Crime Prediction Patterns using Hybrid K-Means Hierarchical Clustering

Geeta Chhabra
Department of Computer Science and Engineering
Amity School of Engineering, Amity University
Noida – 201313, Uttar Pradesh, India
geeta_chhabra@rediffmail.com

Vasudha Vashisht
Department of Computer Science & Engineering
Amity School of Engineering
Noida, Uttar Pradesh, India

ABSTRACT: *Data clustering in data mining has become an increasingly important research area in recent days. The proposed hybrid algorithms k-means hierarchical clustering uses k-mean clustering combined with the hierarchical cluster centres to analyze the crime patterns. The representative points will then construct the hierarchical tree using agglomerative hierarchical clustering algorithm called dendrogram. We have performed experiments to evaluate our approach on crime data available from National Crime Record Bureau, Ministry of Home Affairs, Govt. of India, website. Clustering is used for grouping the similar patterns to identify crime pattern. The experimental result shows that the proposed hybrid algorithm is more effective. This hybrid approach performs better than hierarchical clustering algorithm in terms of accuracy of clusters.*

Keywords: Hierarchical Clustering, K-Mean Clustering, Hierarchical K-means Clustering, Dendrogram, Cluster plot, Data Mining, Data Clustering, Algorithm

Received: 10 September 2019, Revised 4 November 2019, Accepted 2 December 2019

DOI: 10.6025/jitr/2020/11/1/1-11

1. Introduction

Nowadays most of the law enforcement agencies have volume of data which needs to be processed to transform into useful information. Data mining can greatly help in identifying and analyzing crime patterns and helps the enforcement agencies in reducing and preventing crime. It recognises and analyzes crime patterns to prevent further occurrence of such instances.

An appropriate data mining technique needs to be selected to perform analysis of crime related dataset. One of the techniques of data mining is clustering, which organises the data so that the data in same group is more similar than those in other groups. Such clusters are useful in identifying a crime pattern or crime spree.

In this paper, the unsupervised classification has been used using hierarchical and k mean clustering for predicting crime zone. Also, the hybrid approach using the best of hierarchical and k-mean clustering has been used to determine crime zone. Various statistical techniques have been used to assess the clustering tendency and to estimate optimal number of clusters. Results show that Hybrid K-Means Hierarchical Clustering performs better in terms of accuracy of clusters.

This technique is used to process the high volume of crime dataset to extract and interpret the data which will help the law enforcement agencies to analyze crime patterns in order to identify similar incidence which will reduce the further occurrences of such incidences. We have implemented, the proposed algorithm using packages like cluster, factoextra for simplifying cluster flow and ggplot for visualization from R. (Maechler et al, 2018; Wickham, 2016; Kassambara & Mundt, 2017).

1.1. Clustering Algorithms

It is used to split a dataset into several meaningful groups (i.e. clusters) so that similar objects are grouped together. One of the important data mining techniques is clustering that is used to place data elements of a given dataset into related groups without having any advance knowledge of the group definition. The similar clusters help in easy retrieval of information. It is a technique of partitioning data into related groups so that patterns and order are visible. The researchers are committed to develop new techniques as traditional querying methods are not sufficient to meet the raised requirements. Increase in volume of data, complexity of data and non-availability of a well-planned approach used for data computation has given rise to a number of computational challenges. Clustering methods are mainly either partitioning or hierarchical methods. The most popular clustering algorithms are:

1.1.1. K-means Clustering

This is a most widely used partitioning method due to its easy understanding and fast speed, where dataset is split into a set of k clusters. The main disadvantage of this algorithm is that it is difficult to predict the value of k. Also, the randomly chosen initial points may result in different final clusters. It is completely unstructured approach which is sensitive to outliers and results in an unorganized collection of clusters which does not provide useful information (Agarwal & Upadhyay, 2014; Gajawada & Toshniwal, 2012; Baiwal & Raghuvanshi, 2016). In k-mean clustering, the number of clusters are to be specified in advance. Hence it selects randomly initial centroids.

1.1.2. Hierarchical Clustering

It is an alternative to k-means clustering for specifying clusters. It uses pair-wise distance matrix between observations as clustering criteria. (Gholamian et al, 2013). The number of clusters needs not to be known in advance. It results in tree-based representation of the observations known as dendrogram. It can further be defined as agglomerative and divisive. (Bhagwat, 2013).

• Agglomerative Clustering

It is also known as **AGNES** (**Agglomerative Nesting**). It works in bottom-up manner. Each object in this is considered a one-element cluster **(**leaf**)**. The clusters which are more similar to each other are combined into a new bigger cluster **(**nodes**)** at each step. This method is iterated till all data points are member of just one single big cluster **(**root**)**. This results in a tree which is known as dendrogram. This algorithm is good at identifying small clusters but not large ones.

