

# Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches

S.Sivaranjani

Assistant Professor  
Department of Computer Science  
and Engineering  
Avinashilingam University  
Coimbatore, India  
sivaranjanicse@gmail.com

Dr.S.Sivakumari

Professor and Head  
Department of Computer Science  
and Engineering  
Avinashilingam University  
Coimbatore, India  
prof.sivakumari@gmail.com

Aasha.M

Assistant Professor  
Department of Computer Science  
and Engineering  
Avinashilingam University  
Coimbatore, India  
arathil800@gmail.com

**Abstract**—Crime is one of the most predominant and alarming aspects in our society and its prevention is a vital task. Crime analysis is a systematic way of detecting and investigating patterns and trends in crime. In this work, we use various clustering approaches of data mining to analyse the crime data of Tamilnadu. The crime data is extracted from National Crime Records Bureau (NCRB) of India. It consists of crime information about six cities namely Chennai, Coimbatore, Salem, Madurai, Thirunelveli and Thiruchirapalli from the year 2000-2014 with 1760 instances and 9 attributes to represent the instances. K-Means clustering, Agglomerative clustering and Density Based Spatial Clustering with Noise (DBSCAN) algorithms are used to cluster crime activities based on some predefined cases and the results of these clustering are compared to find the best suitable clustering algorithm for crime detection. The result of K-Means clustering algorithm is visualized using Google Map for interactive and easy understanding. The K-Nearest Neighbor (KNN) classification is used for crime prediction. The performance of each clustering algorithms are evaluated using the metrics such as precision, recall and F-measure, and the results are compared. This work helps the law enforcement agencies to predict and detect crimes in Tamilnadu with improved accuracy and thus reduces the crime rate.

**Keywords**— *Crime, clustering, classification, Google Maps, K-Means, KNN, DBSCAN, Agglomerative.*

## I. INTRODUCTION

Crime is a violation against the humanity that is often accused and punishable by the law. Criminology is a study of crime and it is interdisciplinary sciences that collects and investigate data on crime and crime performance. The crime activities have been increased now-a-days and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are voluminous data of crime exist. So we need methodologies to

predict and prevent crime. Data Mining provides clustering and classification technique for this purpose. Clustering is used for grouping the similar patterns to identify crimes. Cluster refers to a geographical collection of crime that can be visualized using the geo-spatial plot in the map. It is a group of clusters which specifies the relationship among the clusters. The clusters that are densely populated are called hotspots. Clustering in crime is mainly used to identify the patterns in crime and also used to predict the crime. Classification is a technique of data analysis that is used to extract and predicts future trends in data based on similarity measures.

The objective of this work is to predict crime in six cities of Tamilnadu by using clustering methods and to identify criminals by using classification methods. For this purpose we use KNN classification, K-Means clustering, Agglomerative hierarchical clustering and DBSCAN clustering algorithms..

## II. RELATED WORK

Kurt Hornik et al., (2012) introduced the R-extension package that delivers a computational environment containing fixed point and genetic algorithm solvers with interfaces to two external solvers (CLUTO and Gmeans) for spherical k-means clustering. The large scale benchmark has been used to analyze the performance of the solvers. The authors concludes that the solvers provides better solution with the interface Gmeans and CLUTO that gives fast and enhanced results.

Brian Kulis and Michael I. Jordan (2011) analysed the links rising among the DP mixture models and hard clustering algorithms by revisit the k-means clustering algorithm from a Bayesian nonparametric viewpoint. They proposed hierarchical Dirichlet process provides high precision results and reduce the time computation complexity.

Navjot Kaur et al., (2012) presented an overview to k-means clustering. They used ranking method for k-means clustering and compared its performance with the traditional k-means clustering. As a result of the comparison the authors concludes that the ranking based k-means clustering had less execution time and provides better results than traditional method.

Xingan Li and Martti Juhola (2014) had applied self-organizing map (SOM) for mapping crime data of 56 countries with 28 different crime situations and conclude that SQM would be a new crime mapping tool that would process large amount of data.

