k 평균 군집화의 그리드 기반 효율적인 k 추정 기법

문철한 [1], 김민형 [1], 민준기 [2]
[1] 한국기술교육대학교 컴퓨터공학과 박사과정
[2] 한국기술교육대학교 컴퓨터공학과 교수

pneuma@koreatech.ac.kr, kimexcel2@koreatech.ac.kr, jkmin@koreatech.ac.kr

A Grid-based Efficient k-estimation Method for k-means Clustering

Cheolhan Moon[1], Min Hyung Kim[1], Jun-Ki Min[2]
[1,2]Dept. of Computer Science and Engineering, Korea University of Technology and Education

요 약

k-평균 군집화 분석은 주어진 데이터 집합에서 유사한 데이터의 군집을 찾고 이를 분석하는 기법으로 널리 사용되어 왔다. 그러나, 데이터 집합에 대해 적절한 k를 찾는 방법에 대한 연구는 부족하며 평가 계수의 전수 비교에 의존해 왔다. 본 연구에서는 이를 경감하기 위한 그리드 기반 근사 군집 분석을 통한 k 추정 기법을 제안한다. 제안 기법의 성능을 10,000~100,000 개의 데이터를 갖는 생성 데이터 집합을 활용한 검증 결과 다양한 k와 데이터 집합의 크기에 대해 효과적임을 보였다.

1. 서론

최근 데이터 마이닝을 활용한 소비자 분석이 다양한 분야에서 연구되고 있으며, 데이터 마이닝의 일종으로 주어진 데이터 집합에서 서로 다른 특징을 갖는 군집들을 찾는 군집화 분석 기법이 있다. 대표적인 군집화 분석 기법인 k-평균 군집화 분석 기법은 k 개의 군집을 생성하여 군집 내 데이터들의 특징을 분석하여 산업 및 경제 등 다양하게 응용된다 [1].

이러한 k-평균 군집화 분석 기법의 주요한 문제점으로는 데이터 집합에 대하여 최적인 k 의 값을 사전에 알 수 없다는 점이 있다. 따라서, 기존의 연구에서는 분류 작업처럼 정확도 기반의 평가를 수행하거나 몇 가지 평가 계수들을 적용하여 적합한 k 를 선정한다. 그러나, 이와 같은 선정 방법은 서로 다른 다양한 k 에 대해 k-평균 군집화를 모두 수행, 그 결과들을 한다. 즉, 불필요한 군집화 과정을 추가로 과도하게 수행하는 문제점이 있다 [2].

따라서, 본 논문에서는 이러한 문제점을 줄이기 위하여 그리드 기반의 k 추정 기법을 제안한다. 기존의 모든 범위의 k 에 대하여 k-평균 군집화를 수행하고 이를 비교해야 했던 것과 달리, 후보를 빠르게 도출하고 이를 기반으로 k를 빠르게 추정할 수 있다.

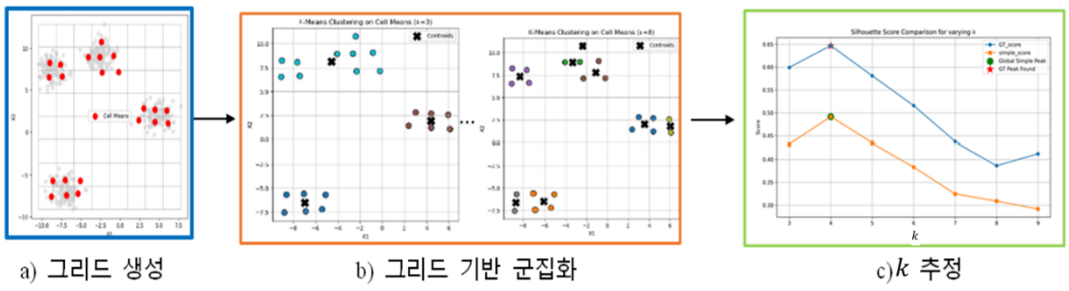
2. 배경지식 및 관련 연구

k-평균 군집화 분석은 하나의 데이터 집합 D 에 대하여 설정된 k 개의 군집을 찾기 위해 다음과 같은 과정을 수행한다 [3].

