K-Means 알고리즘을 이용한 다차원 데이터 클러스터링 기법 구현

임선영[1], 신현순[2], 박영호[1,*]
[1]숙명여자대학교 멀티미디어과학과
[2]한국전자통신연구원
*교신저자
e-mail: sunnyihm@sm.ac.kr, hsshin@etri.re.kr, yhpark@sm.ac.kr

An Implementation of Clustering Method using K-Means Algorithm on Multi-Dimensional Data

Sun-Young Ihm[1], HyunSoon Shin[2], Young-Ho Park[1,*]
[1]Dept. of Multimedia Science, Sookmyung Women's University
[2]Electronic and Telecommunications Research Institute
*Corresponding Author

요 약

K-Means 클러스터링 기법은 데이터마이닝 분야 중 클러스터링 분야에서 가장 널리 쓰이는 방법 중 하나로 주어진 데이터 셋에서 k개의 클러스터를 중심으로 데이터를 분할하는 기법이다. 최근의 데이터는 여러개의 속성을 고려해야 한다. 따라서 본 논문에서는 K-Means 클러스터링 기법을 소개하고, 또 K-Means 클러스터링 기법을 여러 개의 속성을 고려하기 위하여 다차원 데이터에 적용한 실험을 소개한다.

1. 서론

클러스터링 기법은 데이터마이닝 분야에서 가장 기본적이고 중요한 문제 중 하나로 인식되어 왔다. 클러스터링 기법은 서로 관련이 있는 데이터들을 클러스터로 형성하는 방법이다 [4] K-means 클러스터링 기법은 클러스터링 기법 중 가장 유명하고 널리 쓰이는 방법 중 하나이다 [1]. K-Means 기법은 가장 일반적으로 사용되는 분할 클러스터링 기법 중 하나로, 기본적인 개념은 데이터와 그 데이터가 속하는 클러스터의 중심과의 평균 유클리디안 (Euclidean) 거리를 최소화 하는 것이다. K-Means 기법은 구현이 쉽고, 데이터의 수가 n 일 때, 시간 복잡도가 $O(n)$ 인 장점을 가지고 있다 [4].

최근에 사용되는 데이터는 여러 개의 속성을 가지고 있다. 예를 들어, 중고차 검색을 할 때에도 검색 속성으로 차종, 연식, 연비, 가격, 색상 등 여러 개의 속성을 고려해야 한다. 이러한 속성들은 공간상의 차원으로 표현될 수 있다. 예를 들어, 첫 번째 속성은 1차원을 뜻하는 x축으로, 두 번째 속성은 2차원을 뜻하는 y축으로 표현하면, 데이터를 공간상에 표현할 수 있다. 위의 중고차 검색 예제에서는 총 다섯 개의 속성을 가지므로, 5차원의 공간에 표현이 가능하다. 최근에는 데이터의 속성이 많아져 다차원 데이터에 대한 관심이 높아지고 있다. 따라서 본 논문에서는 여러 개의 속성을 고려하기 위하여 다차원의 데이터에서 K-Means 클러스터링 기법을 적용해 보기로 한다. 본 논문에서는 K-Means 클러스터링 기법에 대해 소개하고, K-Means 클러스터링 기법을 다차원의 데이터에 실험해 본다. 본 논문의 구성은 다음과 같다. 2장에서는 K-Means 클러스터링 기법에 대해 설명하고, 3장에서는 K-Means 클러스터링 기법을 다차원 데이터에 대하여 실험한 결과를 소개하고 마지막으로 4장에서는 결론을 짓는다.

2. K-Means 클러스터링 기법

K-Means 클러스터링 기법은 데이터를 k개의 클러스터로 자동적으로 나누는 가장 일반적인 방법 중 하나이다. 이 기법은 초기에 k개의 중심 클러스터를 설정한 후, 이 클러스터들을 계속 반복적으로 계산해 나간다 [3]. 예를 들어 d-차원을 가지는 공간 R^d 상에 n개의 데이터로 이루어진 집합 P가 있다고 하고, 클러스터의 개수 k가 주어졌다고 했을 때, K-Means 클러스터링 기법은 먼저 k개의 중심점을 구한다. 그리고 각 k개의 중심점으로부터 유클리디안 (Euclidean) 거리가 가까운 데이터 포인트들을 구하여 클러스터 집합에 포함시킨다 [2].