A. Malathi and Dr. S. Santhosh Baboo (2011) developed a crime analysis tool for Indian scenario by using various data mining techniques like k-means clustering and DBSCAN clustering. The authors concludes that the tool can be used by the Indian police and law agencies for crime detection and prevention as it provides faster analysis results and identifies common crime patterns.

Raphael Obi Okonkwo and Francis O. Enem (2011) analyzed various data mining techniques that been adopted by various law agencies to identify and prevent terrorism. The authors had also studied the restrictions of data mining in fighting crime in Nigeria and conclude that the data mining can only be used to assist the law agencies to analyze crime.

Uttam Mande et al.,(2012) presented a new methodology for identifying the criminals who committed the crime and proposed a method that uses Generalized Gaussian Mixture Model to maps the criminals based on the eyewitness specified features and concludes that the proposed model had given a unique and accurate result.

Malathi et al., (2011) analyzed the crime data of police department using the data mining techniques such as clustering and classification and also identified the crime trends and suggested that this method can be used to reduce and prevent crime for the upcoming years.

### III. METHODOLOGY

#### Data Set

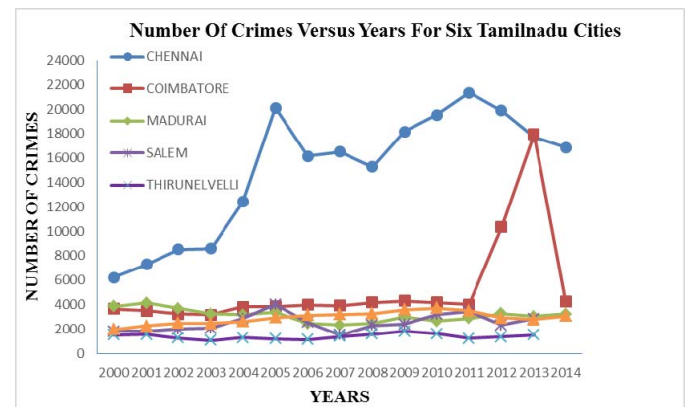
The crime dataset is extracted from the National Crime Records Bureau (NCRB) of India. The crime dataset consists of crime information about six cities of Tamilnadu namely Chennai, Coimbatore, Salem, Madurai, Thirunelveli and Thiruchirapalli during the year 2000-2014. The crime dataset consists of 1760 crime instances and 9 attributes. The attribute crime type consists of 20 different crime types like murder, rape, robbery, burglary and so on. Figure 1 shows the crime dataset.

Fig. 1. Crime dataset

CRIME_YEAR	CRIME_STATE	CRIME_CITY	CRIME_CITY_LATITUDE	CRIME_CITY_LONGITUDE	CRIME_TYPE	NUMBER_OF_CRIMES_IN_THE_CRIME_TYPE
2014	TAMILNADU	CHENNAI	13.082660	80.270718	MURDER	161
2014	TAMILNADU	CHENNAI	13.082660	80.270718	ATTEMPT TO COMMIT MURDER	235
2014	TAMILNADU	CHENNAI	13.082660	80.270718	RAPE	65
2014	TAMILNADU	CHENNAI	13.082660	80.270718	KIDNAPPING AND ABDUCTION	36
2014	TAMILNADU	CHENNAI	13.082660	80.270718	DRACONY	8
2014	TAMILNADU	CHENNAI	13.082660	80.270718	ROBBERY	72
2014	TAMILNADU	CHENNAI	13.082660	80.270718	BURGLARY	365
2014	TAMILNADU	CHENNAI	13.082660	80.270718	THEFT	1520
2014	TAMILNADU	CHENNAI	13.082660	80.270718	RIOTS	118
2014	TAMILNADU	CHENNAI	13.082660	80.270718	CRIMINAL BREACH OF TRUST	6
2014	TAMILNADU	CHENNAI	13.082660	80.270718	CHEATING	553
2014	TAMILNADU	CHENNAI	13.082660	80.270718	COUNTERFEITING	74
2014	TAMILNADU	CHENNAI	13.082660	80.270718	ARSON	17
2014	TAMILNADU	CHENNAI	13.082660	80.270718	HUNT	46
2014	TAMILNADU	CHENNAI	13.082660	80.270718	DOWNY DEATH	10
2014	TAMILNADU	CHENNAI	13.082660	80.270718	ABDUCTION OF WOMEN WITH INTENT TO OUTRAGE HER MODESTY	59
2014	TAMILNADU	CHENNAI	13.082660	80.270718	INSULT THE MODESTY OF WOMEN	157
2014	TAMILNADU	CHENNAI	13.082660	80.270718	CRUELTY BY HUSBAND OR HIS RELATIVES	302
2014	TAMILNADU	CHENNAI	13.082660	80.270718	CAUSING DEATH BY NEGLIGENCE	1083
2014	TAMILNADU	CHENNAI	13.082660	80.270718	OTHER CRIMES	3201
2014	TAMILNADU	COIMBATORE	11.016944	76.955632	MURDER	21
2014	TAMILNADU	COIMBATORE	11.016944	76.955632	ATTEMPT TO COMMIT MURDER	43