먼저, k 개의 중심점의 좌표를 D 의 도메인 내에서 임의로 선정한다. 이후, 각 데이터로부터 가장 가까운 중심점을 찾아 t=0 시점의 초기 클러스터를 선정한다. 선정된 초기 클러스터에 속한 데이터들의 평균 지점으로 새로운 k 개의 중심점을 선정한다. 새롭게 선정된 중심점을 기준으로 t=1 시점의 클러스터를 선정한다. 이와 같은 방식으로 정해진 반복 횟수를 만족하거나 중심점의 위치 변동이 일정한 수준 이하일 경우 반복을 중단한다. 중단 시점에 각 데이터가 속한 클러스터가 군집 결과가 된다.

k-평균 군집화 분석은 대표적인 비지도 학습으로서 산업 및 경제 등 다양한 분야에 널리 활용되고 있어, 그 결과에 대한 평가 방법 및 효율성 증대 방안 역시 다양하게 연구되고 있다.

기존의 군집화 분석 결과를 평가하는 방법으로는 평가 계수들을 활용하여 군집화 결과를 평가하는 방법들이 있으며 대표적 인 평가 계수로는 실루엣 계수 (Silhouette Coefficient)가 있다 [4]. 실루엣 계수의 계산 방식은 각 클러스터마다 a_i와 b_i를 통해 계산된 값들



(그림 1) 그리드 기반의 k 추정 기법 단계별 시각화

의 평균이며, 다음 식과 같이 계산된다.

$$S = \frac{1}{k}\sum_{i=1}^{k}\frac{b_i - a_i}{max(b_i, a_i)}$$

식 1 에서 a_i는 각 군집에 속한 데이터 x마다 그 군집의 중심점 μ_i까지 거리를 모두 합한 것을 의미하며, b_i는 각 데이터마다 자신이 속하지 않은 가장 가까운 클러스터의 중심점까지의 거리의 합이다. 이와 같은 평가 방법들 또한 전체 데이터들을 대상으로 하여 계산이 수행되기 때문에 각각의 계산 비용이 데이터 집합의 크기에 비례하여 증가한다.

기존의 그리드 기반 k-평균 군집화 분석 방법이 다양하게 연구되어 왔지만, 이러한 연구들은 자체적인 각각의 군집화 최적화 방법에 대하여 연구했을 뿐, k의 추정 방법이 존재하지 않아 서로 다른 k 에 대한 군집화를 실행해볼 필요가 있다. 따라서, 최선의 군집 결과를 생성하는 k 에 대한 추정 방법에 대한 연구가 필요하다. 본 연구에서는 그리드 기반의 최적인 k 추정 기법을 제안한다.

3. 그리드 기반의 k 추정 기법

본 연구에서 제안하는 그리드 기반의 k 추정 방법은 그리드 생성 및 대푯값 추출, 그리드 기반 군집화, 그리고 k 추정의 총 세 개의 단계로 구성된다. 최적의 k 가 4 일 때 제안 기법의 단계별 작업 과정을 그림 1 에 도식화해 나타냈으며, 각각의 단계마다 수행 과정은 다음과 같다.

그리드 생성 및 대푯값 추출 단계에서는 주어진 데이터를 $m \times m$개의 셀로 구성된 그리드 G 에 분배한다. 데이터가 존재하지 않는 셀은 무시하며, 데이터가 존재하면 존재하는 데이터들의 중심을 각 셀의 대푯값으로 선정한다. 그림 1-a 에서는 이와 같은 그리드 기반 대푯값 선정 결과를 시각화 했으며, 일부 셀의 경우 내부에 값이 없어 무시된 것을 확인할 수 있다.

그리드 기반 근사 군집화 단계에서는 이전 단계에서 계산된 각 대푯값들을 원본 데이터 대신에 활용하여 k-평균 군집화를 수행한다. 그림 1-b 는 대푯값들을 활용한 k-평균 군집화 과정을 간략히 나타낸다. 일반적인 k-평균 군집화와 마찬가지로 임의의 중심점을 기준으로 각 데이터(대푯값)들을 자동으로 k 개의 군집으로 나누는 과정을 수행한다. 이때 k 의 관찰 범위는 원본을 기준으로 관찰하고자 하는 범위와 동일하게 수행한다.

실제 군집화 수행 단계는 생성한 임시 군집 결과들에 대해, 실루엣 계수를 활용하여 각 군집 결과들을 평가하며 최종 k 를 추정하는 단계이며, 다음과 같은 차례로 이루어진다.

