2-Dimensional Clustering of Crime Dataset

Geetheswar Reddy Baddela, Pranav Sunil, Nishanth Gajjala
CSE-572: Data Mining Project Report
Arizona State University, Tempe, Arizona
{gbaddela, psunil, ngajjal4}@asu.edu

***Abstract*—In the field of bioinformatics and molecular biology, the accurate classification of proteins based on their sequences plays a crucial role in understanding their functions and interactions. This research paper presents a novel deep learning architecture, named HybridSeqNet, that leverages the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) to enhance the accuracy of protein structure classification from sequence data. The proposed HybridSeqNet architecture capitalizes on the hierarchical features captured by CNNs in local regions of protein structure sequences and the temporal dependencies captured by LSTMs in longer-range contexts. This joint architecture synergistically integrates these two neural network paradigms, enabling the model to effectively learn intricate patterns within protein sequences. The dataset used contained more than 4,00000 protein structure sequences. Various types of protein structures like DNA, RNA, hybrid DNA-RNA, etc. are considered for classification. We also compare our model with various existing models. HybridSeqNet achieves remarkable classification accuracy of 92.86% when compared to traditional models. The architecture's effectiveness is demonstrated through comprehensive experiments on a huge dataset of protein structure sequences.**

Keywords: Crime Analytics, Los Angeles Crime Data, Clustering, Classification

I. Introduction

Activities of crime are a facet of society that has turbulent implications that are punishable by the law. Such activities not only affect an individual but have large-scale implications for society as a whole. The crime recordings in urban demographics like Los Angeles show important safety and socio-economic insights. In spite of the abundance of public crime data, it can be a difficult subject for analysis as its variety and size make it difficult for effective interpretation. However, analyzing previous criminal data could help individuals be more aware of the crimes occurring in and around their Neighborhood and navigate their day-to-day lives being more informed. This can also extend to lawmakers and public servants tasked with keeping the Neighborhood safe and providing more insights into prevention planning, resource allocation, and protocols for situations that may arise in regions with high crime risk.

Current studies by authors like Ubon Thongsatapornwatana [2] show us that data mining has emerged as a powerful analytical technique for examining crime data stored from various sources to uncover patterns and trends. They have also been able to prove its effectiveness in enhancing the speed and accuracy of crime-solving efforts, while also facilitating automated crime notifications. With a wide range of data mining techniques available, selecting the most appropriate method is critical to improving the efficiency of crime detection. This survey paper provided a comprehensive review of the literature on data mining applications, focusing specifically on their role in addressing crime-related challenges. It highlighted to us the existing research gaps and challenges in the field of crime data mining and offered valuable insights into leveraging data mining techniques for identifying crime patterns and trends. This study serves as a resource for initial exploration of data mining in crime analysis. Further work done by S. Sathyadevan, et al. [3] reviews the application of data mining in crime analysis, emphasizing its potential to uncover patterns and trends from extensive crime datasets. It highlights the diverse techniques employed, such as association rule mining, clustering, and classification, to enhance the efficiency of crime detection and prevention. The survey identifies key challenges like data integration, visualization, and the dynamic nature of crime patterns while offering insights into addressing these issues. We work towards addressing some of the underlying issues of crime location. In the paper by Tong et al. [4], they focussed on identifying which crimes are committed by the same individual or group, which is an essential task for crime analysts, as it aids in prevention and investigation. They introduced Series Finder, a novel algorithm designed to identify patterns of crimes by growing a sequence from an initial "seed" of related incidents. By incorporating both general characteristics and unique aspects of each pattern, Series Finder offers a dynamic approach that adapts to shifts in criminal behavior. These findings have been important in Crime Prevention and investigation. Yadav and Kumari (2018) [5] explore the use of clustering techniques to analyze criminal behavior and uncover patterns in crime data. Their study demonstrates how grouping similar criminal activities based on shared characteristics can help identify behavioral trends and improve crime prevention strategies. By leveraging computational approaches, the authors emphasize the potential of clustering to support law enforcement in understanding and addressing criminal patterns efficiently. Agarwal et al. (2013) [6] investigate the application of the K-Means clustering algorithm for crime analysis. The study highlights how clustering can be used to group similar crimes based on attributes such as location, time, and type, enabling the identification of hotspots and patterns. The authors demonstrate the effectiveness of K-Means in simplifying crime data and aiding decision-making for law enforcement agencies. Hajela et al. (2020) [7] propose a clustering-based approach to identify crime hotspots for predictive analysis. By leveraging clustering techniques, the study aims to detect areas with high crime intensity, enabling law enforcement to

allocate resources more effectively. The authors emphasize the role of such methods in enhancing crime prediction and prevention strategies through data-driven insights. Tian (2018) [8] explores the application of cluster analysis for detecting criminal communities. The study demonstrates how clustering techniques can group individuals or activities based on shared characteristics, facilitating the identification of underlying networks within criminal organizations. The research highlights the effectiveness of cluster analysis in uncovering hidden patterns and supporting law enforcement in community detection efforts.