The total number of crimes in the six cities of Tamilnadu during 2000-2014 is analyzed by using line graph (Figure 2). The attributes crime year (Y-axis) and total number of crimes in the year (X-axis) is used to generate graph.

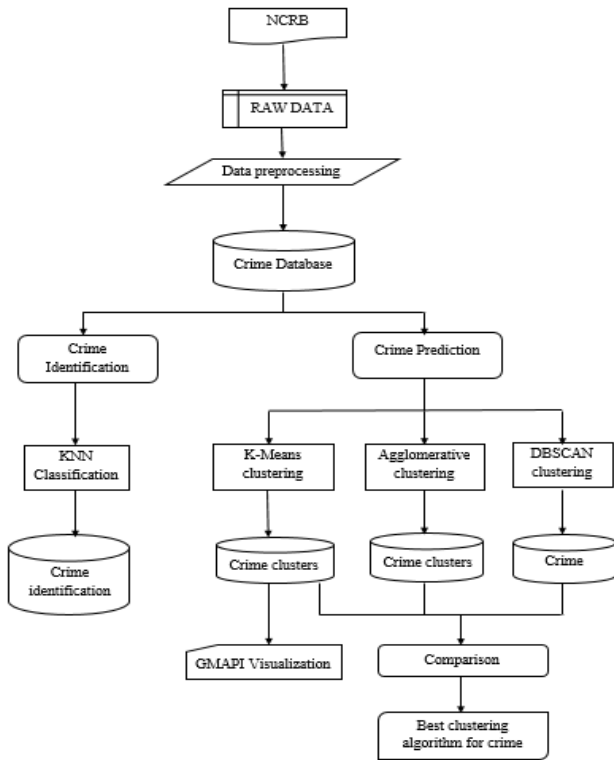
Fig. 2. Number of crimes versus years for six Tamilnadu cities



#### Crime Forecasting and Prediction

This section explains the work flow of our approach (Figure 3). The work flow starts by extracting the data from NCRB and preprocessing it to form the crime database. The database is then provided to crime identification and crime prediction. The crime identification is done using KNN classification which will also retrieve data from dataset based on similarity. The crime prediction is done by various clustering approaches like K-Means, Agglomerative and DBSCAN. These three clustering approaches are compared to discover the best accurate clustering algorithm to predict crime for the specified dataset.

Fig. 3. Work-Flow Diagram



#### IV. EXPERIMENTAL RESULTS

##### A. KNN Implementation:

The KNN classification searches through the dataset to find the similar or most similar instance when an input is given to it. The input to KNN is the query i.e. the attribute values of crime dataset. Based on that query KNN gives result that assist to analyze the large crime database and also helps in predicting the crime future in various cities. It draws crime patterns for various cities.