그림 1은 K-Means 알고리즘의 수도 코드를 나타내고 있다. 행 4-5열에서는 각각의 데이터를 가장 가까운 중점에 할당하고, 행 7-8열에서는 각각의 중점들을 할당된 점들의 평균값으로 다시 계산한다. K-Means 알고리즘에서는 이러한 두 단계를 반복적으로 수행한다. K-Means 클러스터링 기법에서 중심점은 각 클러스터에 속한 데이터의 대푯값을 뜻한다. K-Means 알고리즘의 대부분의 수행

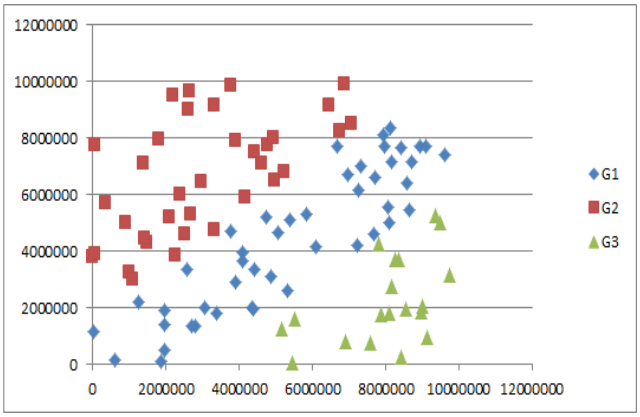
시간은 점 n개와 중심점 k개 사이의 거리를 계산하는데 쓰인다 [5]. 이러한 K-Means 클러스터링 기법의 결과는 그림 2에 나타나 있다. 4차원 이상에서의 K-Means 클러스터링 결과는 표현할 수가 없다. 따라서 그림 2에서는 2차원의 데이터 셋을 사용하여 K-Means 클러스터링 결과를 보여주고 있다. 데이터 셋은 100개의 데이터로 구성되어 있으며 두 개의 속성을 가지고 있다. 총 3개의 그룹 G1, G2, G3로 분할하였으며, 빨간색 네모는 G1을 파란색 마름모는 G2를 초록색 세모는 G3를 뜻한다.

Algorithm KMEANS(dataset X, initial centers C_{init})

X: a set of N data points
C_{init}: initial centers of k clusters
C: cluster centers of k clusters
P= {p(i) | i= 1, …, N} is the cluster label of X

1: $C \leftarrow C_{init}$
2: while ($C \neq C_{perv}$) do {
3: $C_{prev} \leftarrow C$
4: for each data point x_i in X do
5: $p(i) \leftarrow argmin_c \in c^d(xi, c)$
6: for each centers c_j in C do
7: $c_j \leftarrow$ average of x_i, whose p(i) = j
8: }
9: return (C, P)

(그림 1) K-Means 알고리즘의 수도코드 [5]



(그림 2) K-Means 클러스터링 기법의 예

3. 다차원 데이터에서의 K-Means 클러스터링 기법

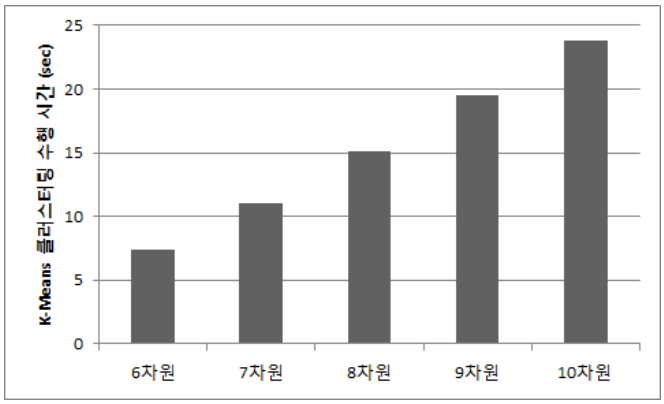
본 논문에서는 여러 개의 속성을 고려하기 위하여 K-Means 클러스터링 기법을 다차원 데이터에 적용하여 실험하였다. 실험환경으로는 인텔 i5-760 쿼드 코어 프로세서를 탑재한 2.80GHz의 리눅스 PC를 사용하였다. 16GB 메인 메모리를 사용하였으며, c++를 사용하여 구현되었다 [6].

