군집화 및 거듭제곱 반복 군집화를 비교한 결과를 나타내는 도면이고, 도 4b는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화를 이용하여 뉴스 그룹 데이터세트를 군집화한 결과를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0010] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세하게 설명한다. 본 명세서에서 사용되는 용어는 실시예에서의 기능 및 효과를 고려하여 선택된 용어들로서, 그 용어의 의미는 사용자 또는 운용자의 의도 또는 업계의 관례 등에 따라 달라질 수 있다. 따라서 후술하는 실시예들에서 사용된 용어의 의미는, 본 명세서에 구체적으로 명시된 경우에는 명시된 정의에 따르며, 구체적으로 명시하지 않는 경우, 당업자들이 일반적으로 인식하는 의미로 해석되어야 할 것이다.

[0011] 도 1은 본 발명에 따른 거듭제곱 반복 알고리즘을 나타내는 도면이다.

[0012] 도 1을 참조하면, 벡터들의 세트 $X=\{x_i\}_{1\ldots n}$(각 x_i는 데이터세트에서의 데이터 포인트(Data Point)를 표현하는 공간 안에 존재함)를 가정한다. 그리고 x_i와 x_j 사이의 유사도(similarity)를 나타내는 유사도 함수(Similarity Function) $s(x_i,x_j)$를 정의한다. 그리고 친화도 행렬(Affinity Matrix) $A=\{a_{ij}\}_{ij=1\ldots n}$(각 $a_{ij}=s(x_i,x_j)$임)를 정의한다. 데이터 포인트 x_i를 표현하는 꼭짓점(Vertex) V_i를 가지는 무방향 그래프(Undirected Graph)와 같은 친화도 그래프(Affinity Graph)인 G=(V,B)를 정의한다. vi와 vj 사이의 변(Edge)의 가중치(weight)는 그들 사이의 유사도를 표현한다. 데이터 포인트 i와 가장 근접한 이웃 데이터 사이의 유사도만을 고려하면, 유사도 행렬과 유사도 그래프는 서로 연관된 요소이 비율이 극단적으로 작은 희소(Sparse)해진다. 반면에 i와 다른 데이터 사이의 유사도 설정은 0이다. 대각 행렬(Diagonal Matrix)은 수학식 1과 같이 정의된다.

수학식 1

[0013] $$D_{ij}=\Sigma A_{ij}$$

[0014] 그리고 정규화된 친화도 행렬(Normalized Affinity Matrix) W를 수학식 2와 같이 정의한다.

수학식 2

[0015] $$W=D^{-1}$$

[0016] 라플라스 행렬(Laplacian Matrix)은 L=I-W에 의해 계산된다(I는 단위 매트릭스(Unit Matrix)임).

[0017] 스펙트럼 군집화의 목표는 데이터를 데이터 포인트들의 군집(Cluster)들로 나누는 것이다. 각 그룹의 데이터 포인트는 유사한 특성(Property)을 갖는다. 스펙트럼 군집화에는 여러 종류가 있으나, 여기서는 한가지만을 고려한다. 우선, L의 k번째로 작은 고유벡터들을 찾고, W의 k번째로 큰 고유벡터들을 찾는다. 후술하는 k 고유벡터들은 W의 k번째 큰 고유벡터들을 의미한다. 이 후, 마지막 군집화(Clustering) 결과를 찾기 위해 이 고유벡터들에 K-평균 알고리즘을 적용한다.

[0018] 상술한 바와 같이, k번째로 큰 고유벡터를 계산하는 시간은 오랜 시간이 소모된다. 개별 k 고유벡터들을 계산하는 대신에, 거듭제곱 반복 군집화(PIC)는 개별 k 고유벡터들의 선형 결합인 하나의 유사 고유벡터를 찾는다. 유사 고유벡터를 계산하는 것은 k번째 고유벡터를 계산하는 것보다 시간과 연산량을 줄일 수 있다. W의 가잔 큰 고유벡터를 계산하는 방법인 거듭제곱 반복(Power Iteration)은 거듭제곱 반복 군집화의 주요 기술이다. 우선, 반복 벡터(Iteration Vector)는 랜덤 초기 벡터(Random Initialization Vector) v_0와 동일하게 설정된다. 각각

의 반복에서, 반복 벡터는 반복 벡터에서의 변화가 없을 때까지 반복 벡터와 W의 곱에 근거하여 갱신된다. 즉 수학식 3과 같이 정의된다.