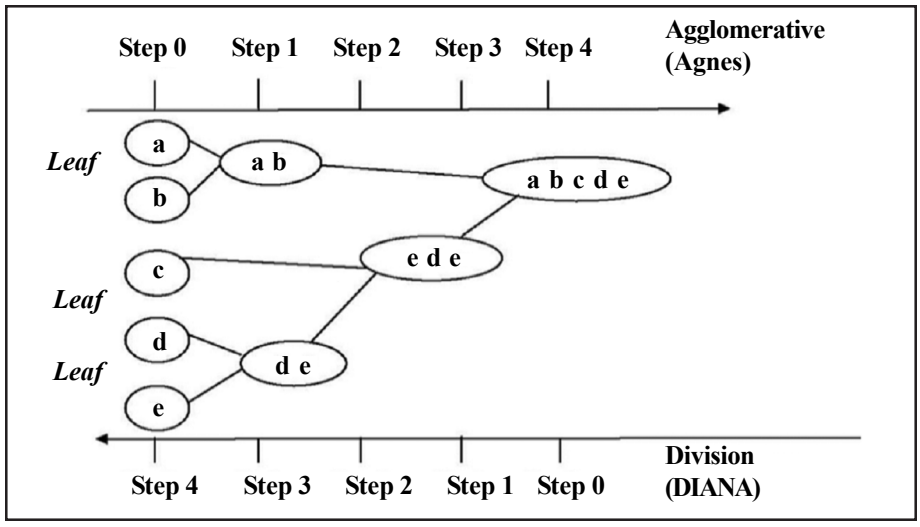

Figure 1. Agglomerative and Divisive hierarchical clustering

• Divisive Hierarchical Clustering

It is also known as **DIANA** (**Divise Analysis**) and it works in a top-down manner. The algorithm is an inverse of AGNES. It starts with the root, which is one big cluster of all objects. The most heterogeneous cluster is divided into two till all objects are in their own cluster.

2. Related Work

Dubey & Chaturvedi (2014) have extensively analysed and compared in their article "A Survey Paper on Crime Prediction Technique Using Data Mining", the various data mining techniques in the spree of crime detection and prediction. They have studied various computational techniques to predict future patterns of crime. They have compared the various data mining techniques for detection & prediction of future crime patterns. They have compared these techniques in terms of concept, predictive accuracy, performance and their disadvantages.

Malathi & Baboo (2011) have proposed in their article "Evolving Data Mining Algorithms on the Prevailing Crime Trend – An Intelligent Crime Prediction Model" missing value imputation techniques and clustering algorithm for crime data from Indian scenario on two major crimes burglary and murder.

Pante et al (2016) in their article "Crime Detection using Data Mining" have predicted crime patterns by applying regression and classification algorithms. They have also applied decision tree, bayesian network, forecasting models such as ARIMA (Auto-regressive integrated moving average) and Artificial neural networks (ANN) and visualized the results using K-means clustering in WEKA tool.

Gholamian et al (2013) have introduced in the article "A New Method for Clustering in Credit Scoring Problems", a clustering method viz self-organizing map neural network to cluster defaulter customers. They have applied neural network on UCI machine learning data for Australian credit data set. This data has 690 instances with 14 attributes & one class with two category default and Non-default. They have compared results with K-means, SOM, PAM against different external and internal measures. The time adaptive self-organizing map (TASOM) neural network used by them is a modified self-organizing map (SOM) neural network. It has adaptive learning rates and neighbourhood sizes. The experimental results show that the TASOM is better in performance for customers clusters.

Agarwal & Upadhyay (2014) in their article "A Fast Fraud Detection Approach using Clustering Based Method" have proposed clustering for detecting fraud in credit card transactions dataset. It has been increased considerably in recent years. In this paper, outlier detection and clustering techniques have been described which are used to find these mischief activities. Datasets are partitioned by clustering. The fraudulent data is identified by outlier detection technique. To find the outliers from the dataset, both techniques are combined efficiently, in this paper. The experimental results show that proposed method has less computational cost. It performs better than distance-based method on real dataset.

From the literature study, it could be concluded that crime data is increasing which in turn requires the need for advanced and efficient techniques for data analysis. The crime data can be efficiently analysed using data mining techniques; to help law enforcement agencies to achieve a truly preventative approach. They get the new understanding from data, recognizing and identifying suspicious behavior and activities and are enabled to get a head start on the criminals. The crime analytic field has lot of potential to identify better techniques & needs further investigation.

