# CRIME PREDICTION USING CLUSTERING ALGORITHMS

V.Yamuna[1], Prof.B.Prajna[2], 1, 2 Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Andhra University, AP, INDIA,

***Abstract*—Analyzing the data is prominent in any field. Analyzing data helps us to produce substantial outputs. Most companies already collect and refine massive quantities of data. So using data analysis we integrate to real time governance to streamline crime rates. The main Motivation of this project is data analysis can preliminary do Prediction based on the early data captured. The captured data is structured into classification and regression; classification will give discrete values, while regression is used to predict ordered values. Using these biases we implement in real time governance.**

**In the proposed system we will collect the data of crimes and is put forward for analysis , on the obtained results we classify the city in the form clusters using clustering algorithm which is sub categorizing a clusters on a point. So we use crime data and perform analysis on type of crimes we separate the areas in the city as clusters. So this will help the cops to know the crime rate in specific areas. Basing on early analysis of crime in a certain time period we will predict the occurrence of crime so that cops can patrol an area to reduce the crime area.**

***Index Terms*—Clustering Analysis, K-Means Clustering, Agglomerative Clustering, Mean Shift, PAM, Internal Validation.**

## I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. As these data mining methods are almost always computationally intensive. We use data mining tools, methodologies, and theories for revealing patterns in data.

Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity) [14]

The K mean algorithm was first projected by Stuart Lloyd, as a technique for pulse-code modulation in 1957. It is a classical and well known clustering algorithm. It is the most commonly used partitioned clustering algorithm because it can be easily implemented. It is efficient in terms of the execution time. Its time complexity is O (tKn) where n data point numbers, K is the cluster number and t is the iteration number. It is used to partition data points into discoverable K (non-overlapping) clusters by finding K centroids or center points and then assigning each point to the cluster associated with its nearest centroid [10].

The Hierarchical clustering algorithm (HCA) is also called as connectivity based clustering, which is mainly based on the core idea of objects that are being more relative to the nearby objects than to the objects far away. It is a method of cluster analysis which seeks to build a hierarchy of clusters. Its result is usually presented in a dendrogram. It is generally classified as Agglomerative and Divisive methods that depended upon how the hierarchies are formed.

- **Agglomerative**: It is a "bottom up" approach. It starts by placing each object in its own cluster. Then merges these minute clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Its complexity is O ($n^3$) which makes then too slow for large data sets.
- **Divisive**: It is a "top down" approach. It starting with all objects in one cluster. Then splits are performed recursively as one move down the hierarchy. Its complexity is O (2n) which is worse. These algorithms join the objects and form clusters by measuring their distance. These algorithms cannot provide a particular partitioning in the dataset, but they provide a widespread hierarchy of clusters that are merged with each other at accurate distance [16].

The PAM algorithm is very similar to K-means, mostly because both are partitioned algorithms, in other words, both break the dataset into groups (clusters), and both work by trying to minimize the error, but PAM works with Medoids, that are an entity of the dataset that represent the group in which it is inserted, and K-means works with Centroids, that are artificially created entity that represent its cluster. The PAM algorithm partitions the dataset of n objects into k clusters, where both the dataset and the number k is an input of the algorithm. This algorithm works with a matrix of dissimilarity, whose goal is to minimize the overall dissimilarity between the represents of each cluster and its members [12].

Mean shift is a non-parametric feature-space analysis technique for locating the maxima of a density function, a so-called mode-seeking algorithm. Application domains include cluster analysis in computer vision and image processing Use those LATEX files for formatting, but please follow the instructions in this template as they are specific to IEEE TMI [1].

### A. *Challenges in Data mining*

Define tremendous amount of data – Algorithms must be highly scalable to handle such as tera-bytes of data.

- High-dimensionality of data – Micro-array may have tens of thousands of dimensions.
- High complexity of data – Noisy and unreliable – Dynamically evolving – High dimensionality – Multiple heterogeneous sources [3].