수학식 3

[0019] $$v^{t+1} = W \circ v^{t}$$

[0020] 수학식 3에서 v_t는 반복 벡터이다.

[0021] 그러나 W는 일반화된 행렬이기 때문에, W의 가장 큰 고유벡터는 군집화 제안에 사용할 수 없는 상수 벡터이다. W의 가장 큰 고유 벡터에서의 위의 반복 과정이 두 단계(Phase)를 가질 것을 요구한다. 첫 번째 단계에서, 만약 두 개의 데이터 포인트들이 같은 군집 안에 있다면, 그것들의 대표값(Representation Value)들은 반복 벡터가 같다. 만약 두 개의 데이터 포인트들이 서로 다른 군집에 속해 있다면, 두 개의 데이터 포인트의 대표값은 서로 반복 벡터가 다르다. 그러므로, 이 반복 벡터는 군집화 제안에 유용하다. 두 번째 단계에서, 반복 벡터는 서서히 상수 벡터인 가장 큰 고유 벡터가 된다. 이러한 경우에, 반복 벡터는 유용하지 못하다.

[0022] 상술한 내용과 같이, 반복 벡터를 위한 랜덤 초기화 벡터 v_0를 생성한다. 그리고, 랜덤 초기화 벡터는 새로운 반복벡터를 생성하기 위하여, 현재의 친화도 행렬과 곱해진다. 이 후, 새로운 반복벡터와 현재의 친화도 행렬을 곱하는 과정을 반복적으로 수행한다. 다음으로 반복벡터가 너무 커지는 것을 방지하기 위하여 각 반복 과정에서 정규화 단계가 필요하다.

[0023] 다음으로 반복 과정 정지를 위한 지역 수렴 단계(Local Converge Phase) 확인을 위하여, 가속(Accelaration)을 사용한다. 가속은 수학식 4와 같이 정의된다.

수학식 4

[0024] $$\sigma^{t+1} = \left| v^{t+1} - v^{t} \right|$$

[0025] $$\left| \sigma^{t} - \sigma^{t-1} \right| < \varepsilon$$

[0026] 수학식 4에서 σ^{t} 는 t번째 반복의 가속도이고, ε 은 반복 횟수를 결정하기 위해 미리 설정된 임계값(Threshold)이다. 가속이 미리 정의된 임계값(Threshold)보다 작을 경우, 반복 과정은 정지된다.

[0027] 이러한 거듭제곱 반복은 스펙트럼 군집화를 위한 강력한 방법이지만, 멀티 클래스(Multi Class)를 갖는 데이터세트(Dataset)에는 적합하지 못하다. 이 데이터세트에서 거듭제곱 반복에 의해 생성된 의사 고유 벡터는 클래스간(Inter-class) 충돌 문제를 갖는다. 서로 다른 군집에 속해 있는 서로 다른 두 클래스는 동일한 값을 가지며, 상호 병합된다. 의사 고유벡터에 K-평균 알고리즘을 적용한 경우, K-평균 알고리즘이 잘못된 군집 결과를 찾는다면 그 결과는 고유(Original) 데이터세트와 달라진다.

[0028] 도 2는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법을 나타내는 흐름도이다.

[0029] 도 2를 참조하면, 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 종래의 거듭제곱 반복 군집화에서 유일한 의사 고유벡터를 계산하는 대신에, 다중 의사 고유벡터들을 계산한다. 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 W의 k 고유벡터들의 선형결합에 해당하는 다른 의사 고유벡터를

찾는다. 뿐만 아니라, 이 새로운 의사 고유벡터는 상호 직교한다. 따라서 의사 고유벡터를 중복하여 계산하는 과정을 피할 수 있다. 마지막으로, 새로운 의사 고유벡터를 계산하는 시간은 거듭제곱 반복의 의사 고유벡터를 계산하는 시간과 동일하다. 그래서 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 변화없는 스펙트럼 군집화의 종래 방법과 달리 거듭제곱 반복의 장점을 유지한다.