3. Proposed Methodology

The most widely used clustering algorithms are hierarchical and k-means clustering algorithms. k-means is computationally faster and produces tighter clusters. On the other hand, the hierarchical clustering is more informative because it produces the complete hierarchy of clusters. Beside these benefits, both of these have some limitations. The performance of k-means clustering depends on the initial number of clusters. The hierarchical clustering is less efficient. Inspite of this, both of algorithms are used in analysis of crime data. As a solution to this, in this article, we have documented hybrid approach that combines the advantages of both hierarchical and k-means clustering. Popular approach in research is to combine different algorithms so as to overcome their limitations and produce better results. (Singh & Kaur, 2013; Paithankar & Bharat, 2014; Verma et al, 2016). In k-means algorithm, a random set of observations are chosen as the initial centers. As we know that with initial

random selection of the clusters, the final k-means clustering solution is very sensitive. So, we might get a slightly different result each time we compute k-means. To avoid this, an ideal solution is to use a hybrid approach to combine the best features of both hierarchical and k-means clustering. This process is named hybrid hierarchical k-means clustering (hk-means).

The hybrid approach procedure is as follows:

Step 1: Compute the data using hierarchical clustering and cut it into k-clusters

Step 2: Compute mean that implies the center of each cluster.

Step 3: By using the set of cluster centres (computed in Step 2), compute k-means. It will improve the initial partitioning generated at step 1 of the algorithm. Hence, the initial partitioning can be slightly different from the final partitioning obtained in the step 3.

4. Experimental Analysis

The data on this has been taken from publication, Crime in India, 2015 of National Crime Record Bureau (NCRB), Ministry of Home Affairs, Govt of India. The dataset has variables namely, number of murders, rape, robbery, theft per lakh in 36 states of India along with population in lakhs. The aim is to classify each state in crime zone according to the similarity of the crime. The unsupervised classification has been used using hierarchical and k means clustering for predicting crime zone. Also, the hybrid approach using the best of hierarchical and k-means clustering has been applied to determine the crime zone. The results of hierarchical clustering and hybrid hierarchical k-means clustering has been compared.

4.1. Method of Accessing the Cluster Tendency

Clustering algorithms, which include partitioning methods such as PAM, CLARA, FANNY, hierarchical and k-means clustering (Jafar & Sivakumar, 2013) is used to split the dataset into clusters of similar objects. For any dataset, before applying any clustering method, it is important to check whether the dataset has tendency to form meaningful clusters. A major issue with such clustering methods, is that it will return clusters even if there are not any clusters which means, if someone blindly applies a clustering analysis, it will by default divide the data into different clusters. Hence, before choosing an appropriate clustering approach, the analyst has to take a decision whether the dataset has any meaningful clusters (i.e. non-random structures) and then the optimal number of clusters possible. This process is defined as the assessing of clustering tendency or the feasibility of the clustering analysis. Cluster tendency assessment determines whether a given dataset has meaningful clusters (i.e., non-random structure).

There are mainly two methods for determining the clustering tendency:

i) Statistical method **(**Hopkins statistic**)**: In this method, cluster tendency is checked by measuring the probability of dataset whether it is uniformly distributed (Banerjee & Dave, 2004). In other words, it checks the spatial randomness of the dataset. A value of about 0.5 means that data is close to each other and is uniformly distributed.

The null and the alternative hypotheses are defined as follow:

- **Null hypothesis**: The data is uniformly distributed with no meaningful clusters.
- **Alternative hypothesis**: The data is not uniformly distributed or has meaningful clusters.

If hopkins statistic has value close to 0, the null hypothesis can be rejected. It can be concluded that the dataset is statistically significant and is clusterable. In the given dataset of NCRB, the value of *H* is 0.1932429. So, it is concluded that the dataset is highly clusterable as the value of H is 0.19 far below the threshold value 0.5

ii) Visual Assessment of cluster Tendency (VAT) algorithm:

This approach can be used to visually inspect the clustering tendency. The algorithm of VAT is as follows (Yingkang H., 2012):
1. Compute the dissimilarity (DM) matrix between the objects using **Euclidean distance measure**

2. Reorder the DM to make the similar objects close to each other to create an **ordered dissimilarity matrix** (ODM)

3. The ODM is displayed as an **ordered dissimilarity image** (ODI), which is the visual output of VAT

The dissimilarity matrix image confirms that there is a cluster structure in the crime data set. The VAT detects the clustering tendency in a visual form by counting diagonally the number of dark square shaped blocks. The figure 2, suggest three number of clusters which are represented by three well-formed blocks in ordered dissimilarity image.