실험을 위해 100K의 크기의 Uniform 분포를 가지는 데이터셋을 사용하였으며, 6-10 차원의 데이터 셋에 대하여 K-Means 클러스터링 기법을 적용하였다. 클러스터의 개수 k는 10개로 설정하였다. 표 1은 K-Means 클러스터링 기법을 적용했을 때, 분할된 각 그룹에 속한 데이터의 개수를 나타낸다. G1, G2, …, G10은 총 10개의 그룹을 나타낸다. 표 1을 통해 각각의 데이터 셋이 K-Means 클러스터링 기법을 적용하여 모두 6개의 공간으로 분할 된 것을 알 수 있다. 각 분할된 그룹에 속한 데이터의 개수를 살펴보면 Uniform 분포의 데이터 셋을 사용하였으므로 비교적 균등하게 분할된 것을 볼 수 있다.

<표 1> 다차원 데이터에서의 K-Means 클러스터링 기법 실험 결과

	6차원	7차원	8차원	9차원	10차원
G1	9,233	7,258	10,729	9,461	10,262
G2	10,409	10,740	11,622	10,326	9,050
G3	11,328	10,199	10,780	10,201	10,951
G4	10,381	10,153	10,632	10,765	10,921
G5	10,444	11,565	10,223	11,318	10,667
G6	10,629	9,961	10,630	8,292	9,997
G7	9,157	9,551	10,375	10,841	10,829
G8	8,595	11,168	5,843	9,101	6,809
G9	9,635	11,038	10,371	10,477	9,989
G10	10,189	8,097	8,795	9,218	10,519

그림 3은 K-Means 클러스터링 기법의 수행 시간을 나타내고 있다. 차원이 높아질수록 수행 시간이 길어지며, $O(n)$의 시간 복잡도를 보이고 있다.



(그림 3) K-Means 클러스터링 기법의 수행 시간

4. 결론 및 향후연구

본 논문에서는 데이터마이닝 기법 중 하나인 K-Means 클러스터링 기법을 소개하고 이를 다차원 데이터에 적용하는 실험을 설명하였다. 다차원 데이터에서의 분할 기법은 데이터마이닝을 포함하여 다양한 분야에 적용될 수 있다.

향후 연구로는 다차원 데이터에서의 K-Means 클러스터링 기법을 사용한 빠른 분할 기법 및 검색 방법에 대한 연구를 진행하고자 한다.

5. 사사문구

본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041854, 안전한 주거환경을 위한 실시간 위험요소 예측/방지용 스마트 홈 서비스 플랫폼 기술 개발]

참고문헌

[1] Dhillon I.S. Guan Y. Kulis B. "Kernel k-means: Spectral Clustering and Normalized Cuts", ACM Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.551-556 2004.

[2] Kanungo T. Mount D.M. Netanyahu N.S. Piatko C.D. Silverman R. Wu A.Y. "A local search approximation algorithm for k-means clustering", Computational Geometry, 28 pp.89-112, 2004.

[3] Wagsta K. Cardie C. Rogers S. Schroedl S. "Constrained K-means Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, pp.577-584, 2001.

[4] 이신원, "K-Means 클러스터링에서 초기 중심 선정 방법 비교," 한국인터넷정보학회논문지 제 13권 제 6호, pp.1-8, 2012.

[5] 윤태식, 심규석, "고차원 대규모 데이터 처리를 위한 K-Means 클러스터링," 정보과학회논문지 제 18권 제 1호, pp.55-59, 2012.

[6] K-Means 소스코드, www.cs.umd.edu/~mount