[0030] 먼저, 입력된 친화도 행렬에 거듭제곱 반복을 적용하여 첫 번째 의사 고유벡터를 생성한다(201). 상술한 도 1에서 설명한 거듭제곱 반복을 이용하여 첫 번째 의사 고유 벡터를 생성한다. 다음으로 디플레이션 방법(Deflation Method)을 이용하여 첫 번째 의사 고유벡터에 해당하는 새로운 정규화된 친화도 행렬을 생성한다(202). 디플레이션 방법은 대칭행렬의 고유벡터를 구하는 한가지 방법으로서 거듭제곱 방법을 통해 계산된 대칭행렬의 고유치와 그에 대응하는 고유벡터를 이용하여 새로운 고유치와 고유벡터를 구할 수 있는 알고리즘이다. 이러한 디플레이션 방법을 이용하여 첫 번째 의사 고유 벡터에 해당하는 정규화된 친화도 행렬을 계산한다. 수학식 5는 다음과 같다.

수학식 5

$$W_l = W_{l-1} - \frac{W_{l-1} v_l v_l^T W_{l-1}}{v_l^T W_l v_l}$$

[0031]

[0032] 수학식 5에서, W_l 는 l번째 반복의 친화도 행렬이고, W_{l-1} 은 l-1번째 반복의 친화도 행렬이고, v_l 은 l번째 반복된 반복 벡터이고, l은 반복 횟수이고, k는 반복 횟수를 결정하는 임계값(Threshold)이다.

[0033] 다음으로 새롭게 계산된 정규화된 친화도 행렬을 이용하여 두 번째 의사 고유벡터를 생성한다(203). 생성된 첫 번째 의사 고유벡터를 이용하여 새로운 정규화된 친화도 행렬을 생성하면, 생성된 새로운 정규화된 친화도 행렬에 거듭제곱 반복 방법을 이용하여 두 번째 의사 고유벡터를 생성한다. 다음으로 생성된 두 번째 의사 고유벡터에 해당하는 정규화된 친화도 행렬을 생성한다(204). 이러한 과정은 l의 값이 k의 값과 같이 질 때까지 반복된다(205).

[0034] 그리고 생성된 모든 의사 고유벡터를 K-평균 알고리즘을 이용하여 군집화한다(206). K-평균 알고리즘을 이용하여 모든 의사 고유벡터를 구분하여 군집화 할 수 있다. 생성된 의사 고유벡터는 상호 직교(Orthogonal)한다. 따라서 의사 고유벡터를 중복하여 계산하는 과정을 피할 수 있다.

[0035] 도 3은 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 알고리즘을 나타내는 도면이다.

[0036] 도 3을 참조하면, 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 정규화된 친화도 행렬 W를 입력하고, 거듭제곱 반복 알고리즘을 이용하여 v_l 을 계산하고, 디플레이션 방법을 이용하여 새로운 W_l 을 생성한다. 그리고 l값을 증가시키면서 반복수행하고 l값이 k값과 같아지면 반복을 중단하고, 생성된 의사 고유벡터를 K-평균 알고리즘을 이용하여 군집화한다. C_k는 k번째 분류된 군집(Cluster)이다.

[0037] 도 4a는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법과 스펙트럼 군집화 및 거듭제곱 반복 군집화를 비교한 결과를 나타내는 도면이고, 도 4b는 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화를 이용하여 뉴스 그룹 데이터세트를 군집화한 결과를 나타내는 도면이다.

[0038] 도 4a 및 도 4b를 참조하면, 문서, 손으로 적은 숫자, 얼굴 등에 대한 데이터세트에 대한 비교 결과를 포함한다. 6개의 데이터세트에서 13가지 실험을 수행하였다. 6개의 데이터세트는 숫자 데이터세트인 USPS 데이터세트, 필기체 숫자 데이터인 MNIST 데이터세트, 문서 데이터세트인 20 뉴스그룹(News-Group) 데이터세트,

TDT2 데이터세트, 6개 항목을 포함하는 문서 데이터세트인 로이터(Reuter) 데이터세트 및 얼굴 인식을 위한 UMist 데이터세트를 이용하였다. 또한, 이러한 데이터세트들 중에서 더 작은 단위의 데이터세트를 이용하여 실험을 진행했다. 도 4a 및 도 4b에서 y축은 정확도를 나타내고, x축은 t값을 나타낸다.