##### Algorithm: KNN classification

##### Input:

- 1) A finite set D of points to be classified,
- 2) A finite set T of points,
- 3) A function  $c: T \rightarrow \{1, \dots, m\}$ ,
- 4) A natural number k.

**Output:** A function  $r: D \rightarrow \{1, \dots, m\}$

##### Method:

- 1) Begin
- 2) For each x in D do
- 3) Let  $U \leftarrow \{\}$
- 4) For each t in T add the pair  $(d(x, t), c(t))$  to U;
- 5) Sort the pairs in U using the first components;

- 6) Count the class labels from the first k elements from U;
- 7) Let  $r(x)$  be the class with the highest number of occurrence;
- 8) End For each
- 9) Return r
- 10) End

KNN method stores all available objects and classifies new objects based on the similarity measure. It is used for criminal identification by considering the past crimes and discovering similar crimes that match the current crime based on number of nearest neighbors matched. The attributes crime type and crime city are considered as crime input attributes for KNN. All the attributes in the crime dataset can be considered but for the sake of brevity only two attributes are considered here. KNN method searches for nearest neighbor of the input values and filter those values from the dataset. The result of this method helps to retrieve data from the database and also assists in understanding the crimes.

##### B. K- Means Implementation

K-Means clustering is used to find internal patterns and relations in crime dataset. This method provides an outline of large crime data and simplify in handling, searching and retrieving of the preferred crime information.

The crime dataset is given as an input to this algorithm. The attributes crime\_year, crime\_city, crime\_type, total\_number\_of\_crimes\_in\_year are considered. The k-means clustering is performed for the following four cases,

*Case-1: Crime detection in during 2006-2014* The two attributes crime\_year and total\_number\_of\_crimes\_in\_year are used to generate clusters but independent of crime\_city and crime\_type. Here the number of clusters is given as 9 as there are nine years from 2000-2014.

*Case-2: Crime detection based on city during 2006-2014* The attributes crime\_year and total\_number\_of\_crimes\_in\_year are used to generate clusters, but it depends on the attribute crime\_city and independent of the attribute crime\_type. Here the number of clusters is given as 6 as there are six different cities of are considered in our dataset.

*Case-3: Crime detection of type burglary in during 2006–2014* The attributes crime\_year and total\_number\_of\_crimes\_in\_year are used to generate clusters, but it depends on attribute crime\_type and independent of attribute crime\_location. Here the crime\_type is considered as burglary and in the same way clusters can be generated for other crime types.

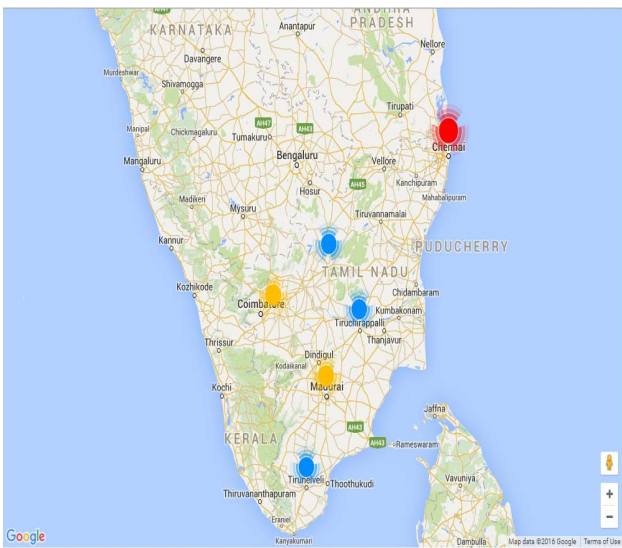
*Case-4: Crime detection of type burglary in Chennai during 2006–2014* The attributes crime\_year and total\_number\_of\_crimes\_in\_year are used to generate clusters, but it depends on attribute crime\_type and crime\_location. This case helps to detect which crime type is at peak in a given location in the particular year.

