미국 LA 범죄 집중지역 분석: 사회경제적 지표를 활용한 데이터 기반 클러스터링 접근법

허재민[01] 이수민[1] 박소정[2] 이어진[1] 최유민[3] 조석헌*
[1]충남대학교 [2]한남대학교 [3]계명대학교 *University of California, San Diego (UCSD)
hjm20720@gmail.com, suminwow1@gmail.com, hisj5674@gmail.com,
win092909297@gmail.com, choeyouminte@gmail.com, *justinshcho@gmail.com

Analysis of Crime Hotspots in Los Angeles: A Data-driven Clustering Approach Using Socioeconomic Indicators

Jaemin Heo[01] Sumin Lee[1] Sojeong Park[2] Eojin Lee[1] Yumin Choi[3] Seokheon Cho*
[1]Chungnam National University [2]Hannam University [3]Keimyung University
*University of California, San Diego (UCSD)

요 약

As crime becomes increasingly sophisticated in modern society, public concern and fear of crime continue to grow. One effective approach to crime reduction involves identifying and analyzing crime hotspots, thereby enabling the development of region-specific prevention strategies. In this study, we applied the K-Means clustering algorithm to analyze crime hotspots effectively across ZIP code-based areas within the city of Los Angeles, California. To support this analysis, we constructed a comprehensive dataset named Crime-related Socioeconomic Indicators (CSI) by integrating and reorganizing datasets collected from multiple sources, including LA City's Open Data portal, the U.S. Census Bureau, and Geohub, all mapped to ZIP code-based regions. From the CSI dataset, six key features were selected through a rigorous evaluation of multicollinearity and their correlation with crime incident counts. To determine the optimal number of clusters for the K-Means model, we emplyed the Elbow method and the mean Silhouette score. The final clustering analysis showed that segmenting Los Angeles into three clusters based on the selected six features most effectively identified distinct crime hotspots. These findings can provide valuable insights for law enforcement agencies in Los Angeles by supporting the formulation of targeted, ZIP code-specific crime prevention strategies.

1. 서 론

현대 사회에 범죄가 지능적으로 발전하고 있어 범죄에 대한 두려움이 증가하고 있다. 2023 년 워싱턴 D.C.에서 Gallup이 실행한 여론 조사 결과 조사 인구의 40%가 범죄에 대한 두려움을 느낀다고 보고했고 이 수치는 30년만에 최고치이다 [1]. 시민들의 치안 불안을 해소시키기 위해서 경찰 기관의 수사 기법 및 프로파일링이 필요하다. 이에 따라 클러스터링 기법을 이용한 범죄 데이터 및 연관 데이터의 범죄 집중지역 분석은 수사 기관에서 지리적인 정보를 토대로 범죄 예방 정책을 펼칠 수 있는 기반이 된다.

J. Bonam *et al.* 는 캐글에서 취득한 UCI Crime and Community 데이터세트를 이용하여 범죄 기록과 경제 요소를 이용해서 범죄 집중지역을 제안하였다 [2]. 최적의 K-means 클러스터링을 이용하여 범죄 집중지역을 제시하였다. 제안한 범죄 집중지역 결과를 라벨링한 후 이를 기계학습 기반 모델들을 사용하여 범죄 집중지역을 분류하였고 준수한 성능의 결과를 보였다. S. Sivaranjani *et al.* 는 National Crime Records Bureau (NCRB) 기관에서 제공하는 인도 6개 도시의 범죄 데이터를 추출하여 K-Means 클러스터링, 병합 군집 분석 및 밀도 기반 클러스터링을 이용해서 범죄 탐지 분석에 대한 연구를 진행하였다 [3]. L. Saroja *et al.* 은 NCRB 에서 수집한 인도 타밀 나두 (Tamil Nadu) 주에서 발생하는 범죄 데이터에 K-Means 클러스터링 알고리즘을 적용하였다. 이를 통해 범죄 집중지역 지역을 클러스터링하였고 그 결과를 지도 위에 도시하였다 [4]. 본 연구에서는 미국 캘리포니아의 LA 도시에 속한 지역에 대해 ZIP 코드 구역 (미국에서 사용하는 우편번호 기반 구역 분할 코드)에 따라 클러스터링 기반 모델을 이용하여 범죄 집중지역을 분석하고자 한다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 범죄 집중지역을 클러스터링하기 위해서 범죄 관련 사회경제적 정보를 포함하는 데이터세트로 생성 과정을 설명한다. 또한, 범죄

* This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) in 2025 (2021-0-01435) & (2024-0-00062).
* Following are results of a study on the "Leaders in Industry-university Cooperation 3.0 "Project, supported by the Ministry of Education and National Research Foundation of Korea.