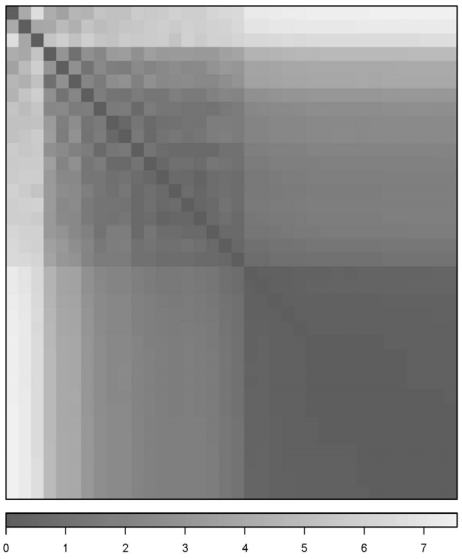

Figure 2. Ordered Dissimilarity Image for crime data

Both the methods, hopkins statistics and a visual method, shows that crime data is clusterable.

4.2. Optimal Number of Clusters

To determine the optimal number of clusters, the direct and statistical testing methods have been used.

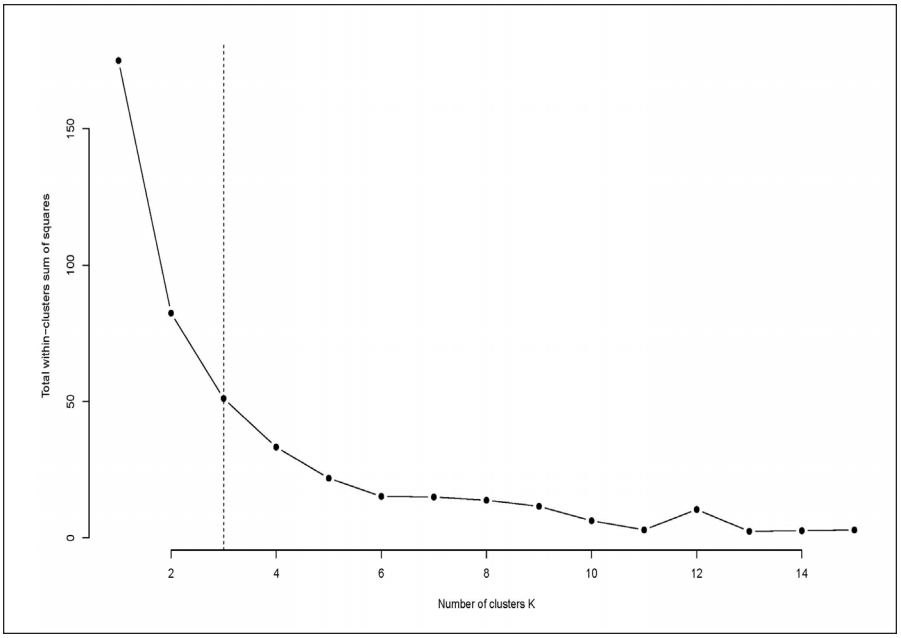

Figure 3. Optimal number of clusters using Elbow method for K-Means

• **Direct Methods.** Most widely used direct methods are *elbow*(Syakur et al, 2018; Bholowalia & Kumar, 2014) and *silhouette* (Thinsungnoena et al 2015) which optimize a criteria such as average silhouette or within cluster sums of squares.

• **Statistical testing methods** such as gap statistic which consists of comparing evidence against null hypothesis.

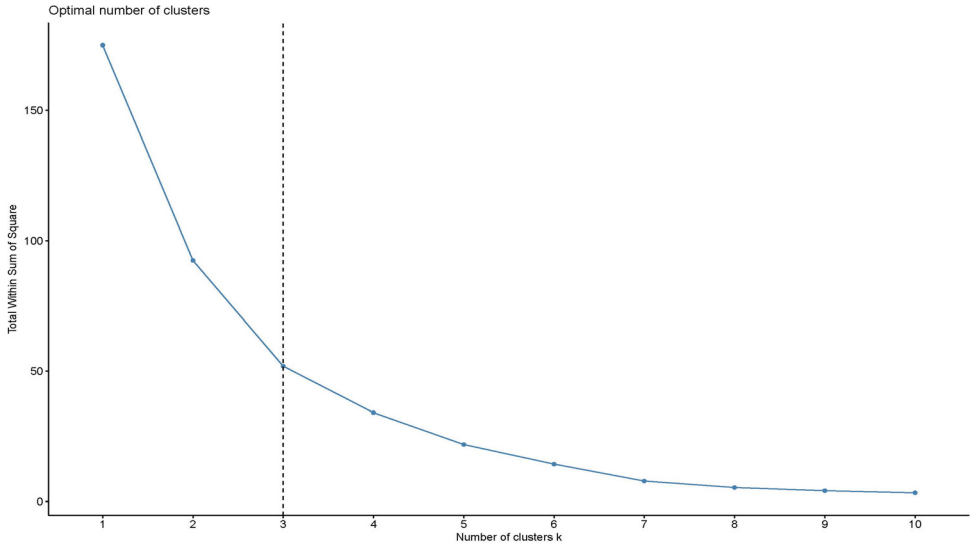

Figure 4. Optimal number of clusters using Elbow method for Hierarchical clustering

Thus, using elbow method, figure 3 & figure 4 shows that for crime data, the optimal number of clusters are 3 using k-means and hierarchical clustering.