Although these existing methods work extensively towards predicting crime, and its investigation as a process. However, in our proposed approach, we look to introduce clustering on a two-dimensional scale with a broader superclass of different types of crime, aiding in clustering and classification. This is in contrast to the previous methods and work done to predict crime by providing insights into the prevalent crimes in the neighborhood, allowing citizens to be aware of their surroundings and build caution. To understand this, our project uses a comprehensive dataset 'Crime Data from 2020 to Present," sourced from data.gov, to interpret and analyze the crime reports with a special emphasis on location. With this project, we aim to provide a better understanding and insight into this prevalent issue by understanding the distribution of crime in urban settings like the city of Los Angeles.

The project was primarily run on the Python 3 Google Compute Engine backend (TPU); RAM: 120.77 GB/334.56 GB; Disk: 17.44 GB/225.33 GB after our final iteration of the code. It was also run on a local system with Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz, 2592 Mhz, 6 Core(s), 12 Logical Processor(s), OS: Microsoft Windows 11 Home, RAM: 32GB, GPU: NVIDIA RTX 2070 MaxQ.

II. Data Exploration and Problem Statement

The dataset 'Crime Data from 2020 to Present' used for the model was retrieved from data.gov, an official open-source website of the United States government. It contains close to 850,000 entries with 28 total features of crime from 2020 till early Jan 2024. This data is transcribed from original crime reports that are typed on paper. This data is as accurate as the data in the LAPD database during that span. This is a real-world dataset that gives information on crime reports in the City of Los Angeles dating back to 2020. The data includes features like date reported, unique crime id, date occurred, time occurred, area code, area name, rd, crime code, crime description, status, status description, location, cross street, latitude and longitude. The dataset obtained from data.gov contained many erroneous data, which required data cleaning and pre-processing. Among these features, we have extensively used time occurred, crime description, latitude, and longitude for our analysis. However, insights into other parts of the data are in situ for our analysis outside the model.

In figure 1, we see the distribution of crime in correspondence to area

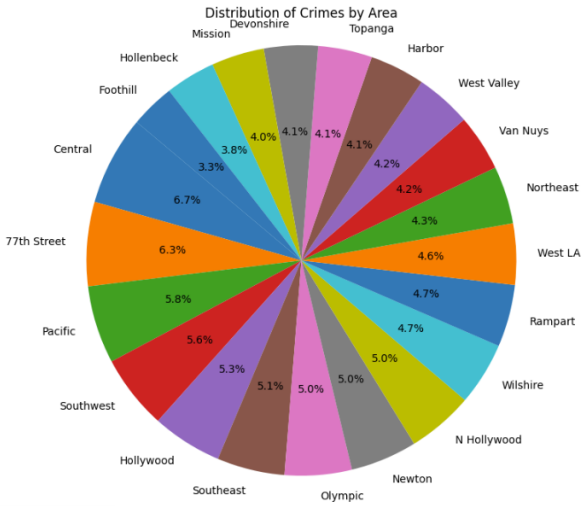

Fig. 1: Distribution of Crime by area

From this, we see that areas like "Central" and "77th Street" areas exhibit the highest proportions, with 6.7% and 6.3% of total crimes, respectively, indicating these regions may have higher crime rates compared to others. Conversely, areas like "Foothill" and "Mission" report lower proportions, suggesting these are relatively safer regions.

Figure 2. shows the distribution of crime in the form of a heatmap of the LA Crime dataset.