집중지역 분석을 위해서 최종적으로 사용하는 특성들을 선정하는 과정을 언급한다. 제 3 장에서는 본 연구에서 사용한 K-means Clustering 알고리즘과 최적 군집 수를 찾는 지표를 설명한다. 제 4 장에서는 범죄 집중지역 클러스터링 모델 결과를 비교 및 분석한다. 마지막으로, 제 5 장에서는 본 연구의 결과에 대한 요약과 향후 연구 방향을 제안한다.

2. 범죄 관련 데이터세트 구성 및 데이터 전처리

2.1 범죄 관련 사회경제적 지표 데이터세트 설명

본 연구에서는 LA 도시의 범죄 집중지역을 분석하기 위해 3개의 기관에서 데이터를 수집하였다: LA Open Data에서 제공하는 LA Crime 관련 데이터세트에는 2020년부터 LA Police Department (LAPD) 관할 지역 내 발생한 범죄 사건에 대한 정보 및 LA 시의 가로등 위치 정보를 포함하고 있다 [5, 6]. 발생한 범죄 위치 (위도와 경도) 정보를 ZIP 코드로 변환하였다. 또한, 가로등 위치 정보를 ZIP 코드로 변환하여 ZIP 코드별 가로등 수를 추출하였다. 미국 인구 조사국 (U.S. Census Bureau)에서 제공하는 다수의 사회경제적 지표 관련 데이터세트를 ZIP 코드 단위로 재구성하여 사용한다 [7, 8]. 이 데이터세트에는 인구 수, 빈곤율, 중위소득, 실업률, 교육 수준 및 노숙자 수 관련 정보들을 포함하고 있다. 마지막으로 Geohub에서 제공하는 경찰서 데이터세트를 추가로 고려하였다 [9].

표 1은 위와 같이 수집한 데이터세트들에 포함된 특성 (Feature)들 중에서 범죄 집중지역을 분석하기 위해서 필요한 사회경제적 지표 정보들만을 선정한 Crime-related Socioeconomic Indicators (CSI) 데이터세트를 보여주고 있다. 이는 2022 년의 LAPD 관할하는 지역 내 ZIP 코드를 기준으로 구성된 통합 데이터세트로 ZIP 코드 개수와 동일한 총 134개의 데이터 샘플과 다음과 같은 22개의 특성을 포함하고 있다.

표1. Crime-related Socioeconomic Indicators (CSI) 구성

Feature Name	Type
ZIP CODE, Crime Counts, Streetlight Counts, Total Population, Total Male, Total Female, Under 5 Years, Child, Adolescent, Young Adult, Middle Adult, Older Adult, Poverty Level_50%, Poverty Level_100%, Poverty Level_200%, Median Income, Unemployed People, Low Education, Medium Education, High Education, Homeless People	Integer
Police Station Indicator	String

ZIP 코드별 총 범죄 발생건수 (Crime Counts), 가로등 수 (Streetlight Counts), 총 인구 (Total Population), 성별 인구 (Total Male와 Total Female) 5세 이하 인구 (Under 5 Years), 5 ~ 14세 인구 (Child), 15 ~ 19세 인구 (Adolescent), 20~ 44세 인구 (Young Adult), 45세 ~ 64세 인구 (Middle Adult), 65세 이상 인구 (Older Adult) 빈곤 기준선 이하 인구 (Poverty Level_100%), 기준선의 50 % 이하 인구 (Poverty Level_50%), 기준선의 200 % 이하 인구 (Poverty Level_200%), 가구당 중위 소득 (Median Income, [USD]), 실업 인구 (Unemployed People) 저 학력 인구 (Low Education), 중 학력 인구 (Medium Education), 고 학력 인구 (High Education) 노숙자 수 (Homeless People)이 있다. 마지막으로 ZIP 코드별로 경찰서 존재 유무 (Police Station Indicator) 특성이 존재한다. Police Station Indicator는 경찰서가 없으면 0의 값을 경찰서가 최소 1개 이상 있으면 1의 값을 가지도록 설정하였다.