4.3. Cluster Analysis

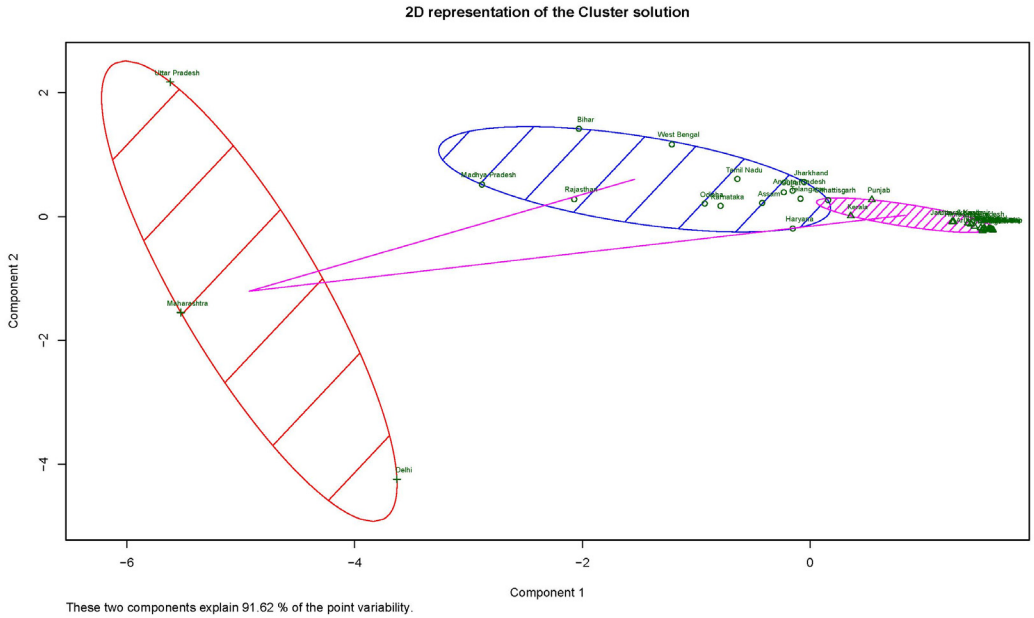

Figure 5. Cluster Plot using K-Means Clustering

After assessing the cluster tendency & optimal number of clusters for crime data, the cluster analysis has been performed using hierarchical, k-means and hierarchical k-means clustering.