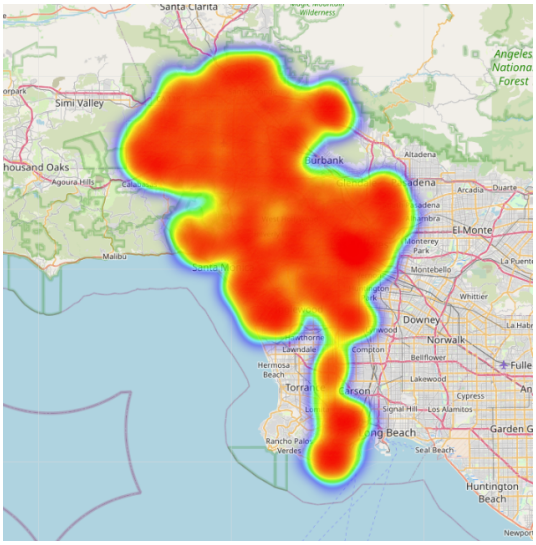

Fig. 2: Heatmap of distribution of crimes

The intensity of the red and yellow regions indicates higher crime densities, while blue and green areas represent relatively lower densities. From the visualization, central and southern parts of the city show significant clustering of crimes, which aligns with the findings in the pie chart.

In figure 3, we see a kernel density estimation (KDE) plot illustrating the distribution of victims' ages. The curve peaks around the 20–30 age range, indicating that most victims fall within this demographic. The density gradually decreases for older age groups but remains notable up to the 60s, suggesting crimes affect a wide age range.

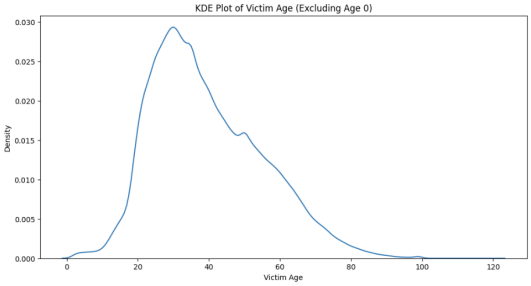

Fig. 3: Heatmap of distribution of crimes

III. Proposed Method

In this section, we describe the proposed two-dimensional clustering method. This method involves clustering crime data first based on geographic locations and subsequently refining the clusters based on crime types. The approach leverages the spatial and categorical features of the dataset, providing a consolidated view of crime patterns.

We begin with the dataset $D = \{x_1, x_2, \ldots, x_n\}$, where each data point x_i contains latitude and longitude coordinates $(\text{LAT}_i, \text{LON}_i)$, a detailed crime description Desc_i, and other auxiliary attributes.

At first we cluster based on location(latitude and longitude respectively). Before moving into clustering we normalize latitude and longitude as follows:

$$x_i' = \frac{x_i - \mu}{\sigma}, \tag{1}$$

where μ and σ are the mean and standard deviation of the coordinates, respectively. We then apply K-Means clustering to partition the crimes into K location clusters C_k by minimizing the within-cluster sum of squared distances:

$$\text{minimize} \quad \sum_{k=1}^{K} \sum_{x \in C_k} \|x - \mu_k\|^2, \tag{2}$$

where μ_k is the centroid of cluster C_k, given by:

$$\mu_k = \frac{1}{|C_k|} \sum_{x \in C_k} x. \tag{3}$$

Each crime x_i is assigned to a specific location cluster C_k, such that:

$$\text{Location\_Cluster}_i = k, \quad \text{if } x_i \in C_k. \tag{4}$$

Once the location clusters are established, we proceed to analyze the distribution of crime categories within each cluster. This is represented by a distribution matrix M, where each element $m_{k,c}$ denotes the count of crimes of category c in location cluster C_k. Formally, this is defined as:

$$m_{k,c} = \sum_{i=1}^{n} 1(\text{Location\_Cluster}_i = k \wedge f(\text{Desc}_i) = c), \tag{5}$$

where $1(\cdot)$ is the indicator function, returning 1 if the condition is true and 0 otherwise. To account for varying cluster sizes, the matrix is row-normalized such that:

$$m'_{k,c} = \frac{m_{k,c}}{\sum_{c \in \mathcal{C}} m_{k,c}}, \quad \text{for each } k \text{ and } c. \tag{6}$$

The normalized distribution matrix M' serves as input for the second clustering stage. We apply L-Means clustering to group the location clusters into L crime clusters L_l based on their crime distributions:

$$\text{minimize} \quad \sum_{l=1}^{L} \sum_{M'_k \in L_l} \|M'_k - \nu_l\|^2, \tag{7}$$

where ν_l represents the centroid of cluster L_l. This step groups location clusters with similar crime type distributions, yielding crime clusters L_l.

Finally, each crime x_i is associated with both a location cluster and a crime cluster:

$$x_i \to (\text{Location\_Cluster}_i, \text{Crime\_Cluster}_{\text{Location\_Cluster}_i}). \tag{8}$$

This dual-clustering approach provides a consolidated framework for analyzing crimes spatially and categorically.