2.2 다중 공선성 분석 및 최종 특성 선정

CSI 데이터세트에는 범죄 집중지역 분석 모델의 성능을 향상시키기 위해서 특성들 간 다중 공선성 (Multicollinearity)을 분석하였다. 표 2는 CSI 데이터세트의 22개 특성들 중에서 분산 팽창 계수 (Variance Inflation Factor: VIF)가 10이하인 특성들을 나열하였다. 표 2의 선정 (Selection)에 나타낸 것처럼 Crime Counts 특성뿐만 아니라 Crime Counts와 상관관계의 절대값이 0.5 이상인 5개의 특성을 범죄 집중지역 분석을 위해 선정하였다.

표 2. 특성별 분산 팽창 계수 (VIF) 분석 및 특성 선정

Feature Name	VIF	Selection
Total Population	8.45	Yes
Unemployed People	7.92	Yes
Poverty Level_50%	6.51	Yes
Crime Counts	4.96	Yes
Streetlight Counts	3.25	Yes
Median Income	2.03	No
High Education	1.78	No
Homeless People	1.68	Yes
Police Station Indicator	1.22	No

3. 클러스터링 알고리즘 및 성능 평가 지표

3.1 K-means Clustering 알고리즘

K-means Clustering은 비지도 학습으로 각 데이터 샘플들이 특성의 값에 따라 유사한 군집을 이루도록 하는 클러스터링 기법이다. 구현이 간단하고 직관적이라는 장점이 있다.

3.2 최적 클러스터링 군집 수 도출 지표

K-means Clustering은 특성의 값을 통해 구해진 거리를 이용하는 거리 기반 알고리즘이다. Elbow method는 군집 수 (k)에 따른 Within-cluster Sum of Squares (WCSS)의 변화 추이를 통해 최적의 군집 수 (k^*)를 결정하는 기법이다. 일반적으로 k의 값이 증가할수록 WCSS는 감소하는 경향을 보이지만, 특정 k값 다음에 감소하는 폭이 급격히 완화된다. 이 지점을 엘보우 (Elbow)라 부르며 해당 지점의 군집 수를 k^*로 선정한다. WCSS는 다음과 같이 계산된다.

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} dist(x_j^i, c_i) \quad (1)$$

여기에서, k와 n_i는 각각 총 클러스터 개수와 i번째 클러스터에 포함된 총 데이터 샘플 수를 나타낸다. c_i 와 x_j^i는 각각 i번째 클러스터의 중심과 i 번째 클러스터에 포함된 j 번째 데이터 샘플을 의미한다. mean Silhouette score ($\bar{s}$)는 전체 클러스터의 분리도 및 응집도를 평가하는 지표이다. 데이터 샘플 x 의 Silhouette 값인 $s(x)$는 다음과 같이 정의된다.

$$s(x) = \frac{\mathrm{b}(x) - a(x)}{\max\{a(x),\ b(x)\}} \quad (2)$$

여기에서, $a(x)$는 데이터 샘플 x와 같은 군집 내에 속한 다른 데이터 샘플들 간의 평균 거리이고 $b(x)$는 데이터 샘플 x와 데이터 샘플 x가 속하지 않은 군집들에 속한 다른 데이터 샘플들 간의 평균 거리 중에서 가장 작은 값을 의미한다. mean Silhouette score ($\bar{s}$)는 모든 데이터 샘플들의 $s(x)$값의 평균으로 산출된다. $\bar{s}$ 는 1에서 -1 사이의 값을 가지며 값이 1에 가까울수록 모든 데이터 샘플들이 K-means 알고리즘 기반 모델이 제시한 군집들에 평균적으로 잘 할당되었음을 나타낸다.

