A K-means-like Algorithm for K-medoids Clustering

이 종 석 박 해 상 전 치 혁
{jongseok, shoo359, chjun}@postech.ac.kr
포항공과대학교 기계산업공학부
경상북도 포항시 남구 효자동 산31

Abstract

Clustering analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. In this paper we propose a new algorithm for K-medoids clustering which runs like the K-means algorithm. The new algorithm calculates distance matrix once and uses it for finding new medoids at every iterative step. We evaluate the proposed method using real and synthetic data and compare with the results of other algorithms. The proposed algorithm takes reduced time in computation and better performance than others.

1. Introduction

하나의 객체가 일정한 수의 속성을 갖고 이러한 객체가 다수 있다고 할 때 군집 분석이란 유사한 속성들을 갖는 객체들을 묶어 전체의 객체들을 몇 개의 군집으로 나누는 것을 말한다. 이를 위한 비계층적 방법으로 K-means 군집 방법, K-medoids 군집 방법이 많이 사용되고 있다.

사전에 정해진 군집의 수 k 를 바탕으로 K-means 군집 방법은 k 개의 중심 좌표를 선정하여 각 객체와의 거리를 산출한 후 가장 가까운 군집에 그 객체를 배정하는 방법이고 ([1], [2]) K-medoids 군집 방법은 k 개의 군집의 대표 객체(medoids)를 정하고 객체와 그가 속하는 군집의 대표 객체와의 거리의 총합을 최소로 하는 방법이다. 여기서 군집의 대표 객체란 그 군집에 속하는 객체 중 다른 객체와의 거리가 최소가 되는 객체를 의미한다.

그러나 K-means 군집 방법은 이상치가 있을 경우 성능이 떨어진다는 단점이 있고 K-medoids 군집 방법은 이상치에는 덜 민감하나 계산 시간이 오래 걸린다는 단점이 있다.

이 논문은 이 두 방법의 특징을 혼합한 새로운 군집 방법인 K-means like K-medoids를 제안하고 알고리즘의 성능을 비교 분석하였다.

2. Proposed Algorithm

각 객체가 p 개의 변수를 갖고 객체 i 의 j 번째 변수를 X_{ij} (i =1,..., n ; j = 1,..., p) 라 하고 모든 객체를 k 개의 군집으로 분류한다고 할 때 K-means like K-medoids 방법의 알고리즘은 다음과 같다.

단계 0 : (초기 대표 객체를 선정)

0-1. 비유사성 척도로 유클리드 거리(Euclidean distance)를 생각하였을 때 모든 객체 사이의 거리를 다음과 같이 계산한다.

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (X_{ia} - X_{ja})^2} \quad i = 1,...,n \quad j = 1,...,n \qquad (1)$$

0-2. 중심에 위치하는 객체를 선정하기 위해 다음을 계산한다.

$$p_{ij} = \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \quad i = 1,\ldots,n \quad j = 1,\ldots,n \qquad (2)$$

0-3. 각 객체마다 $\sum_{i=1}^{n} p_{ij}$ 을 계산하여 값이 증가하는 순으로 정렬한 뒤 가장 작은 값을 가지는 k 개의 객체를 초기 대표 객체로 선정한다. i 번째 객체가 대표 객체일 때 $b_i = 1$로 정의한다.

$$b_j = 1, \quad \sum_{i=1}^{n} b_i = k \qquad (3)$$

0-4. 대표 객체로 선정되지 않은 객체에 대하여 k 개의 대표 객체와의 거리를 산출한 후 가장 가까운 군집에 그 객체를 배정한다. j 번째 객체가 대표 객체인 군집에 i 번째 객체가 속한다면 a_{ij} = 1로 정의한다.

$$a_{iv_i} = 1 \quad for \quad v_i = \arg\min(d_{ij}) \qquad (4)$$

0-5. 각 군집 내의 모든 객체와 대표 객체와의 거리의 총합을 현재의 최적 값으로 정한다.

$$Z = \sum_{i=1}^{n} d_{iv_i} \qquad (5)$$

단계 1 : (새로운 대표 객체 산출)

1-1. 같은 군집 내에서 객체 사이의 거리의 합이 최소가 되는 객체를 새로운 대표 객체로 선정한다.

1-2. k 개의 군집에 대해, k 개의 새로운 대표 객체를 선정한다.

단계 2 : (객체의 군집 배정)

2-1. 대표 객체로 선정되지 않은 객체에 대하여 k 개의 대표 객체와의 거리를 산출한 후 가장 가까운 군집에 그 객체를 배정한다.