[0039] USPS3568 데이터세트의 실험결과(301)에서는 전반적으로 스펙트럼 군집화의 정확도가 낮게 나왔으며, 초기에는 디플레이션 기반의 거듭제곱 반복 군집화의 정확도가 가장 높았으나 이후 스펙트럼 군집화의 정확도와 거의 동일하게 감소하였으며, 후기를 제외하고 전반적으로 거듭제곱 반복 군집화의 정확도가 가장 높다.

[0040] MNIST3568 데이터세트의 실험결과(302)에서는 디플레이션 기반의 거듭제곱 반복 군집화의 정확도가 가장 높게 나타났으며 스펙트럼 군집화의 정확도가 가장 낮게 나타났다.

[0041] USPS0127 데이터세트의 실험결과(303)에서는 디플레이션 기반의 거듭제곱 반복 군집화와 스펙트럼 군집화의 정확도가 유사하게 나타났으며, 거듭제곱 반복 군집화는 중간 부분을 제외하고 전체적으로 낮은 정화도를 나타냈다.

[0042] MNIST0127 데이터세트의 실험결과(304)에서는 전반적으로 디플레이션 기반의 거듭제곱 반복 군집화의 정화도가 가장 높게 나타났으며, 초반에는 스펙트럼 군집화의 정확도가 디플레이션 기반의 거듭제곱 반복 군집화와 비슷하게 나타났으나 이후 급격히 감소한다. 거듭제곱 반복 군집화는 중간부분을 제외하고 전반적으로 가장 낮은 정확도를 나타낸다.

[0043] 로이터 데이터세트의 실험결과(305)에서는 디플레이션 기반의 거듭제곱 반복 군집화 방법의 정확도가 가장 높게 나타났으며, 거듭제곱 반복 군집화 방법의 순서로의 정확도가 가장 낮게 나타났다.

[0044] UMist 데이터세트의 실험결과(306)에서는 디플레이션 기반의 거듭제곱 반복 군집화 방법과 스펙트럼 군집화 방법의 정확도가 유사하게 나타나며, 거듭제곱 반복 군집화 방법은 상대적으로 낮은 정확도를 나타낸다.

[0045] 뉴스그룹a 데이터세트의 실험결과(307)에서는 초반이후, 세 가지 방법의 정확도가 모두 유사하게 나타났다.

[0046] 뉴스그룹b 데이터세트의 실험결과(308)에서는 디플레이션 기반의 거듭제곱 반복 군집화 방법과 스펙트럼 군집화 방법의 정확도가 유사하게 나타났으며, 거듭제곱 반복 군집화 방법의 정확도가 가장 낮게 나타났다.

[0047] 뉴스그룹c 데이터세트의 실험결과(309)에서는 디플레이션 기반의 거듭제곱 반복 군집화 방법의 정확도가 약간 높게 나타났으며, 거듭제곱 반복 군집화 방법의 정확도가 약간 낮게 나타났다.

[0048] 뉴스그룹d 데이터세트의 실험결과(310)에서는 불규칙적이지만 디플레이션 기반의 거듭제곱 반복 군집화 방법의 정확도가 약간 높게 나타났으며, 거듭제곱 반복 군집화 방법의 정확도가 가장 낮게 나타났다.

[0049] 몇몇 데이터세트의 경우 다른 경향이 나타나거나 세 가지 방법의 정확도가 유사하게 나타난 경우도 발생하였다. 이러한 특징은 각 데이터시트의 데이터 경향과 종류가 다르기 때문에 발생한 결과이다. 대부분의 데이터세트에서 디플레이션 기반의 거듭제곱 반복 군집화 방법과 스펙트럼 군집화 방법의 정확도가 상대적으로 높게 나타났으며, 거듭제곱 반복 군집화 방법의 정확도가 가장 낮게 나타나는 경향을 보인다. 결국, 본 발명에 따른 디플레이션 기반의 거듭제곱 반복 군집화 방법은 스펙트럼 군집화 방법의 정확도와 유사한 수준의 정확도를 가진다고 할 수 있다.