4. 범죄 집중지역 클러스터링 모델 분석 및 결과

표 2에서 보여준 것처럼 선정한 6개의 특성들의 값들이 범위가 다양하므로 각 특성들의 값을 [0, 1] 범위로 변환한 Min-Max 정규화를 수행한 것과 수행하지 않은 것을 비교하였다.

4.1 Elbow Method 활용 범죄 집중지역 클러스터링 분석

그림 1은 정규화 시도 유무와 클러스터 수에 따른 WCSS값을 보여주고 있다. 정규화를 적용 유무 경우에 있어서 최대 Elbow 지점이 k값이 각각 5와 4로 나타났다. 이 다음의 Elbow 지점은 정규화 적용 유무와 관계없이 k값이 3일 때로 나타났다.

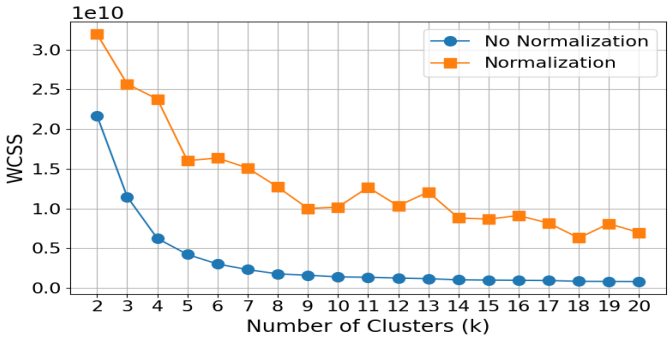

그림 1. K-means Clustering 기반 Elbow Method

4.2 Mean Silhouette Score 활용 클러스터링 분석

그림 2는 정규화 시도 유무와 클러스터 수 (k) 값에 따른 mean Silhouette score 값을 보여주고 있다. 정규화를 적용 유무 경우에 있어서 과적합일 수 있는 k=2일 때를 제외한다면 다음으로 높은 mean Silhouette score를 갖는 k값은 공통적으로 3 인 것을 확인할 수 있다. 이 때의 mean Silhouette score는 정규화를 수행한 경우에는 0.371이지만 정규화를 수행하지 않은 경우에는 0.540이다. 따라서, 본 연구에서는 정규화를 수행하지 않은 경우에 높은 mean Silhouette score를 가지기 때문에, 정규화를 하지 않는 것이 더 효과적이라고 판단을 하였다.

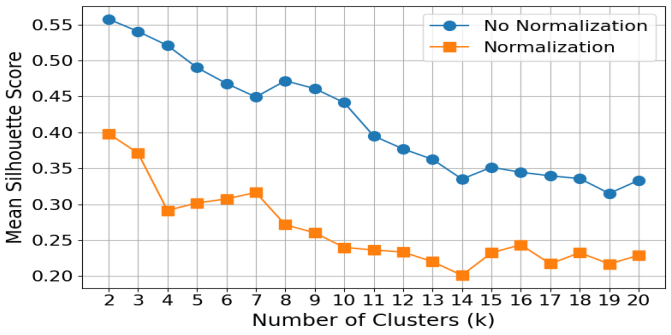

그림 2. K-means Clustering 기반 Mean Silhouette Score

4.3 최적 클러스터 반영 범죄 집중지역 클러스터링 분석

K-means Clustering 기반 모델의 최적의 클러스터를 도출하기 위해 사용한 Elbow method와 mean Silhouette score 결과들을 함께 고려하였을 때 최적의 클러스터값은 3으로 나타났다. 그림 3 은 K-means Clustering 기반 모델의 최적 성능을 보이는 클러스터링 결과를 그림으로 도시한 것이다. 즉, 최적 클러스터 개수가 3이고 표 2에 나열하였던 최종적으로 선택한 6개의 특성들에 대하여 정규화를 수행하지 않은 경우에 있어서 클러스터링 결과인 것이다. 여기에서, 각 클러스터의 Centroid 정보를 분석한 결과 Cluster 0에 속하는 ZIP 코드 지역들이 가장 높은 범죄 집중지역을 나타내며 Cluster 2번에 속하는 지역들이 가장 낮은 범죄 집중지역을 보인다.