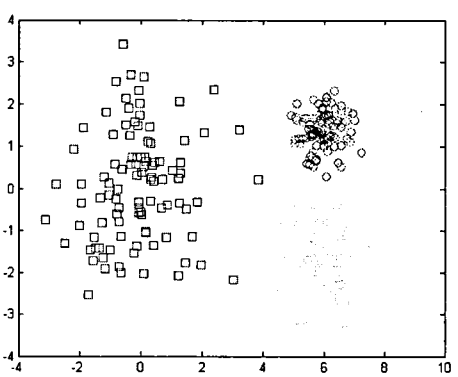
2-2. 새로운 군집에서 대표 객체와 객체 사이의 거리의 총합을 계산한다. 이전 최적 값과 비교하여 같으면 마치며, 그렇지 않으면 단계 1을 반복한다.

3. Examples

제안된 알고리즘의 성능을 알아보기 위하여 K-means 군집 방법과 비교하였다.

3.1. Artificial data

[그림 1]과 같이 총 세 개의 군집을 갖는 2차원 데이터를 생성하였다. 같은 군집의 객체는 같은 모양으로 표시하였고 편의상 사각형으로 표시된 군집을 군집 A , 원으로 표시된 군집을 군집 B , 삼각형으로 표시된 군집을 군집 C 로 명하겠다. 군집 A는 평균 (0, 0), 분산 (1.3, 1.3), 군집 B는 평균 (6, 1.5), 분산 (0.5, 0.5), 군집 C는 평균 (6, -1.5) 분산 (0.7, 0.7)인 다변량 정규분포에서 각 100개씩 임의로 추출하였다.

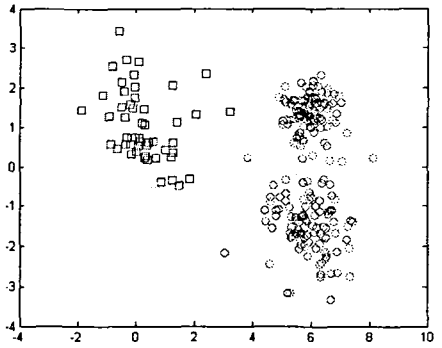


[그림 1] 알고리즘 비교를 위한 데이터

Matlab 7.0 소프트웨어를 이용하여 K-means 군집 방법 (k =3)으로 군집 분석을 수행하여 [그림 2]와 같은 결과를 얻었다.

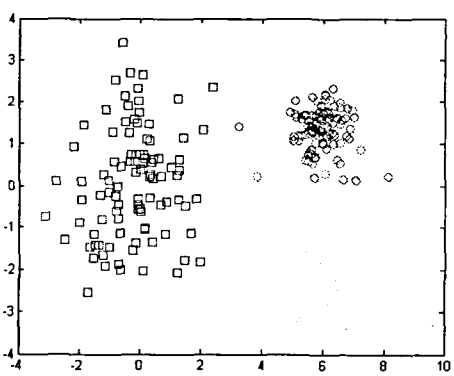
[그림 2]에서 보면 군집 B와 군집 C를 제대로 구분하지 못하고 있다. 군집의 개수 k 를 3으로 정하였기 때문에 군집 B와 군집 C를 나누는 대신 군집 A를 두 개의 군집으로

구분하는 것을 볼 수 있다.



[그림 2] K-means로 분석한 결과

그러나 [그림 3]에서 보듯이 제안된 알고리즘으로 분석하면 거의 완벽하게 구분하고 있음을 알 수 있다. 두 군집이 비록 밀집되어 있더라도 평균의 차이가 크지 않고 또 어느 한 군집의 분포는 많이 퍼져 있을 때 제안된 방법이 좋음을 알 수 있다.



[그림 3] K-means like K-medoids 방법으로 분석한 결과

[표 1]은 위의 데이터에 대해 K-means 군집 방법, K-means like K-medoids 방법, K-medoids 군집 방법의 하나인 PAM(Partitioning Around Medoids) 군집 방법([3])으로 분석하였을 때 객체 사이의 거리의 총합을 구한 것이다. 중심 거리란 각 군집의 중심에서 객체까지의 거리의 총합을 의미하고 대표 거리란 각 군집의 대표 객체에서 객체까지의 거리의 총합을 의미한다. K-means liked K-modoids의 거리의 합이 더 작은 것을 알 수 있다.

[표 1] 군집 결과 거리 비교

	중심 거리	대표 거리
K-means	430.18	429.22
제안된 방법	299.93	301.16
PAM	299.93	301.16

3.2. Iris data

1935년 Anderson에 의해 수집된 붓꽃 데이터([5])로 K-means 군집 분석과 비교하였다. 붓꽃 데이터는 총 150개로 4개의 측정값을 가지며 50개씩, 3개의 종으로 구분된다.

K-means 군집 방법으로 분석한 결과를 보면 각각 50개, 61개, 39개로 나뉘며 실제 값과 비교할 때 총 150개 데이터 중 133개를 정확히 분류하고 있다.