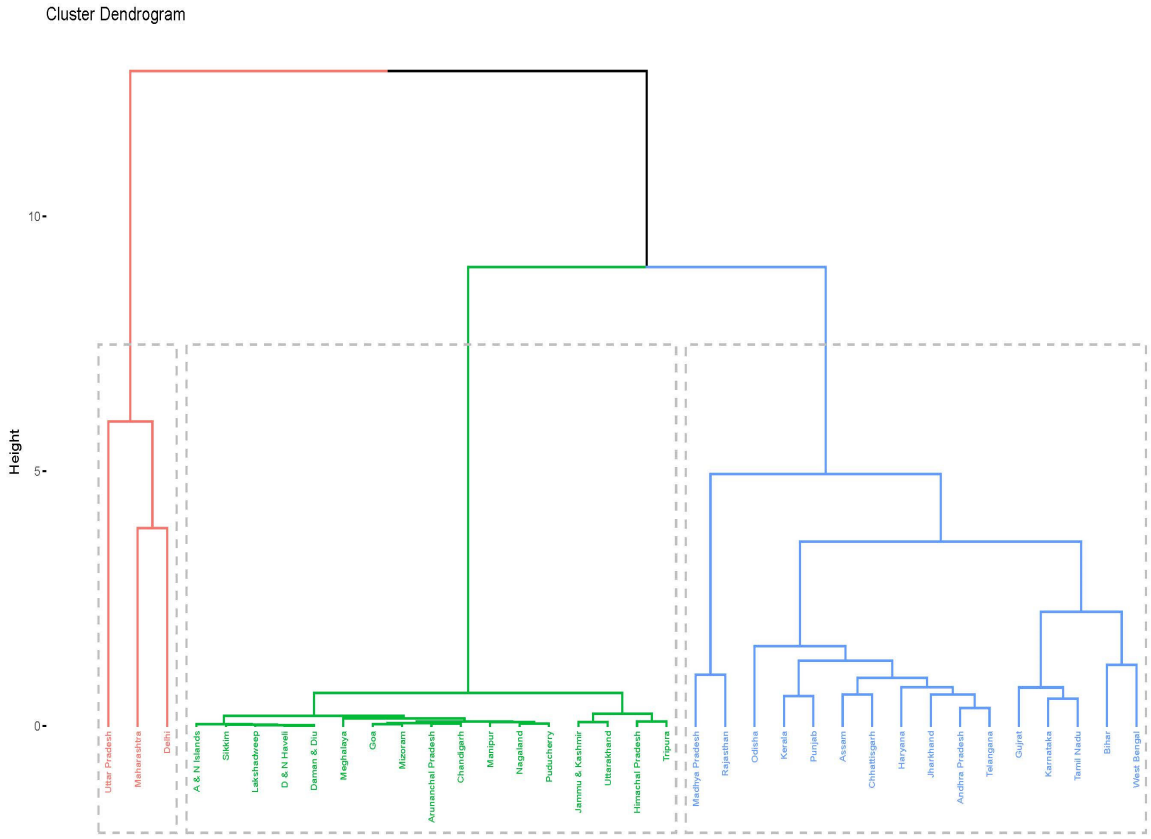

Figure 6. Dendogram for hierarchical Clustering

The cluster plot for k mean, dendrogram for hierarchical clustering and hierarchical k-means clustering has been generated.

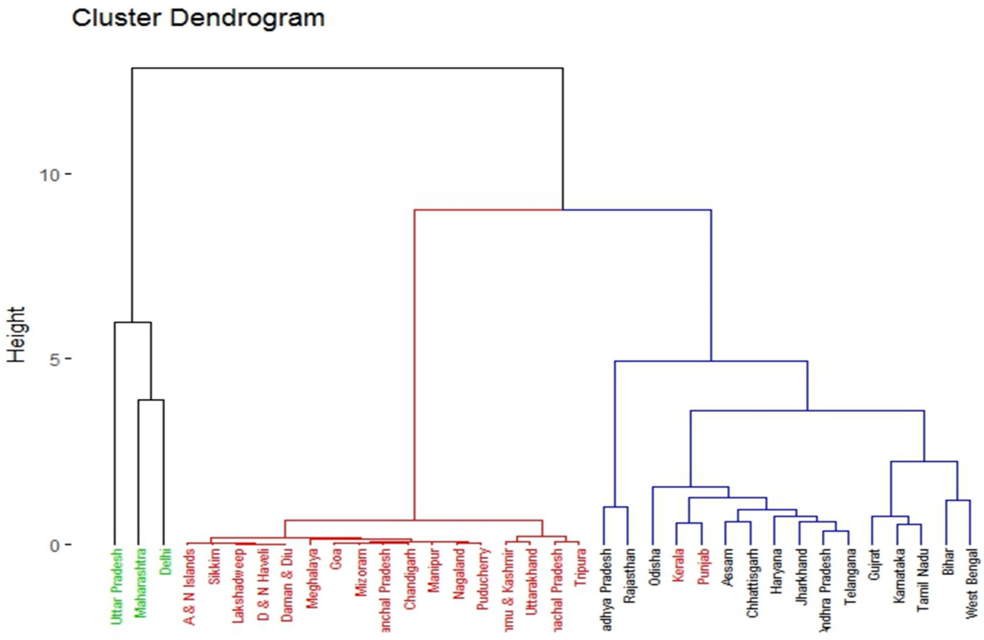

Figure 7. Dendrogram for hierarchical k-means clustering

4.4. Cluster Validation:
To check the accuracy of clusters

1. Silhouette plots have been generated for hierarchical, k-means, clustering and hierarchical k-means clustering. It measures the average distance between clusters to determine how well the observations are clustered. It displays the closeness of points in one cluster to the point in neighbouring clusters.

2. The confusion matrix defined between initial clusters using only hierarchical clustering and hierarchical k-means clustering has been generated.