5. 결론

본 연구에서는 LA 도시의 범죄 예방에 대한 분석을 위해 ZIP 코드 구역별 클러스터링 기반 모델을 개발하여 범죄 집중지역을 분석하였다. 범죄 집중지역을 클러스터링하기 위한 알고리즘으로 K-Means Clustering을 사용하였다. 또한, LA 시의 Open Data, 미국 인구 조사국 및 Geohub 등에서 수집한 데이터세트들에 대하여 ZIP 코드 구역별로 재구성 및 전처리 과정을 수행한 후 Crime-related Socioeconomic Indicators (CSI) 라는 새로운 데이터세트를 생성하였다. 또한, 분석의 정확성을 위해 CSI 데이터세트에 포함된 특성들 간의 다중 공선성 (Multicollinearity) 분석과 범죄 발생건수와의 상관계수 값을 고려하여 특성 6개를 최종적으로 선정하였다. 이 6개의 특성들을 고려하는 K-Means Clustering의 최적의 군집 수를 도출하기 위해 Elbow method와 mean Silhouette score를 중요한 지표로 사용하였다. 그 결과 두 지표 모두에서 최적 클러스터 개수가 3일 때 우수한 성능을 보였다. 또한, 특성 6개에 대해서 정규화를 수행하지 않았을 때 mean Silhouette score가 0.540로 높았다. 이 결과들을 바탕으로 LA 범죄 집중지역 클러스터링 결과를 지도위에 나타내었으며 이는 LA시의 범죄 관련 정책을 세우는데 도움을 될 것으로 예상한다.

향후 연구에서는 다양한 클러스터링 기반 모델들을 고려하여 범죄 유형에 따른 범죄 집중지역을 분석하고자 한다.

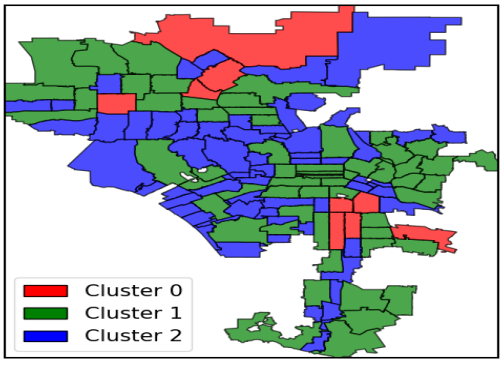

그림 3. K-means Clustering 기반 LA 범죄 집중지역 모델 결과

참 고 문 헌

[1] L. Saad, “Personal Safety Fears at Three-Decade High in U.S.,” GALLUP, Nov. 2023.

[2] J. Bonam, L. R. Burra, G. S. Susheel, K. Narendra, M. Sandeep, and G. Nagamani, “Crime Hotspot Detection using Optimized K-means Clustering and Machine Learning Techniques,” ICESE, pp. 787-792, Jul. 2023.

[3] S. Sivaranjani, S. Sivakumari, M. Aasha, and K. Narendra, “Crime Prediction and Forecasting in Tamilnadu Using Clustering Approach,” ICETT, pp. 1-6, Oct. 2016.

[4] L. S. Thota, M. Alayan, A. A. Khalid, F. Fathima, S. B. Changalasetty, and M. Shiblee, “Cluster Based Zoning of Crime Info,” ICACC, pp. 87-92, Mar. 2017.

[5] City of Los Angeles Open Data, “Crime Data from 2020 to Present,” Available: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8.

[6] City of Los Angeles Open Data, “Streetlight Locations,” Available: https://data.lacity.org/City-Infrastructure-Service-Requests/Streetlight-locations/9ei6-svt8.

[7] US Census Open Data, “2022: ACS 5-Year Estimates Data Profiles,” Dec. 2023, Available: https://data.census.gov/table?q=DP05&g=040XX00US06$8600000&y=2022.

[8] Los Angeles Homeless Services Authority, “2022 Greater Los Angeles Homeless Count Data,” Oct. 2022.

[9] P. Jula, “LAPD Police Stations,” Feb. 2018, Available: https://geohub.lacity.org/datasets/1dd3271db7bd44f28285041058ac4612_0/about?layer=0.