K-means like K-medoids 방법으로 분석한 결과를 보면 각각 50개, 44개, 56개로 나뉜다. 실제 값과 비교할 때 총 150개 중 138개를 정확히 분류하고 있으므로 K-means 군집 방법보다 성능이 더 뛰어나다고 할 수 있다. [표 2]는 각 방법의 거리를 비교한 것이다.

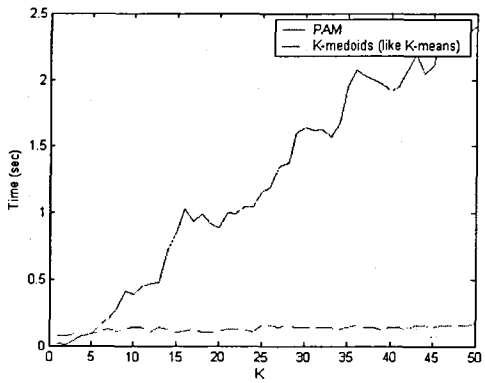
[표 2] 군집 결과 거리 비교

	중심 거리	대표 거리
K-means	101.85	102.60
제안된 방법	102.19	103.40
PAM	101.82	102.56

4. Simulation

R 2.1.1 소프트웨어를 이용하여 PAM 군집 방법과 K-means like K-medoids 방법의 계산 시간을 비교하였다.

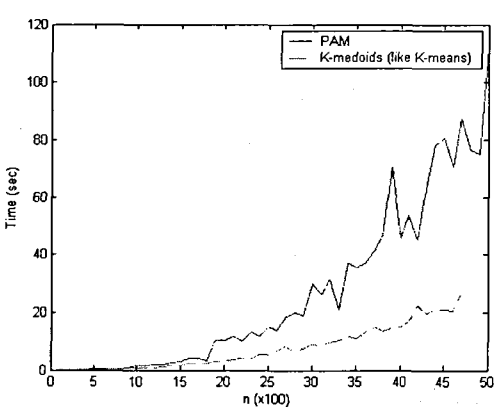
[그림 4]는 임의로 추출한 객체가 300개일 때 군집의 개수를 증가시켰을 때의 계산 시간을 그래프로 나타낸 것이다.



[그림 4] k의 증가에 따른 소요 시간 비교

그림에서 알 수 있듯이 K-means like K-medoids 방법의 계산 시간은 k의 증가에 따른 영향을 받지 않는다는 것을 알 수 있다. 이에 반해 PAM 군집 방법은 k 가 증가함에 따라 계산 시간이 크게 증가하고 있다.

[그림 5]는 변수의 수를 2로, 군집의 수를 3으로 고정하고 객체 수를 증가시켰을 때의 계산 시간을 그래프로 나타낸 것이다. 모든 계산이 끝날 때까지 소요 시간을 측정하는 것이므로 객체는 임의 추출하였다.



[그림 5] n의 증가에 따른 소요 시간 비교

역시 K-means like K-medoids 방법의 계산 시간이 PAM 군집 방법에 비해 적게 걸리는 것을 알 수 있다.

5. Conclusion and Future Works

본 논문은 새로운 군집 방법을 제안하였다. 제안된 방법은 모든 객체간의 거리를 단계 0에서 한번만 계산하고 반복 단계에서는 이미 계산된 거리를 이용한다. 따라서 알고리즘을 수행하는 과정에서 더 이상 복잡한 계산이 필요 없는 특징이 있다. 이러한 특징으로 인해 매 단계마다 거리를 계산하여야 하는 기존의 PAM 군집 방법에 비해 계산 시간을 크게 줄일 수 있다. 또한 군집의 중심 좌표를 기준으로 객체와의 거리를 계산하는 K-means 군집 방법의 특성상 잘 분리해내지 못하는 데이터에 대해서도 뛰어난 성능을 보여준다. 그러나 초기 대표 객체를 선정하는 과정에서 제안된 방법에서는 객체들의 가장 중심에 위치하는 k 개를 선택하도록 하였다. 이렇게 하였을 시 데이터의 형태에 따라 더 나은 대표 객체를 찾기 위한 알고리즘 수행 횟수가 늘어날 수 있다. 반복 횟수를 줄일 수 있는 초기 대표 객체를 찾는 것이 추후 연구 분야이다.

References

[1] MacQueen, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,", Berkeley, University of California Press, Vol. 1, 281-297.

[2] Hartigan, J.A. (1975), Clustering Algorithms, New York, NY: Wiley

[3] Kaufman, L. and Rousseeuw, P.J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York

[4] E. Anderson (1935), "The irises of the Gaspé peninsula", Bulletin of the American Iris Society 59, 2-5.