4.4.1 Silhouette Plots

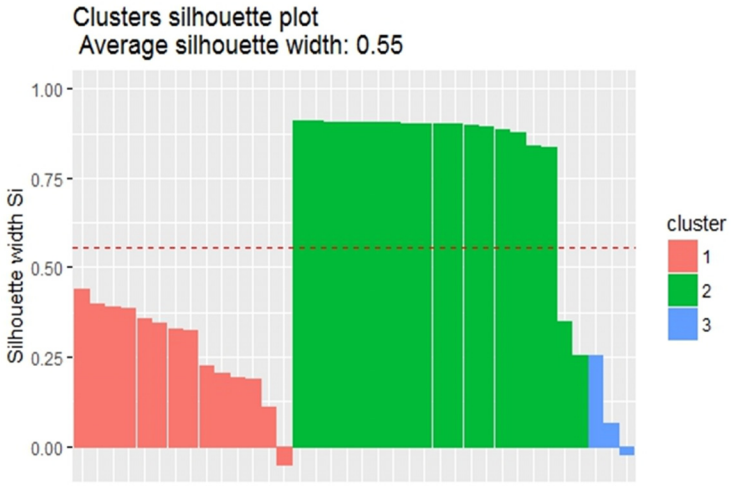

Figure 8. Silhouette Plot for K-Mean Clustering

Silhouette Analysis for k-Mean Clustering says that there are three clusters of size 14, 19 and 3 with the width 0.27, 0.83 and 0.10 and average silhouette width is 0.55(fig. 8). The state of Chhattisgarh and Uttar Pradesh have negative silhouette index means that they have been placed in wrong clusters (in 1 and 3) and should be placed in cluster 2 and 1 respectively.

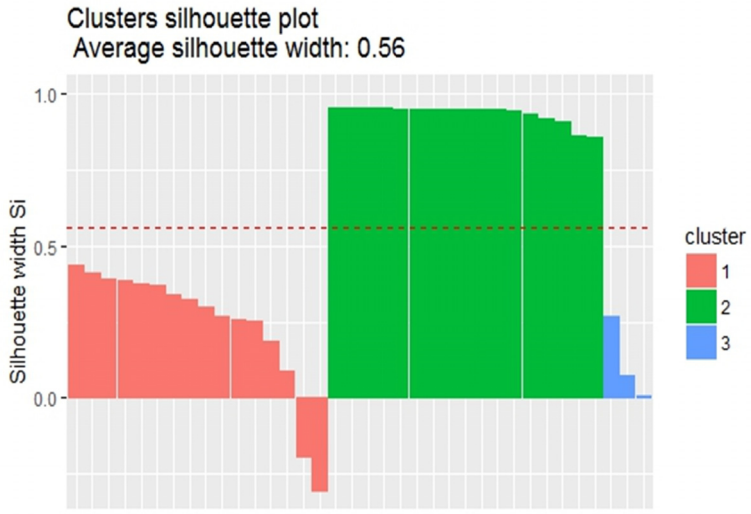

Figure 9. Silhouette Plot for Hierarchical clustering

Silhouette Analysis for hierarchical clustering says that there are three cluster of size 16, 17 and 3 with the width 0.24, 0.93 and 0.12 and average silhouette width is 0.56(fig. 9). The state of Kerala and Punjab have negative silhouette index means that they have been placed in the cluster 1 and should be placed in cluster 2.

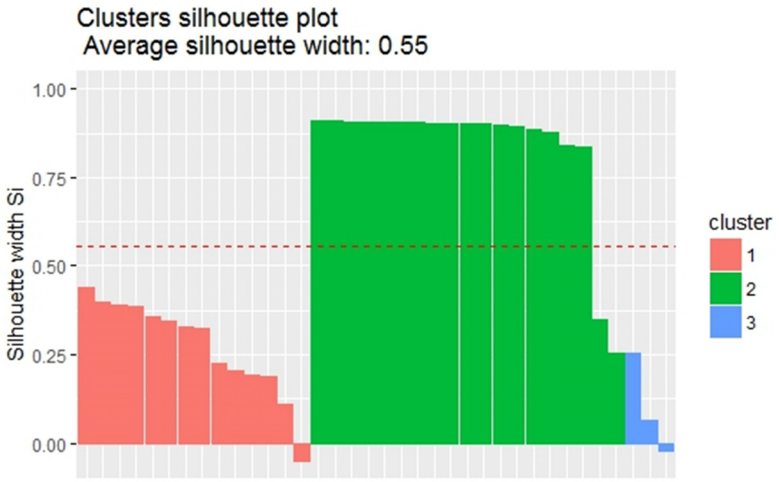

Figure 10. Silhouette Plot for hierarchical k-means clustering

The silhouette analysis for hierarchical k-means clustering says that there are three clusters of size 14,19 and 3 with width of 0.27, 0.83 and 0.10. Average silhouette width is 0.55(fig.10).