The results are visualized on a map using latitude and longitude coordinates. Each marker is color-coded by its crime cluster, with popup information displaying the location cluster, crime cluster, and crime category. This visualization aids in understanding crime patterns and distributions across different regions.

Utilizing this cluster information, we are not only able to provide information on the crime present in the neighborhood but also provide a real use case for people to find the path with the least crime risk and density between their start and destination.

To find an optimal path between the start and the end destination we propose a method to calculate an optimal route between two locations based on crime risk and density. The approach involves constructing a weighted graph, where the weights are derived from a combination of spatial distance, crime risk, and crime density. The shortest path in this graph minimizes these combined weights, resulting in a safer route.

A. Graph Construction

Let $G = (V, E)$ be a graph, where:

- V: The set of nodes, each representing a geographic point $(\text{LAT}_i, \text{LON}_i)$.
- E: The set of edges connecting nearby nodes based on geographic proximity.

The weight of an edge $w(i, j)$ between nodes i and j is defined as:

$$w(i, j) = d(i, j) + r(i) + d_{\text{crime}}(i), \tag{9}$$

where:

- $d(i, j)$: The Euclidean distance between nodes i and j.
- $r(i)$: The normalized crime risk score for node i, derived from the location cluster.
- $d_{\text{crime}}(i)$: The crime density for node i, representing the number of crimes in its location cluster.

The crime risk score $r(i)$ is normalized as:

$$r(i) = \frac{\text{Crime Count in Cluster}(i)}{\text{Maximum Crime Count Across Clusters}}. \tag{10}$$

Similarly, the crime density $d_{\text{crime}}(i)$ is calculated as:

$$d_{\text{crime}}(i) = \sum_{x \in \text{Cluster}(i)} 1, \tag{11}$$

where the sum counts all crimes in the cluster corresponding to node i.

The Nearest Neighbors algorithm is used to connect each node to its closest k neighbors, forming the edge set E.

B. Path Calculation

To find the safest route between two points s (start) and t (end), we compute the shortest path in the graph G using Dijkstra's algorithm. The algorithm minimizes the cumulative edge weights:

$$\text{minimize} \sum_{(i,j) \in P} w(i,j), \tag{12}$$

where P is the path from s to t.

Once the path is calculated it is visualized on a map, with markers indicating the start and end points, as well as the intermediate nodes along the route. The path is displayed on the map as a polyline, where:

- Nodes are color-coded based on their crime clusters.
- The start and end points are highlighted with distinct markers.

This approach ensures that the calculated route not only minimizes geographic distance but also avoids areas with high crime risk and density, providing a safer travel option. Now we look into the experimental results of our model where we compare our approach with various other baseline models.

IV. Experimental Results

We predict a new crime into the four superclasses of Crime: theft, Violent, Misdemeanor, and Miscellaneous. We use Area, Victim age, Victim Sex, Victim Descent, Premises Description, Weapon Description, Latitude, and Longitudinal data to classify the crime into the four following superclasses. Table 1 shows the model comparison using various models to compare the Accuracy, Precision, Recall and F-1 Score.

TABLE I: Comparison of metrics

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes (Baseline)	0.74	0.63	0.74	0.68
XGBoost	0.77	0.68	0.77	0.70
Gradient Boosting	**0.81**	**0.79**	**0.81**	0.77
Random Forest	**0.81**	0.78	**0.81**	**0.78**

From this, we can see that the F-1 score of Random Forest is marginally better than Gradient Boosting and outperforms our baseline model Naive Bayes.

Evaluation Metrics

The classification models were evaluated using the following metrics:

- **Accuracy:** The ratio of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{13}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{14}$$

- **Recall:** The ratio of correctly predicted positive observations to all actual positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

- **F1-Score:** The harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

Post this, we measure the metrics for clustering using three models, namely, K-Means, HDBScan, Deep Embedded Clustering(DEC) using K-Means.

TABLE II: Comparison of metrics

Model	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
HDBScan (Baseline)	0.4135	255.11	1.6637
K-Means	0.4595	65397.90	0.7235
DEC	**0.6604**	**287290.73**	**0.4542**

We can see that DEC consistently outperforms the other methods across all evaluation metrics. DEC achieves the highest Silhouette Score (0.6604) and Calinski-Harabasz Index (287290.73), indicating superior cluster cohesion and separation, along with the lowest Davies-Bouldin Index (0.4542), reflecting compact and well-separated clusters.