실루엣 계수의 극댓값을 보인 k 값들을 모두 후보로 선정한다. 극댓값의 k 값을 후보 k 에 추가한다. 후보 k 값들을 대상으로 실제 군집화를 수행하고 그 결과를 비교, 최선의 군집 결과를 갖는 k 를 최종 추정 k 로 출력한다. 그림 1-c 는 이와 같은 k 후보 선정 및 후보 k 들로부터 최적의 k 도출 과정을 나타낸다.

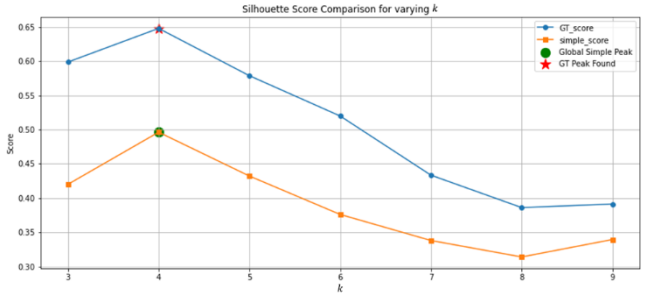
이와 같은 과정들을 통하여 제안 기법에서는 기존 k-평균 군집화 기법들이 가지는 k 추정의 문제를 감소시키는 것이 가능하며, 보다 적은 수의 실제 군집화 과정만으로 효과를 볼 수 있다.

4. 실험 결과 및 분석

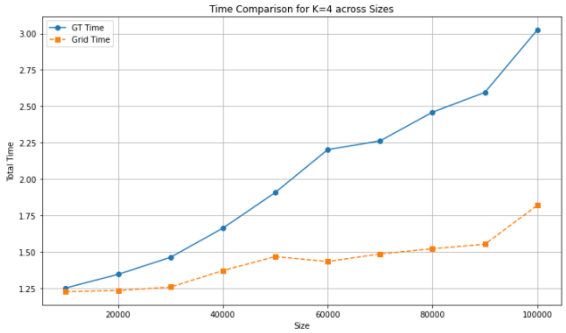
본 절에서는 본 연구에서 수행한 실험 환경과 성능 평가 방법, 실험 결과에 대해 기술한다.

실험 환경: 본 연구에서는 제안 기법의 효과를 보이기 위하여 10,000~100,000 개의 데이터로 구성된 데이터 집합 내에 임의의 군집을 k 개 형성하고, 종래의 k-평균 군집화 분석 기법과 제안 기법을 각각 적용하여 그 결과를 비교 평가하였다. 실험에 활용한 PC 는 intel i7 11700 (2.5GHz), DDR4 24G 로 구성되어 있으며 Python 3.11 환경에서 NumPy 를 활용한 기본적인 계산 프로그래밍 및 scikit-learn 라이브러리의 make_blobs() 함수를 사용한 실험 데이터 생성 및 그리드 구축, 그리고 k-평균 군집화 과정을 구현하였다. (random seed=0) 그리드의 크기를 결정짓는 변수인 m 은 8 로 고정하여 최적의 k 가 3-8 일 경우에 대해 실험들을 수행하였다.

naïve k-평균 군집화 분석을 수행했을 때 계산되는 실제 실루엣 계수를 범위 내의 각 k 에 대해 수집한다. 수집된 실제 실루엣 계수와 그리드의 대푯값들을 사용한 실루엣 계수를 서로 비교하여, 제안 기법이 올바른 최적의 k 를 발견할 수 있는지 검증한다. 또한, 각각의 k 마다 수행 시간을 측정하여 제안 기법에 필요한 총 수행 시간과 naïve 방식 수행시에 필요한 시간에 대하여 비교하는 것으로 제안 기법이 보다 효율적으로 동작함을 보인다.



(그림 2) 최적의 k=4 일 때 실루엣 계수 변화



(그림 3) 데이터 집합의 크기 별 수행시간 비교

그림 2 는 최적의 k 가 4 일 때 그리드 기반 추정과 naïve 기법 각각에 대하여 실루엣 계수 그래프를 시각화한 것이다. 위쪽의 그래프가 GT, 실제 실루엣 계수에 대한 관측 그래프이며 아래의 그래프가 제안 기법을 적용했을 때의 실루엣 계수 그래프이다. 두 그래프의 모습은 전체적으로 유사하며, 그 최댓값의 위치가 동일해 제안 기법의 유효성을 보였다.