4.4.2. Cross Tabulation

The confusion matrix defined between initial clusters using only hierarchical clustering and the final ones defined using hierarchical k-means clustering is

Clusters	**1**	**2**	**3**
1	14	0	0
2	2	17	0
3	0	0	3

Table 1. Confusion Matrix Predicting the Clusters

5. Conclusion & Future Work

It can be seen from the confusion matrix that 2 observations belonging to cluster 1 has been reclassified into cluster 2 in hierarchical k-means clustering. Also, it can be seen from the dendrogram in figure 7 that Kerala and Punjab have been wrongly placed. Silhouette Analysis for hierarchical clustering also states the same. Therefore, there is an improvement using hierarchical k-means clustering over hierarchical clustering. Thus, the state of Uttar Pradesh, Maharasthra and Delhi have been placed in one crime zone which means that they fall in one group and follow the same pattern of crime.

Thus, crime analytics involves systematic analysis to identify & analyze patterns, trends and disorders in crime to help law enforcement agencies to deploy resources in effective manner. It also assists in identifying the suspects. Crime analytics plays an important role in devising solutions to crime problems and formulating crime prevention strategies.

This approach deals with numerical data, it can be used for text mining also. It can also be implemented on more complex and varying datasets. Other than criminology, it can be applied in several other fields like healthcare for predicting the patterns of diseases, in agriculture for finding the various patterns of yields based on particular fertilizers or soil inputs, monitoring traffic in crowded cities etc. Also, it can be used to predict accident prone area by health care agencies; it can be used to find the location of towers by Telephone Company where optimum strength of signals can be received and to place the patrolling vans in the vicinity of the areas where crime rate is high.

References

[1] Agarwal, S., Upadhyay, S. (2014). A Fast Fraud Detection Approach using Clustering Based Method. *Journal of Basic and Applied Engineering Research.*

[2] Bhagwat, K., Dhanshri, M., Sayali, S., Akshay, D., Tornekar, R. (2013). Comparative Study of Brain Tumour Detection Using K means, Fuzzy C Means and Hierarchical Clustering Algorithms. *International Journal of Scientific & Engineering Research*.

[3] Baiwal, S., Raghuvanshi, A. (2016). Imputation of Missing Values using Association Rule Mining & K-Mean Clustering. *International Journal of Scientific Development and Research*.

[4] Banerjee, A., Davé, R. N.(2004). Validating clusters using the Hopkins Statistic. *IEEE International Conference on Fuzzy Systems.*

[5] Bholowalia, P., Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*.

[6] Dubey, N., Chaturvedi, S. K. (2014). A Survey Paper on Crime Prediction Technique Using Data Mining. International. *Journal of Engineering Research and Applications*.

[7] Gajawada, S., Toshniwal, D. (2012). Missing Value Imputation Method Based on clustering and Nearest Neighbours. *International Journal of Future Computer and Communication*.

[8] Gholamian, M. R., Jahanpour, S., Sadatrasoul, S. M. (2013). A New Method for Clustering in Credit Scoring Problems. *Journal of Mathematics and Computer Science*.

[9] Jafar, M. O. A., Sivakumar, R. (2013). A Comparative Study of Hard and Fuzzy Data Clustering Algorithms with Cluster Validity Indices. *Emerging Research in Computing, Information, Communication and Applications.*

[10] Kassambara, A., Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

[11] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2018). *cluster: Cluster Analysis Basics and Extensions. R package version* 2.0.7-1.

[12] Malathi, A., Baboo, S. S. (2011). Evolving Data Mining Algorithms on the Prevailing Crime Trend – An Intelligent Crime Prediction Model. *International Journal of Scientific & Engineering Research*.

[13] Paithankar, R., Bharat, T. (2014). A H-K Clustering Algorithm for High Dimensional Data Using Ensemble Learning. *International Journal of Information Technology Convergence and Services*.

[14] Pande, V., Samant, V., Nair, S. (2016). Crime Detection using Data Mining. *International Journal of Engineering Research & Technology.*

[15] Singh, G., Kaur, N. (2013). Implementation of Hybrid Clustering Algorithm with Enhanced K-Means and Hierarchal Clustering. *International Journal of Advanced Research in Computer Science and Software Engineering.*

[16] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of The Best Customer Profile Cluster. *IOP Conf. Series: Materials Science and Engineering.*

[17] Thinsungnoena, T., Kaoungku, N, Durongdumronchai, P., Kerdprasop, K., Kerdprasop, N. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. *Proceedings of the* 3rd *International Conference on Industrial Application Engineering.*

[18] Verma, V., Bhardwaj, S. , Singh, H. (2016). A Hybrid K-Mean Clustering Algorithm for Prediction Analysis, *Indian Journal of Science and Technology.*

[19] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

[20] Yingkang, H. (2012). VATdt: Visual Assessment of Cluster Tendency Using Diagonal Tracing. American. *Journal of Computational Mathematics.*