## II. Literature Review

Clustering is considered as an unsupervised classification process [4]. A large number of clustering algorithms have been developed for different purposes [4-6]. Clustering techniques can be categorized into partitioning clustering, Hierarchical clustering, Density based methods, Grid-based methods and Model based clustering methods. Partitioning clustering algorithms, such as K-means, K-Medoids PAM, assign objects into k (predefined cluster number) clusters, and iteratively reallocate objects to improve the quality of clustering results. Hierarchical clustering algorithms assign objects in tree structured clusters, i.e., a cluster can have data point's representatives of low level clusters [8]. The idea of Density-based clustering methods is that for each point of a cluster the neighborhood of a given unit distance contains at least a minimum number of points, i.e. the density in the neighborhood should reach some threshold. The idea of the density-based clustering algorithm is that, for each point of a cluster, the neighborhood of a given unit distance has to contain at least a minimum number of points [8].

Literature survey is the basic step in preparing the new methodology for the particular area of subject. Many researchers have been made their work on the latest advancements in the technology for the improvements in the conversion of unstructured data into structured format. This conversion will help the user to take decisions by applying the queries on the structured data. We survey clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts [1].

The results of an experimental study of some common document clustering techniques [19]: agglomerative hierarchical clustering, K-means, Mean shift, Pam. Our results indicate that the mean shift technique is better than the standard K-means approach and hierarchical approaches we tested [2][11] .

We have provided a brief introduction of K-Means algorithm analysis. As we know that clustering is a process for discovering groups and identifying interesting patterns. Data mining refers to extracting knowledge from large database. Today retrieving information from large dataset is very typical task. That's why we need data mining techniques for managing huge dataset [3][17][18].

This paper is referred as the knowledge discovery from data (KDD). It focuses on the feasibility, usefulness, effectiveness, and scalability of techniques of large data sets. After describing data mining, this edition explains the methods of knowing, preprocessing, processing, and warehousing data. It then presents information about data warehouses, online analytical processing (OLAP), and data cube technology [4].

This research essay is aimed to provide a comprehensive overview of predictive analytics and its real world applications in various fields. Predictive Analytics modeling is used in representation of real world situations for rendering or description of reality [5].

Data mining has its origins in various disciplines, of which the two most important are statistics and machine learning. Statistics has its roots in mathematics; therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice. In contrast, the machine - learning community has its origins very much in computer practice [6].

## III. Problem Description

To control crime inside a city is not easy task for the cops due to their adequate human resource and their workflow. So we assist them to patrol citizen on ease of real time governance to reduce the crime rate. Our goal is to analyze the location of a crime and stopped people by using different clustering algorithms were studies and applied to the data sets for analyzing the location of the crime and stopped people in order to reduce city crime rate.

## IV. Design and Methodology

The following figure 4.1 shows Data Mining Architecture containing six components. That is a data source, data warehouse server, data mining engine, and knowledge base. We can say it is a process of extracting interesting knowledge from large amounts of data. That is stored in many data sources. Such as file systems, databases, data warehouses. Also, knowledge used to contributes a lot of benefits to business and individual [6].

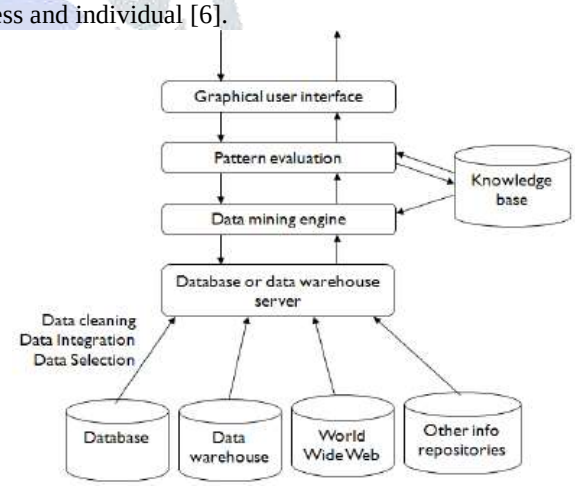



Figure: 1 Data Mining Architecture

### A. *Major Components of Data Mining*

- **Data Sources:** There are so many documents present. That is a database, data warehouse, World Wide Web (WWW). That is the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.
- **Database or Data Warehouse Server**: The database server contains the actual data that is ready to be processed. Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.
- **Data Mining Engine:** In data mining system data mining engine is the core component. As it consists a number of modules. That we used to perform data mining tasks. That

includes association, classification, characterization, clustering, prediction, etc.