[0050] 이상 바람직한 실시 예를 들어 본 발명을 상세하게 설명하였으나, 본 발명은 전술한 실시 예에 한정되지 않고, 본 발명의 기술적 사상의 범위 내에서 당 분야에서 통상의 지식을 가진 자에 의하여 여러 가지 변형이 가능하다.

부호의 설명

[0051] 301: USPS3568 데이터세트의 실험결과

302: MNIST3568 데이터세트의 실험결과

303: USPS0127 데이터세트의 실험결과

304: MNIST0127 데이터세트의 실험결과

305: 로이터 데이터세트의 실험결과

306: UMist 데이터세트의 실험결과

307: 뉴스그룹a 데이터세트의 실험결과

308: 뉴스그룹b 데이터세트의 실험결과

309: 뉴스그룹c 데이터세트의 실험결과

310: 뉴스그룹d 데이터세트의 실험결과

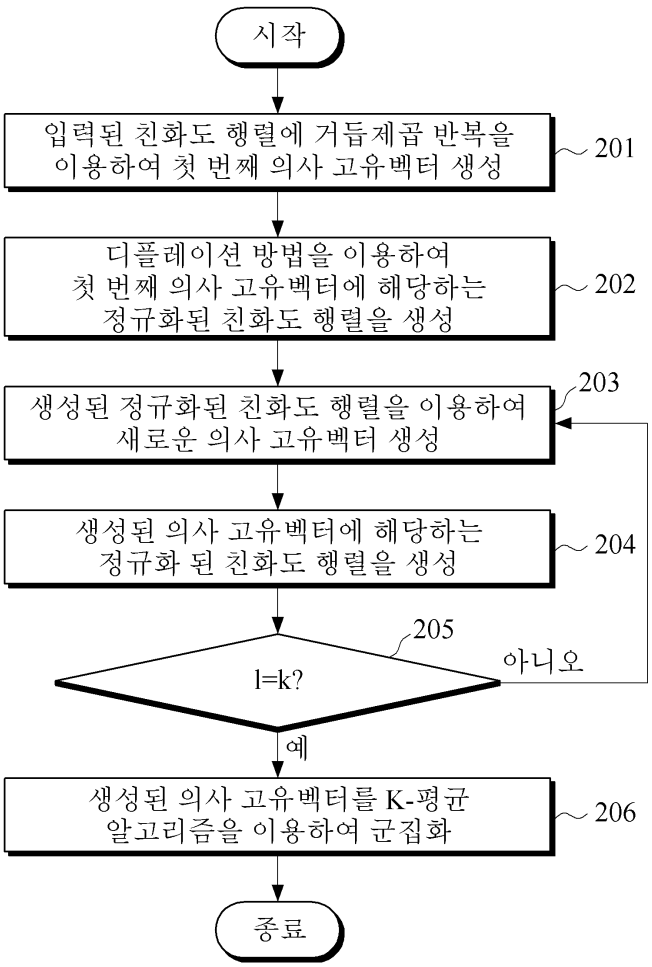
도면

도면1

Algorithm 1 Power iteration

Input: Normalized Affinity Matrix W.
repeat
$\upsilon^{t+1} = (W\upsilon^{t}) / \| W\upsilon^{t} \| 1$.
$\sigma^{t+1} = \| \upsilon^{t+1} - \upsilon^{t} \|$.
Increase t.
until $| \sigma^{t} - \sigma^{t-1} | < \varepsilon$
Output: Pseudo eigenvector υ^{t}.