## Visualization of K-Means clustering




GMAPI is used to improve k-means results. It combines Google maps through Netbeans for user-friendly and improved visual aid to k-means. The attributes `crime_city`, `crime_city_longitude`, `crime_city_latitude` and `total_number_of_crimes_committed_in_city` are used for GMAPI. The attributes `crime_city_longitude` and `crime_city_latitude` are used to locate the crime city on Google map and the attribute `crime_city` and `total_number_of_crimes_committed_in_location` are used for crime detection during 2006–2014.

Now, if we want to know about the crimes in Tamilnadu cities for a particular year like 2014, we can directly locate with GMAPI (Figure 4)

Fig. 4. Crime clusters in GMAPI for year 2014 and crime type Burglary



Clusters with different number of crimes are represented by different color markers as:

-  : Blue color cluster for areas where less number of crimes happened.
-  : Yellow color cluster for areas where moderate number of crimes happened.
-  : Red color cluster for areas where more number of crimes happened.

In GMAPI, the crime year has to be selected to display the result. GMAPI generates crime clusters that recognize the hot spots of crime locations. Thus, GMAPI fastens the crime investigation and suggests implementing security actions in those affected areas.

## C. Agglomerative Hierarchical Clustering

This is a bottom-up strategy and it begins by assigning each object to its own cluster and then integrates these atomic clusters into a big clusters, until all of the objects are in a single cluster or until certain end conditions are fulfilled. The agglomerative clustering will never redo any step i.e. once attached object can never be separated. The steps of agglomerative clustering is,

- 1) Assign each object to an individual cluster.
- 2) Evaluate all pair-wise distances between clusters.
- 3) Construct a distance matrix using the distance values.
- 4) Look for the pair of clusters with the shortest distance.
- 5) Remove the pair from the matrix and merge them.
- 6) Evaluate all distances from this new cluster to all other clusters, and update the matrix.
- 7) Repeat until the distance matrix is reduced to a single element.

The crime dataset is given as an input to this algorithm. The attributes `crime_year`, `crime_city`, `crime_type`, `total_number_of_crimes_in_year` are considered. The agglomerative clustering is performed for the four cases that are considered for k-means. The four cases are,

- Case-1: Crime detection in during 2006-2014
- Case-2: Crime detection based on city during 2006-2014
- Case-3: Crime detection of type burglary in during 2006–2014
- Case-4: Crime detection of type burglary in Chennai during 2006–2014.

## D. DBSCAN Implementation

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm develops areas with appropriately high density into clusters and finds clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of density-connected points.

### Algorithm: DBSCAN Algorithm

**Input:** N objects to be clustered and global parameters Eps, MinPts.

**Output:** Clusters of objects.

**Method:**

- 1) Arbitrary select a point P.
- 2) Retrieve all points density-reachable from P wrt Eps and MinPts.
- 3) If P is a core point, a cluster is formed.
- 4) If P is a border point, no points are density-reachable from P and DBSCAN visits the next point of the database.

Continue the process until all of the points have been processed.



The crime dataset is given as an input to this algorithm. The attributes crime\_year, crime\_city, crime\_type, total\_number\_of\_crimes\_in\_year are considered. The DBSCAN clustering is performed for the four cases that are considered for k-means and agglomerative hierarchical clustering. The four cases are,

- Case-1: Crime detection in during 2006-2014.
- Case-2: Crime detection based on city during 2006-2014.
- Case-3: Crime detection of type burglary in during 2006-2014.

Case-4: Crime detection of type burglary in Chennai during 2006-2014.

#### E. Comparison of Clustering Approaches

The performance of the k-means clustering, agglomerative hierarchical clustering and DBSCAN clustering algorithms are evaluated by means of accuracy using precision, recall and F-measure. The accuracy is calculated by the following formulas.

