IJCAT - International Journal of Computing and Technology, Volume 5, Issue 4, April 2018
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

Crime Prediction and Analysis using Clustering Approaches and Regression Methods

[1] Raghavendhar T.V, [2] Joslin Joshy, [3] Mahaalakshmi R, [4] Ashutosh Soni M

[1] Department of CSE, SRM Institute of Science and Technology,
Vadapalani, Chennai, Tamilnadu, India

[2] Department of CSE, SRM Institute of Science and Technology,
Vadapalani, Chennai, Tamilnadu, India

[3] Department of CSE, SRM Institute of Science and Technology,
Vadapalani, Chennai, Tamilnadu, India

[4] Department of CSE, SRM Institute of Science and Technology,
Vadapalani, Chennai, Tamilnadu, India

Abstract - Crime is one of the biggest violations that has been not yet completely solved ever since the evolution of human race. In order to solve this, crime analysis and prediction is one of the methods. Crime analysis is a scientific way of developing effective strategies to prevent crime in future. In this project the crime analysis and prediction is done using different clustering approaches for and various regression methods. DBSCAN and k-means clustering methods are used for analysis and regression methods such as ridge, naïve Bayes and linear are used for prediction. Silhouette coefficient is used to determine the efficiency of the clustering methods. The error values from the regression are determined using root mean square method. The crime data is extracted from State Crime Records Bureau (SCRB) of Tamilnadu, India. It contains crime information about 38 different cities and districts. With the help of this approach, crime can be predicated and reduced it in the future.

Key words: *DBSCAN, k-means, linear regression, naive Bayes regression and ridge regression methods*

1. Introduction

Crime is a threat to humans caused by fellow beings which can be punishable by the law of government. A study of crime and the sciences that collects and investigate data on it and crime performance is called Criminology [12]. The activities of crime had been constantly increasing and it is the responsibility of the police to control and prevent it. Crime analysis and its prediction is difficult for the police as the volume of data is large. Therefore we need scientific methods to predict and analyze crime. This way the time and space is reduced. There are two methods of data mining for this purpose. They are clustering and regression methods respectively. The grouping of similar data is called clustering [14]. The dense spots are known as clusters and the rest is called noise. With the help of clustering, the data are sorted according to the conditions given. Regression is a method which helps to predict the crime in future with the given cluster of data [15].The objective of our work is to predict and analyze crimes in 38 different cities of Tamil Nadu by using clustering and regression methods. For this purpose K-Means and DBSCAN algorithms are used for clustering the data. The clustered data is then compared using silhouette co-efficient algorithm, from which the most effective clustering method from K-means and DBSCAN is determined for the given data set. The clustered data is then sent to Regression methods such as ridge, naive Bayes, and linear for predicting crime for the given data set and the result is viewed in the form of graphs.

The efficiency of each regression methods is determined using root mean square method. Big data–Hive is used for data storage in order to enhance privacy and security to protect the crime data.

2. Headings and Footnotes

In figure 1 the **raw data** it extracted from **SCRB** and is stored in **hive** .The data is then processed and clustered using DBSCAN and K-means clustering .The clusters are compared using Silhouette method and the best of the two clusters is stored for regression .The stored clustered data is then used in various regression methods. Finally the accuracy of each regression is calculated using root mean square method and is shown graphically.

2.1 Map Reduce:

Map reduce is model that generates data sets on a cluster [13]. The raw data is loaded in Hadoop using map reduce. Later, the same data is loaded into hive to make multiple copies of the data to avoid loss of data and to use the data in R. Code generation is done in eclipse as shown in figure 2.

2.2 Clustering:

2.2.1 DBSCAN:

DBSCAN is density-based spatial clustering of applications with noise. The DBSCAN algorithm is basically based on clustering points within the distance of epsilon with some initial minimum number of points. [13]It requires epsilon (Eps) as one parameter value and minimum number points as the other parameter. (MinPts).It begins with a random point as its starting point. It then identifies and joins all the nearby points within distance Eps of that particular starting point. A cluster is formed when the number of nearby points joined is greater than or equal to MinPts. If the nearby points is less than the minimum number of points the particular starting point is declared as noise. The start point is then marked as visited. The algorithm repeats the evaluation process for all the neighbors' repeatedly. If the number of neighboring nodes is less than MinPts, the point is marked as noise.