수행 시간 분석: 추정을 위해 관찰한 k 들과 후보로 도출된 k 들을 대상으로 한 실제 군집화 수행에 걸린 총 시간을 naïve 상황과 비교한다. 기존 방식의 경우 범위 내의 모든 k 값에 대해 k 평균 군집화를 수행하여야 하므로 전체 수행 시간 합계가 총 수행 시간이 된다. 반면, 제안 기법의 경우엔 그리드 기반 수행 시간의 총합에 추정 실루엣 계수가 최고였던 지점의 k 에 대한 군집화를 실제로 수행하는 시간이 된다.

그림 3 은 달라지는 데이터 집합의 사이즈에 대하여 이와 같이 계산된 총 군집화 수행 시간을 나타낸다. 기존 방식이 대상으로 하는 k 에 따라 수행 시간이 비례하여 빠르게 증가하는 한편, 제안 기법의 수행 시간은 크게 변화하지 않는다. 이는 제안 기법이 일정한 크기의 그리드를 활용하기 때문이다. 따라서, 제안 기법이 데이터 집합의 크기 변화와 무관하게 효율적이며, 특히 그 크기가 커질수록 효과적임을 보였다.

데이터가 10,000 개인 동일 크기의 데이터 집합에 대하여 최적의 k 가 서로 다를 때 각 경우마다의 실루엣 계수 분포를 <표 1>에 나타냈다. 실험을 수행한 최적의 k 가 3 일 때부터 8 일 때까지 실제 최적의 k 위치와 제안 기법으로 추정한 실루엣 계수의 최대치, 즉 최적의 k 위치가 동일함을 확인할 수 있다.

<표 1> 다양한 최적의 k에 대한 실루엣 계수 비교

최적의 k	k	실제_실루엣	제안_실루엣	최적의 k	k	실제_실루엣	제안_실루엣
3	3	**0.653**	**0.500**	6	3	0.557	0.500
	4	0.530	0.413		4	0.555	0.420
	5	0.444	0.372		5	0.606	0.486
	6	0.365	0.312		6	**0.651**	**0.531**
	7	0.384	0.299		7	0.609	0.500
	8	0.409	0.283		8	0.571	0.460
4	3	0.603	0.429	7	3	0.567	0.476
	4	**0.652**	**0.504**		4	0.549	0.397
	5	0.576	0.440		5	0.589	0.461
	6	0.517	0.377		6	0.611	0.505
	7	0.442	0.349		7	**0.627**	**0.548**
	8	0.397	0.292		8	0.593	0.515
5	3	0.623	0.503	8	3	0.604	0.601
	4	0.626	0.534		4	0.605	0.545
	5	**0.677**	**0.595**		5	0.593	0.521
	6	0.625	0.544		6	0.631	0.552
	7	0.564	0.446		7	0.655	0.592
	8	0.549	0.497		8	**0.676**	**0.641**

5. 결론 및 향후 연구

본 연구에서는 그리드를 활용하여 데이터의 지역 대푯값을 찾은 후에 이를 사용한 k-평균 군집화를 위한 k 추정 기법을 제안하였다. 최대 10 만 개 데이터의 다양한 수의 군집이 구성된 상황에서 빠르면서 정확히 k 를 추정할 수 있었으며, 제안 기법을 활용한 추정이 효과적임을 보였다.

향후에는 보다 일반적인 데이터 집합 등 다양한 환경에 대하여 제안 기법을 확장할 것이며, 이를 위하여 제안 기법이 가지는 특징을 분석해 상황에 따라 적절한 m 의 선정 방법 등을 연구할 것이다.

ACKNOWLEDGEMENT

이 논문은 2025 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2022R1I1A3071314)

참고문헌

[1] DU, Xingyu 외, City classification for municipal solid waste prediction in mainland China based on K-means clustering, Waste Management, 144, -, 445-453, 2022.

[2] ZHU, Erzhou 외, Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index, Neurocomputing, 363, -, 149-170, 2019.

[3] 이인식 외, K-평균 군집모형 및 순서형 로짓모형을 이용한 버스 사고 심각도 유형 분석: 측면부 사고를 중심으로, 자동차안전학회지, 13, 3, 69-77, 2021.

[4] 이동환; 임희석, k-평균 군집화 기법을 활용한 SNS의 부적절한 광고성 콘텐츠 탐지, 한국정보처리학회 학술대회논문집, 여수, 2021, 570-573.