- **Pattern Evaluation Modules:** This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value. Also, it interacts with the data mining engine. That's main focus is to search towards interesting patterns.
- **Graphical User Interface:** We use this interface to communicate between the user and the data mining system. Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process. When the user specifies a query, this module interacts with the data mining system. Thus, displays the result in an easily understandable manner.
- **Knowledge Base:** In whole data mining process, the knowledge base is beneficial. We use it to guiding the search for the result patterns. The knowledge base might even contain user beliefs and data from user experiences.

R is a system for statistical analyses and graphics created by Ross Ihaka and Robert Gentleman. R is both software and a language considered as a dialect of the S language created by AT&T Bell Laboratories. S is available as the software S-PLUS commercialized by Insightful.

R is freely distributed under the terms of the GNU General Public License; its development and distribution are carried out by several statisticians known as the R Development Core Team [7].

## V. RESULTS

K-MEANS:

We will be clustered to find out the most frequent locations for this crime. The plot for the location of identifying crimes is shown in Figure, we could see that our dataset not well separated or defined which making it hard to cluster.

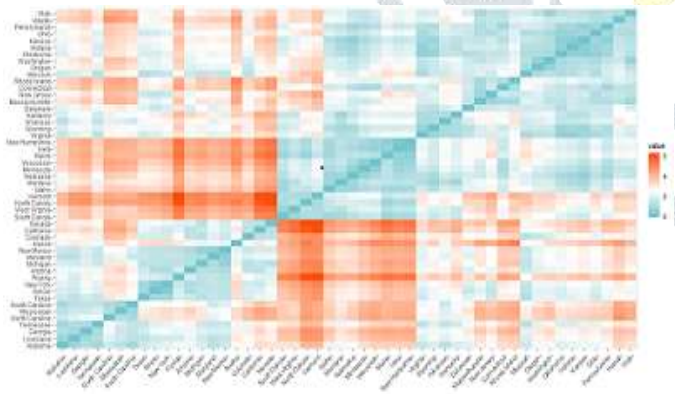
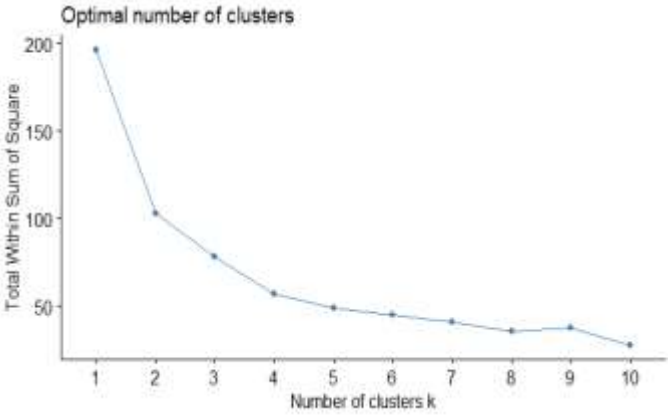


Figure2: Plot K-Means

Now let us apply *K*-Means clustering algorithm. Before doing that, we must determine what the best number of clusters is**.** We used the Sum of squared error to determine the optimal number of clusters.




We can see and notice the suitable number of cluster is 4. *K*-means algorithm clustered and divided the location in plot graph with 4 colors (i.e., areas). Each area indicates to different cluster. As we can see below the map contains all location for doing murder. We circled different clusters with different colors (i.e., areas).