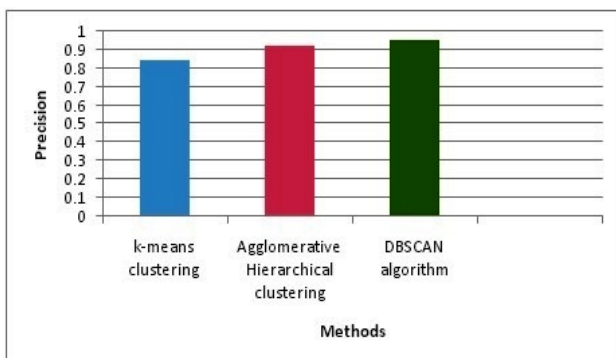
##### a. Precision

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant. In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

The precision is calculated as follows:

$$PRECISION = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

Fig. 5. Comparison using precision



From the figure 5 we can observe that the comparison of existing and proposed system in terms of precision metric. In existing scenario, the precision values are lower and precision of existing system is 0.84 using k-means clustering algorithm. In proposed system, the precision value is increased by using agglomerative hierarchical clustering and DBSCAN algorithm as 0.92 and 0.95 respectively. Thus it proves that proposed system is superior in crime detection performance.

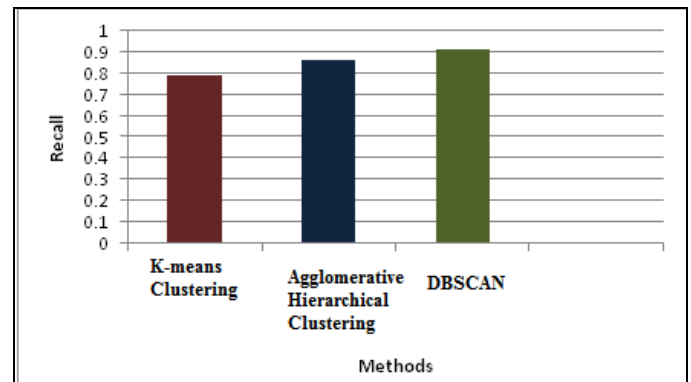
##### b. Recall

Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search. Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

The calculation of the recall value is done as follows:

$$RECALL = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

Fig. 6. Comparison using recall



From the figure 6 we can observe that the comparison of existing and proposed system in terms of recall metric. In existing scenario, the recall values are lower and recall of existing system is 0.79 using k-means clustering algorithm. In proposed system, the recall value is increased by using agglomerative hierarchical clustering and DBSCAN algorithm as 0.86 and 0.91 respectively. Thus it proves that proposed system is superior in crime detection performance.

##### c. F-Measure

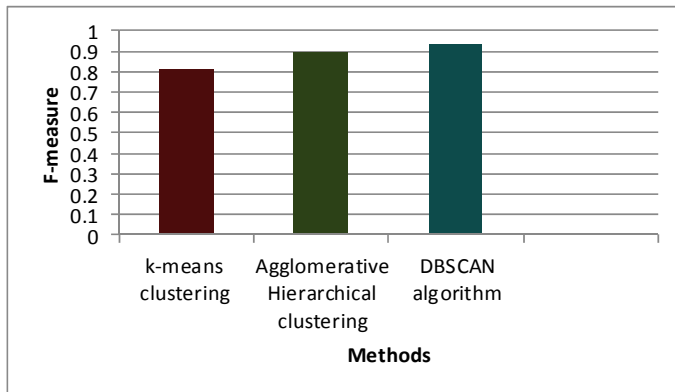
F-measure is an external measure for measuring goodness or accuracy of clustering methods. F-measure depends on two

factors Precision and recall. F-measure is calculated by weight average of recall and precision.

We can calculate the F-measure by using the following formula:

$$F = 2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Fig. 7. Comparison using recall



From the figure 7 we can observe that the comparison of existing and proposed system in terms of f-measure metric. In existing scenario, the f-measure values are lower and f-measure of existing system is 0.81 using k-means clustering algorithm. In proposed system, the f-measure value is increased by using agglomerative hierarchical clustering and DBSCAN algorithm as 0.89 and 0.93 respectively. Thus it proves that proposed system is superior in crime detection performance.