도면2

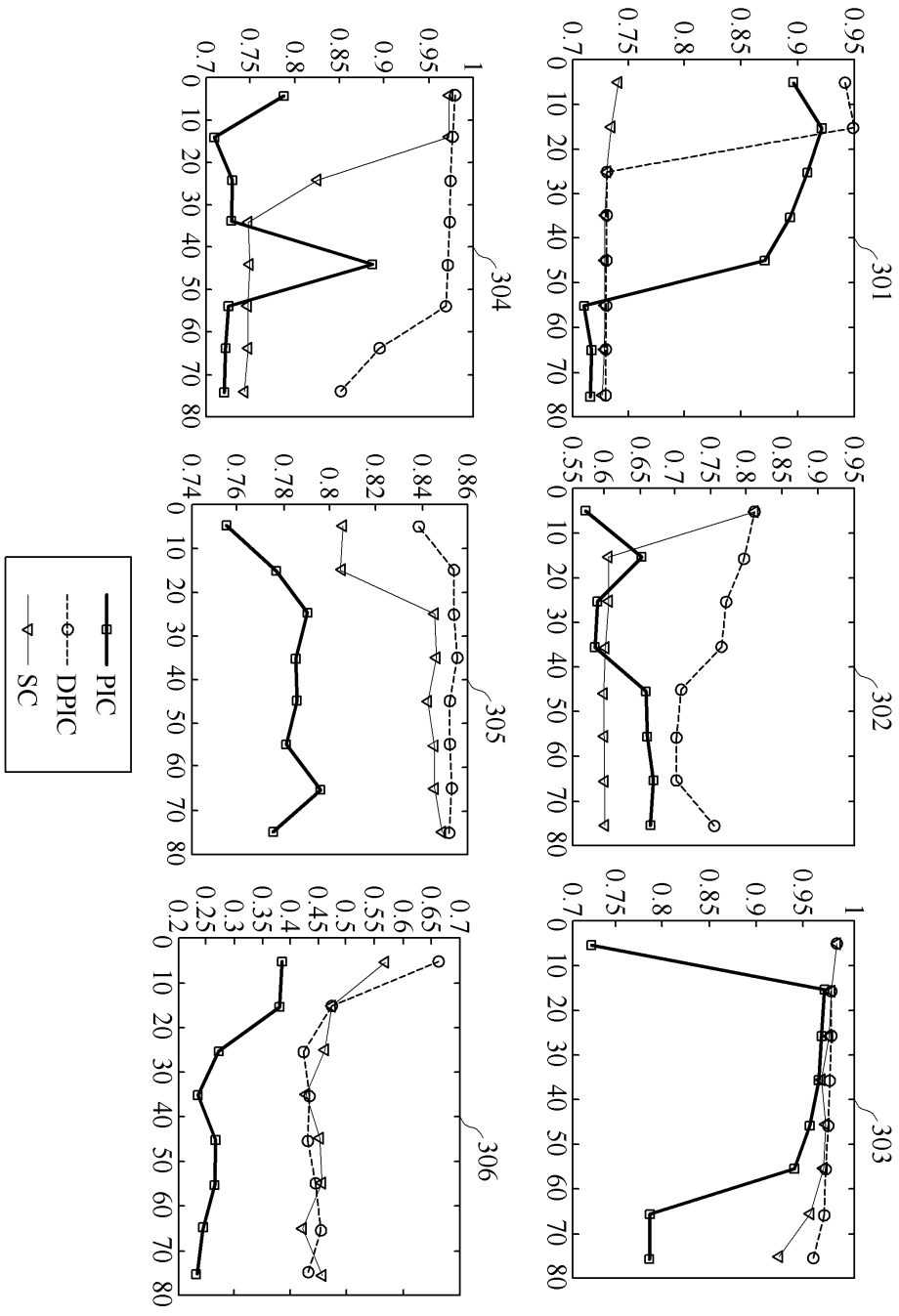


도면3

Algorithm 2 Deflation based Power Iteration Clustering (DPIC)

Input: Normalized Affinity Matrix W.
$W_0 = W$.
repeat
υ_l = PowerIteration(W_{l-1}). //*Power iteration: find* υ_l *from* W_{l-1}
$W_l = W_{l-1} - \frac{W_{l-1}\upsilon_l\upsilon_l^T W_{l-1}}{\upsilon_l^T W_{l-1}\upsilon_l}$.// *Deflation: remove* υ_l *effect on* W_{l-1}
Increase l.
until $l = k$
Use K-Means on pseudo eigenvectors $\upsilon_1, \upsilon_2, \ldots, \upsilon_k$.
Output: Clusters $C_1, C_2, \ldots, C_k$.

도면4a



도면4b