A. Clustering Evaluation Metrics

To evaluate the performance of the clustering models, we utilize the following metrics: *Silhouette Score*, *Calinski-Harabasz Index*, and *Davies-Bouldin Index*. These metrics provide insights into the quality of the clusters by measuring cohesion, separation, and overall cluster compactness.

1) Silhouette Score: The Silhouette Score measures how well each data point is clustered by comparing its distance to other points within the same cluster (cohesion) and to points in the nearest cluster (separation). For a data point i, the score is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{17}$$

where:

- $a(i)$: The average distance of i to all other points in the same cluster.
- $b(i)$: The average distance of i to all points in the nearest cluster.

The overall Silhouette Score for the dataset is the mean of $s(i)$ across all data points. The score ranges from -1 to 1, where higher values indicate better clustering.

2) Calinski-Harabasz Index: The Calinski-Harabasz Index evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. It is defined as:

$$CH = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \cdot \frac{n-k}{k-1}, \tag{18}$$

where:

- n: Total number of data points.
- k: Number of clusters.
- $\text{Tr}(B_k)$: Trace of the between-cluster dispersion matrix.
- $\text{Tr}(W_k)$: Trace of the within-cluster dispersion matrix.

Higher values of the Calinski-Harabasz Index indicate better-defined clusters.

3) Davies-Bouldin Index: The Davies-Bouldin Index measures the average similarity between each cluster and its most similar cluster. It is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right), \tag{19}$$

where:

- k: Number of clusters.
- σ_i: Average distance between points in cluster i and the cluster centroid c_i.
- $d(c_i, c_j)$: Distance between centroids c_i and c_j.

Lower values of the Davies-Bouldin Index indicate better clustering, as clusters are more compact and well-separated.

V. Implementation

VI. Conclusion

In this comprehensive research endeavor, we have delved into the complex field of macromolecule-type classification, focusing extensively on structural protein sequences. By harnessing the combined strength of a modern Hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model, we achieved results that demonstrate the efficacy of current deep learning techniques in bioinformatics. Our approach was grounded on a rigorously prepared dataset, which allowed us to fine-tune and validate our model effectively. The outcomes have been remarkable, with the model showing high precision and recall rates in distinguishing different structural protein types, promising accuracy and reliability in predictions, and facilitating a range of future applications. However, we acknowledge that there is room for further improvements and refinements to enhance the model's performance. The exploration of different architectural configurations and strategies to address class imbalances presents promising directions for future work. In conclusion, our HybridSeqNet project stands as a pivotal tool in structural protein type classification, hinting at the immense potential of AI to redefine our understanding of the biological realm. We move forward with enthusiasm, ready to build upon this foundation and foster further scientific and medical advancements.

References

[1] Cung, B. (2013). Crime and Demographics: An Analysis of LAPD Crime Data. UCLA. ProQuest ID: Cung_ucla_0031N_11420. Merritt ID: ark:/13030/m52b99n3. Retrieved from https://escholarship.org/uc/item/2v76v571

[2] Thongsatapornwatana, Ubon (2016). A survey of data mining techniques for analyzing crime patterns. 2016 Second Asian Conference on Defence Technology (ACDT) DOI: 10.1109/ACDT.2016.7437655

[3] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," in Networks Soft storage(ICNSC), 2019 First International Conference on, Aug 2019, pp. 406–412.

[4] Wang, Tong, Cynthia Rudin, Dan Wagner, and Rich Sevieri. "Learning to Detect Patterns of Crime." European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2013, Prague, 23-27 September 2013. Version: Author's final manuscript

[5] Yadav, Romika Kumari, Savita. (2018). Analysis of Criminal Behavior through Clustering Approach. International Journal of Computer Sciences and Engineering. 6. 341-344. 10.26438/ijcse/v6i11.341344.

[6] Agarwal, Jyoti Nagpal, Renuka Sehgal, Rajni. (2013). Crime Analysis using K-Means Clustering. International Journal of Computer Applications. 83. 1-4. 10.5120/14433-2579.

[7] Hajela, Gaurav et al. "A Clustering Based Hotspot Identification Approach For Crime Prediction." Procedia Computer Science (2020): n. pag.

[8] Junjing, Tian. (2018). Research on the Application of Cluster Analysis in Criminal Community Detection. Journal of Physics: Conference Series. 1060. 012052. 10.1088/1742-6596/1060/1/012052.