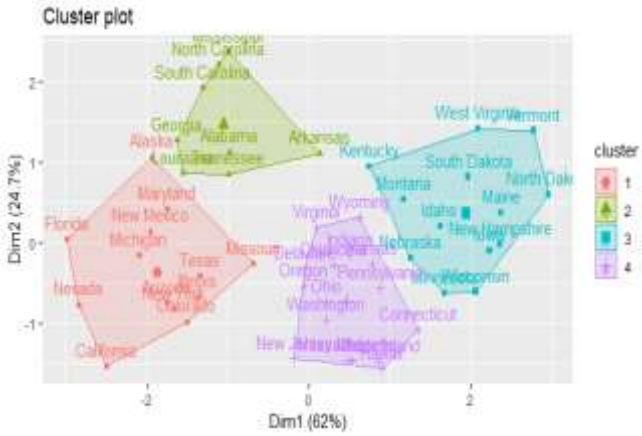



Figure 3: A map contains all locations for doing murder

HIERARCHICAL ALGORITHM:

We clustered the murder crime location using the Group-average clustering for agglomerative clustering. We also applied the Complete-link clustering and Single-link clustering but we found that the Group-average clustering gives the best result among three methods.

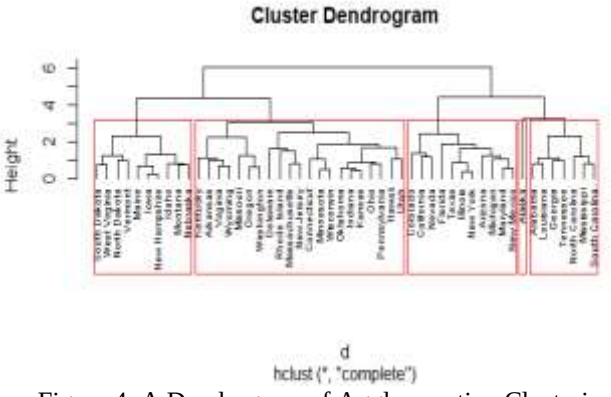



Figure 4: A Dendrogram of Agglomerative Clustering

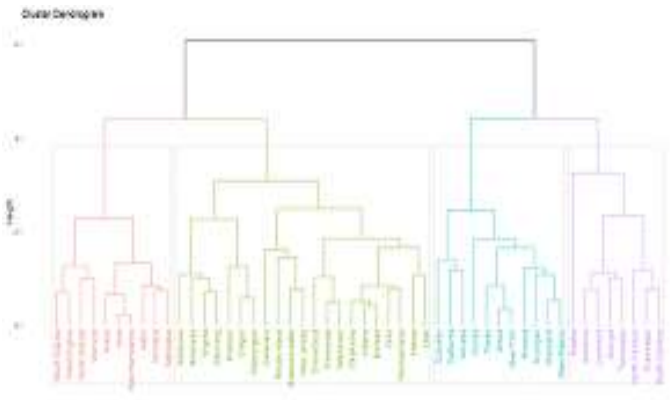


Figure 5: Plot Dendrogram

PARTITION AROUND MEDIODS:

By applying pam algorithm we identify the location where the murders occur frequently.by using this we can stop crime by sending polices before occurring the crime.

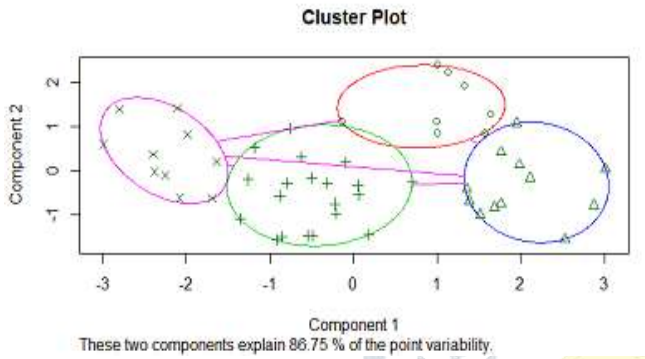



Figure 6: Plot using PAM

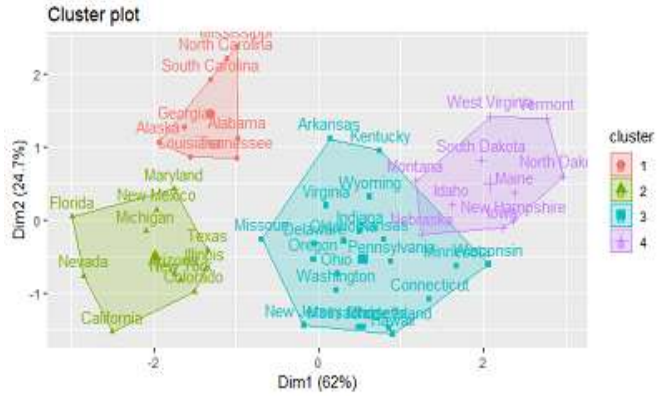



Figure 7: plot PAM

MEAN-SHIFT ALGORITHM:

By using mean shift algorithm we identify the location of the crime and stop the crime before it is occurring by sending polices.