## V. CONCLUSION

Crime detection is the dynamic and emerging research field in the real world environment which aims to prevent the crime rates. Data Mining plays a vital role in law enforcement agencies in crime analysis in terms of crime detection and prevention. This work presents the method to predict and forecast crimes in six cities of Tamilnadu. Clustering techniques are used for crime detection and classification technique is used for crime prediction. The K-Means clustering, Agglomerative hierarchical clustering and DBSCAN clustering are implemented and their performance is evaluated based on accuracy. On comparing their performance the DBSCAN clustering gives result with high accuracy and effectively forms clusters than other two algorithms. The KNN classification is used for predicting crimes based on similarity search. Thus this system assists law enforcement agencies for an improved and accurate crime analysis.

In future, this work can be extended to have improved classification algorithms to identify criminals more efficiently.

We can also enhance privacy and other security measures to protect the crime data.

## REFERENCES

- [1] Kaur N, Sahiwal JK, and Kaur N, "Efficient k-means clustering algorithm using ranking method in data mining", *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 1(3) pp. 85-91, 2012.
- [2] Bajpai D, "Emerging Trends in Utilization of Data Mining in Criminal Investigation: An Overview", *Journal of Environmental Science, Computer Science and Engineering & Technology*, vol. 1(2), pp. 124-131, 2012.
- [3] Hornik K et al. "Spherical k-means clustering", *Journal of Statistical Software*, vol. 50(10), pp. 1-22, 2012.
- [4] Kalaikumaran T and Karthik S, "Criminals and crime hotspot detection using data mining algorithms: clustering and classification" *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 1(10), pp. 225, 2012.
- [5] Kumar J, Mishra S, and Tiwari N, "Identification of Hotspots and Safe Zones of Crime in Uttar Pradesh, India: Geo-spatial Analysis Approach", *International Journal of Remote Sensing Applications*. Vol. 2(1), pp. 15-19, 2012.
- [6] Lei L, "The GIS-based Research on Criminal Cases Hotspots Identifying". *Procedia Environmental Sciences*, pp. 12, vol. 957-963, 2012.
- [7] Mande U et al., "Feature specific criminal mapping using data mining techniques and generalized Gaussian mixture model", *International Journal Computer Science, Communications & Networking*, vol. 2(3), pp. 375-379, 2012.
- [8] Shafeeq A and Binu VS, "Spatial Patterns of Crimes in India using Data Mining Techniques", *International Journal of Engineering and Innovative Technology*, vol. 3(11), pp. 291-295, 2014.
- [9] Wang D et al., "Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns" *Computers, Environment and Urban Systems*, vol. 39, pp. 93-106, 2013.
- [10] Chainey S, Tompson L and Uhlig S, "The utility of hotspot mapping for predicting spatial patterns of crime", *Security Journal*, vol. 21(1), pp. 4-28, 2008.
- [11] Rachel Boba, "Introductory Guide to Crime Analysis and Mapping", Report to the Office of Community Oriented Policing Services, 2011.
- [12] Silvia Ferrari, Kelli C. Baumgartner, George B. Palermo, Roberta Bruzzone, And Marco Strano, "Comparing Bayesian And neural Networks For Decision Support In criminal Investigations", *IEEE Control Systems Magazine*, August 2008.
- [13] Oatley, G., Ewart, B., & Zeleznikow, J. "Decision support systems for police: Lessons from the application of data mining techniques to "soft" forensic evidence", *Artificial Intelligence and Law*, vol. 14(1-2), pp. 35-100, 2006.
- [14] Duan M. Y., Zang Y., Pu P. X., Shi L., "Geo-Info Graph Spectrum Analysis For Representing Distance Relations In Gis", *IEEE transaction on computing*, 2008.
- [15] McCue, C. "Data mining and predictive analysis. Intelligence gathering and crime analysis" (1st ed.,). Butterworth-Heinemann, 2007.