크기가 제한된 분할 시스템 구축을 위한 신호 교신 기반의 군집화 기법 연구

김선호, 이상현*
고려대학교

espirar1025@korea.ac.kr, *sanghyunlee@korea.ac.kr

A Study on the Clustering Algorithm by Passing Messages for Building Capacitated Clusters

Kim Sun Ho, Lee Sang Hyun*
Korea Univ.

요 약

'K-means'는 기법의 특성상 군집의 수용량을 제한하는 발전된 형태의 군집화 문제에 적용되기 부적절하다. 본논문에서는 이에 대안으로 개체간의 신호교신 모델을 활용하여, 일반적인 군집 구축 문제에서 크기 조건을 추가한 상황을 해결할 수 있는 군집화 기법을 제안하고 이를 관련 예제에 적용해본다.

Ⅰ. 서 론

일반적인 군집화 문제는 각 군집의 '알짜유사도(Net Similarity)'의 합을 최대로 만드는 군집형태 형성을 목표로 설계된다. 단, 이러한 문제환경은 각 군집의 크기를 제한할 수 없어서 용량의 상한 혹은 하한이 설정된 군집에 개체를 배치하는 상황에 적용하기엔 적절하지 않다. 따라서 정격화된 군집화 문제를 해결하기 위해서는 기존 군집화 기법을 변형하여 적용해야 한다.

대표적인 군집화 기법인 'K-means'는 군집의 개수 및 초기 중앙값을 입력으로 갖는 기법의 특성상, 동일한 문제에도 소요되는 시간이 비확정적이며, 최적해를 알지 못하는 상황에서는 알고리즘을 반복 수행하여 다른 결과로 수렴하는 모든 경우의 수를 탐색해야 최적해에 도달할 수 있다. 때문에 보다 복잡한 조건을 가진 군집화 문제에선 K-means 에 효율을 기대하기 어렵다.

본논문에서는 크기 조건을 추가한 군집화 문제의 해법으로 개체간의 신호교신 모델을 활용한 '정격 친화성 전파(Capacitated Affinity Propagation, 이하 CAP)'를 제안한다. 먼저 친화성 전파의 기본형을 소개하고, 이를 확장하여 군집용량을 제한시킨 형태의 문제환경에서 적용될 수 있도록 변형한 알고리즘을 설명한다. 후에는 이의 적용 예제로 다변량 정규분포 난수 샘플을 분할한 결과를 기재하였다.

Ⅱ. 기본 모델

'친화성 전파(Affinity Propagation, 이하 AP)'는 Brendan J. Frey 과 Delbert Dueck 에 의해 제안된 신호 교신(Message-Passing)[1] 기반의 군집화 기법이다. AP 는 개체군을 각 개체가 완전히 연결된 네트워크로 대응시켜 문제환경을 정의한 다음, 노드(개체) 간의 반복적인 신호 교신을 통해 각 군집을 대표하는 모범개체(Exemplar)와 각 모범개체를 따르는 일반개체들의 묶음으로 군집을 도출한다.

각 노드는 본인에게 가장 적합한 모범 개체를 선정하기 위해 실수값의 신호를 주고받는다. 신호는 크게 신임(Responsibility)과 유효성(Availability)으로 나눌 수 있다. 먼저 개체 i가 개체 k에게 송신하는 신임 신호 'r(i, k)'는 '개체 k 가 개체 i 의 모범 후보로서 적합한지에 대한 평가치'이다. 유효성 신호 'a(i, k)'는 반대로 개체 k 가 개체 i 에게 송신하며, '모범 후보인 개체 k 가 속한 군집에 개체 i 를 가입시키는 것에 대한 평가치'이다. 두 신호는 모두 반복되는 과정을 통해서 수정된다. 신호 교신의 결과로 각 노드는 자신의 모범 개체가 속한 군집에 가입되는 형태로 개체군이 분할된다(모범 개체는 자기자신을 모범으로 채택한다.)

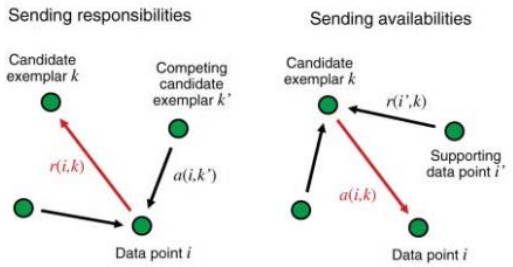

그림 1. 신임 신호와 유효성 신호[2]

'$s(i,k)$'를 개체 i 와 k 간의 유사도(Similarity) 라고 할 때($s(k,k)$는 선호도, 즉 모범으로 채택되려는 성향

이다), AP 에서 신호의 갱신 규칙과 AP 알고리즘의 전체 과정은 아래와 같다.[2][3]

$$r(i,k) \leftarrow s(i,k) - \max_{k's.t.k' \neq k}\{a(i,k') + s(i,k')\}$$

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i's.t.i' \notin \{i,k\}} \max\{0, r(i',k)\}\} \quad (i \neq k)$$

$$a(k,k) \leftarrow \sum_{i's.t.i' \neq k} \max\{0, r(i',k)\}$$

Algorithm 1: Affinity Propagation

Initialize $t \leftarrow 0, r(i,k) \leftarrow 0, a(i,k) \leftarrow 0$ for all (i, k)

Repeat

Update r(i, k) messages and send to neighbors

Update a(i, k) messages and send to neighbors

Increase $t \leftarrow t+1$

Until $|\Delta a(i,k)| < \sigma$ *for all* (i,k) or *iteration count* $> t_{max}$.