Figure 8: Plot Mean Shift

Internal Validation:

The analytic results are shown as follows (as to how well they were clustered). We compared the clustering by first looking at the within. Clusters, which are the low number considered better than high number.

| Algorithm | Accuracy |
|---|---|
| K-Means | 57% |
| Hierarchical | 51% |
| Pam | 76% |
| Mean shift | 95% |

## VI.CONCLUSION

In this paper, we have applied different clustering techniques to get the best result that can help police officers to improve their work. Clustering is a popular. It is intended to identify several clusters discovered in databases using different measures of interestingness. First, we prepared the selected attributes to use them in creating the clusters. Second, we used some measures to determine the optimal number of clusters for each algorithm. Then, we created different clusters by using different methods. Finally, we used several visualization techniques to represent the clusters' profiles.

## REFERENCES

[1] J. K. Author, "Title of R. Xu, B. and D. Wunsch II, "Survey of Clustering Algorithms," IEEE Trans. On Neural Networks, VOL. 16, NO. 3, pp. 645–678, May 2005.
[2] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," KDD Workshop on Text Mining, 2000.
[3] R. Ali, U. Ghani, and A. Saeed, "Data Clustering and Its Applications," Rudjer Boskovic Institute, 2001.
[4] J. Han, M. Kamber and J. Pei and M. Kamber, Data Mining, Concepts and Technologies, 3rd Edition, The Morgan Kaufmann, , 2011.
[5] Sang C. Sug, Practical Applications of Data Mining, Jones & Bartlett, 2012.
[6] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Edition, Wiley-IEEE Press, August 2011.
[7] The R Project for Statistical Computing, The R Foundation.
[8] K.-C. Wong and Z. Zhang, "Snpdryad: predicting deleterious nonsynonymous human snps using only orthologous protein sequences," Bioinformatics, p. btt769, Jan 2014.
[9] T. Gonzalez, "On the computational complexity of clustering and related problems," in System Modeling and Optimization, ser. Lecture Notes in Control and Information Sciences, R. Drenick and F. Kozin, Eds. Springer Berlin / Heidelberg, 1982, vol. 38, pp. 174–182,

10.1007/BFb0006133. [Online]. Available: http://dx.doi.org/10.1007/ BFb0006133.

[10] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: A data-distribution perspective," IEEE Transactions on Systems, Man, and Cybernetics, Part B, vol. 39, no. 2, pp. 318–331, 2009.

[11] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, ser. SODA '07. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. [Online]. Available: http://portal.acm.org/citation.cfm?id= 1283383.1283494

[12] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645–678, May 2005. [Online]. Available: http://dx.doi.org/10.1109/TNN.2005.845141.

[13] K.-C. Wong, C. Peng, M.-H. Wong, and K.-S. Leung, "Generalizing and learning protein-dna binding sequence representations by an evolutionary algorithm," Soft Comput., vol. 15, no. 8, pp. 1631– 1642, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1007/ s00500-011-0692-5.

[14] F. Abascal and A. Valencia, "Clustering of proximal sequence space for the identification of protein families," Bioinformatics, vol. 18, pp. 908–921, 2002.

[15] C. Aggarwal and P. Yu, "Redefining clustering for high-dimensional applications," IEEE Trans. Knowl. Data Eng., vol. 14, no. 2, pp. 210–225, Feb. 2002.

[16] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in Proc. ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 94–105.

[17] Shi Na., Liu Xumin, Guan Yon , "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics(IITSI), pp.63-67, 2-4 April 2010.

[18] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science , Vol.10, pp:1626-1633,July 2006.

[19] S. Prakash kumar and K. S. Ramaswami, "Efficient Cluster Validation with K-Family Clusters on Quality Assessment", European Journal of Scientific Research, 2011, pp.25-36.