Decoding Decisions

The set $\mathcal{K} = \{k | r(k,k) + a(k,k) > 0\}$ is choosen as the exemplar set. And for each non-exemplar point i, $\hat{k}_i = argmax_{k \in \mathcal{K}}(r(i,k) + a(i,k))$.

Ⅲ. 정격 친화성 전파(CAP) 모델

CAP 알고리즘은 위 **Algorithm 1** 의 내용을 그대로 따르되, 추가된 제약조건에 맞춰 신호갱신규칙을 변경한 형태이다. 군집 크기의 상한과 하한을 각각 U 와 L 이라 했을 때, CAP 신호갱신규칙은 다음과 같이 나타낼 수 있다.

$$r(i,k) \leftarrow s(i,k) - \max_{k's.t.k' \neq k}\{a(i,k') + s(i,k')\}$$

$$a(i \neq k, k) \leftarrow \min\left[\begin{array}{c} -\min\{0, r(\tilde{k}_{L-1}, k)\} - \max\{0, r(\tilde{k}_{U-1}, k)\}, \\ \sum_{i' \in (\{k\} \cup \tilde{K}_{L-2})} r(i',k) + \sum_{i'' \in \tilde{K}_{U-2} \setminus \tilde{K}_{L-2}} \max\{0, r(i'',k)\} \end{array}\right]$$

$$a(i=k,k) \leftarrow \sum_{i' \in \tilde{K}_{L-1}} r(i',k) + \sum_{i'' \in \tilde{K}_{U-1} \setminus \tilde{K}_{L-1}} \max(0, r(i'',k))$$

s.t.

$$\tilde{K}_{n+1} = \bigcup\{\tilde{K}_n, \{argmax_{i' \notin (\{i,k\} \cup \tilde{K}_n)} r(i',k)\}\}, \tilde{K}_0 = \varnothing$$

$$\tilde{k}_{n+1} = argmax_{i' \notin (\{i,k\} \cup \tilde{K}_n)} r(i',k)$$

$\tilde{K}_n$는 $k' \notin \{i,k\}$인 k'에 대해 신호 r(i,k')의 가장 큰 값부터 n 번째 큰 값 까지 대응되는 각각의 k'의 집합이다.

Ⅳ. 시뮬레이션 및 결론

	Center	Covariance matrix	# of samples
G_1	(0,0)	[1,0; 0,1]	60
G_2	(2,-5)	[0.5,0; 0,0.5]	20
G_3	(4,0)	[1,-0.5;-0.5,1]	120

표 1.다변량 정규분포 난수 샘플에 관한 정보

고안한 알고리즘의 성능을 확인해보기 위해, 크기가 다른 세 다변량 정규분포 난수를 생성하였다. 각 난수 샘플의 크기와 모수는 **표 1** 에 기재된 값들과 같다.

해당 샘플을 바탕으로 K=3 인 K-means 와 AP, U 와 L 이 각각 40, 20 인 CAP 를 구동하였다. 단, 서로 다른 (i,k)에 대하여 s(i,k)는 두 개체 간의 유클리드 거리에 음의 부호(-)를 붙인 값으로 설정했으며, 선호도 s(k,k)는 모든 k 에 대하여 같은 값을 같게 하되, AP 에서 세 군집이 형성되도록 조정하였다.

시뮬레이션 결과는 아래 **그림 2** 와 같다. K-means 와 AP 에서와는 달리, CAP 에서는 군집이 용량 규제에 맞춰 형성했음을 확인할 수 있다.

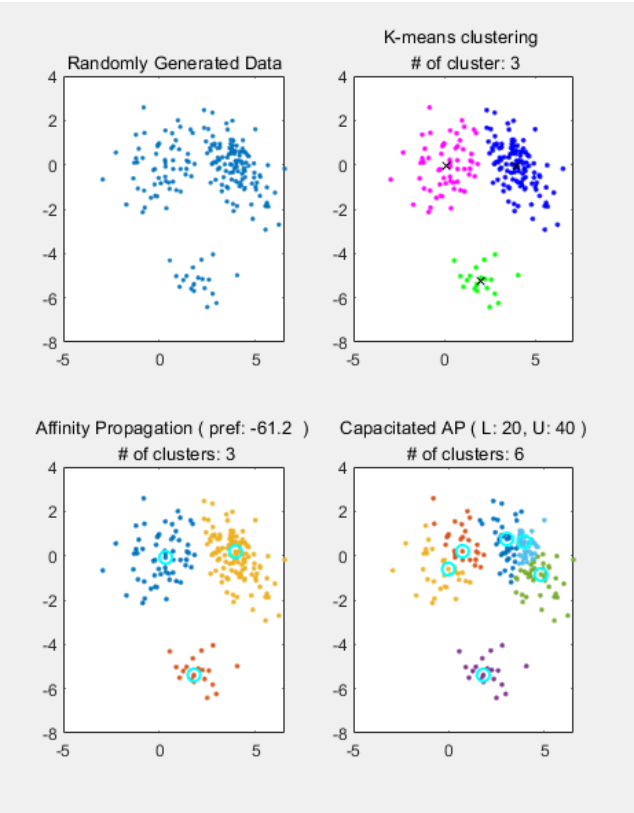

그림 2.난수 샘플, K-means 결과, AP 결과, CAP 결과

ACKNOWLEDGMENT

본 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1A2C1084855).

참 고 문 헌

[1] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Transactions on information theory, vol. 47, no. 2, pp. 498-519, 2001.

[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," science, vol. 315, no. 5814, pp. 972-976, 2007.

[3] I.-E. Givoni, Beyond affinity propagation: Message passing algorithms for clus-tering. Citeseer, 2012