If a cluster is fully expanded then the algorithm continues to iterate through the remaining unvisited points in the dataset. The algorithm is as follows:

- Initiate a graph whose points to be clustered.
- For each starting-point a, create an edge from a to every point p in the range 0 of a.
- Set M to the points of the graph;
- If M has no starting points then terminate.
- Select a starting point in M.
- Let Y be the set of points that can be reached from a by going forward then Create a cluster containing ∪ a { $M = M/(Y \cup \{c\})$ }
- Repeat the steps until all points are visited.

2.2.2 K-Means:

K-means clustering partitions objects into *k* clusters where each object belongs to the cluster with the nearest mean [14].This produces *k* different clusters. The best number of clusters *k* which leads to the greatest distance is not known as a priority and must be calculated from the dataset. The aim of K-Means clustering is used to minimize the squared error function. The algorithm is as follows:

- It clusters the data into k groups. k is predefined
- Selects k points at random as clusters.
- Assigns objects to their closest cluster in accordance with the Euclidean distance function
- Calculates the centroid or the mean of all objects in each cluster.
- The steps are repeated until no data point is reassigned.

2.3.3 Silhouette Coefficient:

Silhouette coefficient has been used to find the efficiency of the cluster formed and helps to form more efficient clusters.[17]
The silhouette coefficient of a data point i is calculated from Eq(1)

Where a(i) is the silhouette coefficient; b(i) is the difference of the data i with all other data in the same cluster; c(i) is the difference of the data i with the closely associated cluster.
The silhouette coefficient of a cluster can be calculated by finding the mean of all the data points in the cluster.
The silhouette coefficient can take values between – 1 and 1. A value closer to one is considered better for the clusters formed.

2.3 Regression:

This section discusses on the different regression technologies used in our work.

2.3.1 Linear regression:

The most straight forward and easy to use regression for any model is linear regression. A linear regression

consists of two fields a dependent value and an independent value. Dependent values can be Dacoity, rape, etc. Year as independent values. Equation of a straight line Eq(2)

2.3.2 Ridge regression:

Linear regression can sometimes have much deviation from the original graph to improve the accuracy. Hence we use ridge regression. To avoid the under fitting and over fitting we use biasing and variance to enhance the accuracy of our model. Under fitting occurs when algorithm does not fit the data very well and over fitting occurs when a model learns the data in detail in a way that it has negative impacts on the performance of the model. Python code is used to find different values of alpha, order to overcome errors and for better accuracy of the crime that is predicted. Alpha was used with values 15, 0.5 and 0.001 for dacoity, murder and rape respectively.

2.3.3 Naïve Bayes:

Naïve Bayes algorithm is a powerful classification algorithm. It works on conditional probability which is used to calculate the probability of an event by using its earlier knowledge. It is faster than many other algorithms. It uses Bayes theorem for classifying different classes of data.
The basic formula for Naïve Bayes is Eq(3)

Where P(X|Y) is the probability of X when an event Y has already happened.
P(Y|X) - probability of Y when an event X has happened.
P(X) - probability of event X.
P(Y) - probability of event Y.

There are three classification of Naïve Bayes algorithm.

1) Gaussian distribution
2) Multinomial
3) Bernoulli

For our case Gaussian distribution is preferred because our model follows a normal distribution and has continuous values. Multinomial and Bernoulli are ruled out because Multinomial is used for discrete counts and Bernoulli is used for if my feature vector are binary i.e. 0 or 1.

In Gaussian distribution we have taken year as our independent variable and crime like Murder, Dacoity, rape, etc. as our dependent variable. We pass it to the formula of Gaussian distribution and calculate the accuracy.

3. Tables, Figures and Equations

3.1 Tables and Figures

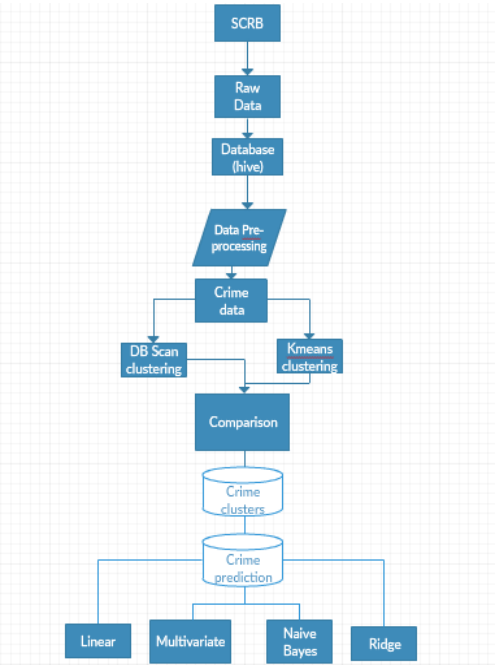

Fig 1. Flow diagram

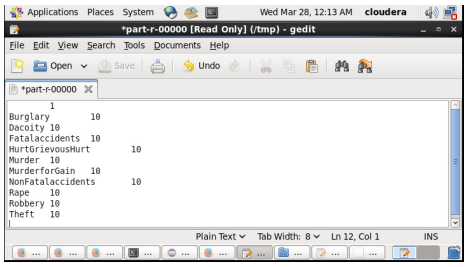

Fig 2. Map Reduce

Fig 3. Word Cloud of Crime dataset

Clustering:

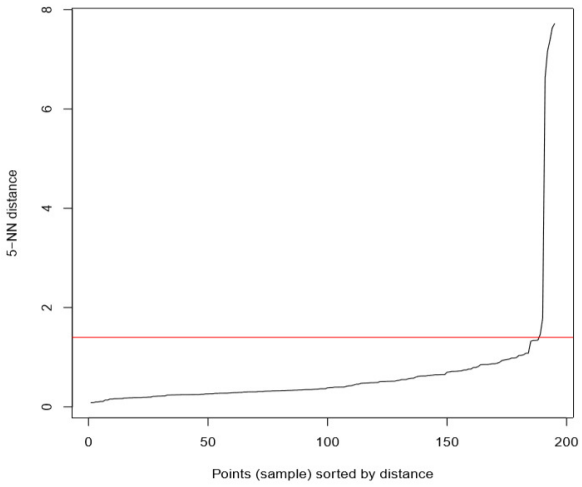

Fig 4. Distance threshold for clustering

DBSCAN:

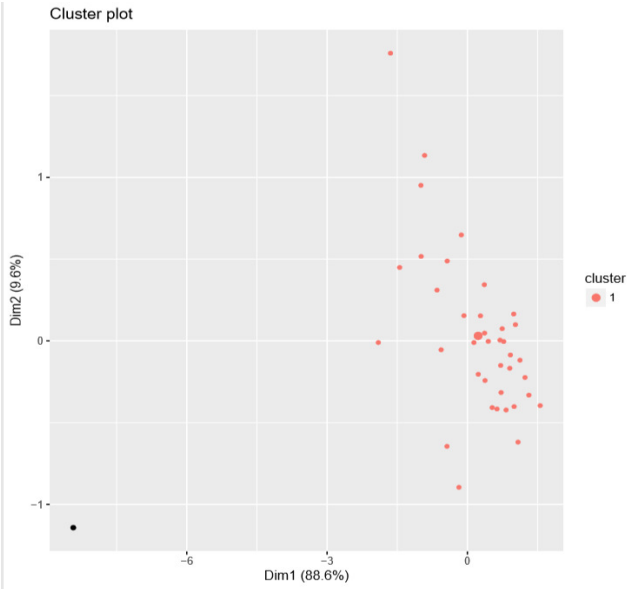

Fig 5. DBSCAN clustering

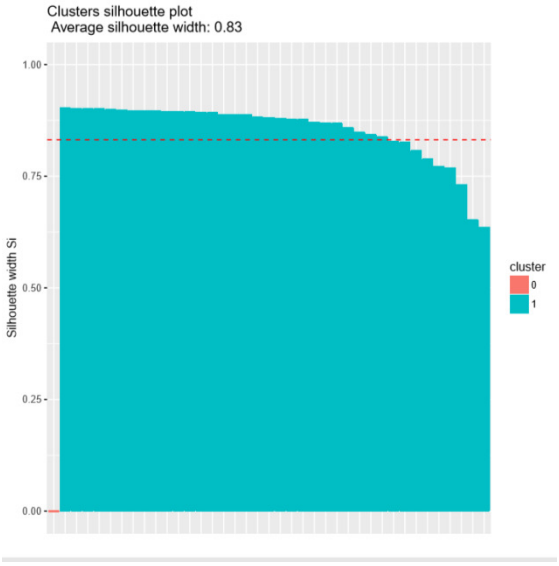

Fig 6. Silhouette coefficient for DBSCAN

Table 1- Types of clustering and its error values:

CLUSTERING	SILLHOUTTE COEFFICIENT'S ERROR VALUE
DBSCAN	0.85
K-MEANS	0.08 ,0.31 ,0.34 ,0.44

K-Means:

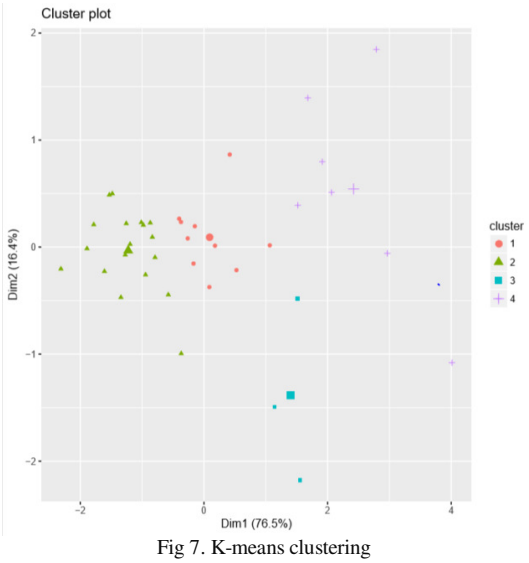

Fig 7. K-means clustering

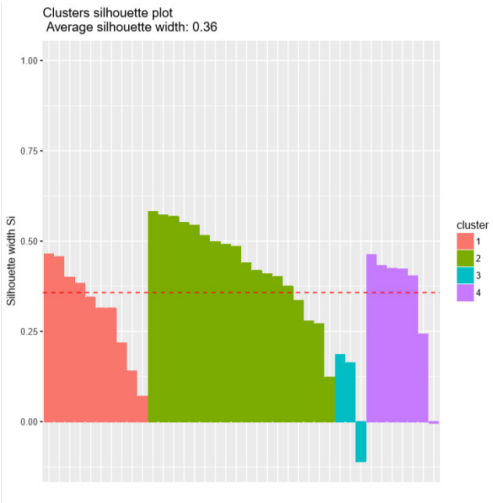

Fig 8. Silhouette coefficient for K-means

Regression:

Linear regression:

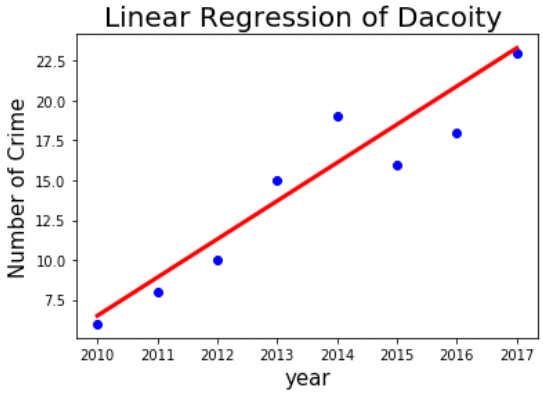

Fig 9. Linear regression for dacoity

Ridge regression:

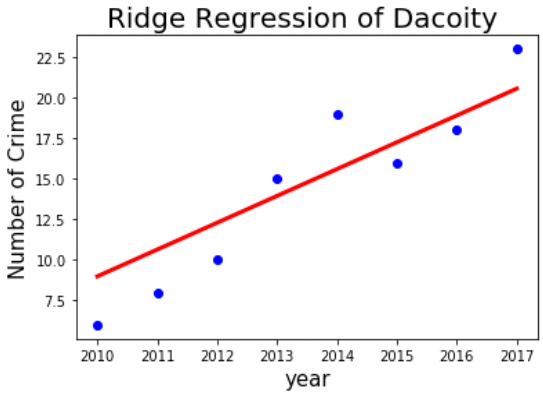

Fig 10. Ridge regression for dacoity

. Naïve Bayes regression:

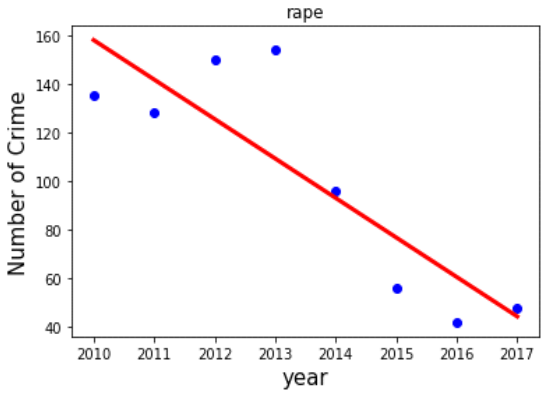

Fig 11. Naïve Bayes regression for rape

Table 2- Types of regression and its error values

REGRESSIONS	R-SQUARED ERROR VALUE
LINEAR	0.6843
RIDGE	0.8095
NAÏVE BAYES	0.9108

3.2 Equations

$$a(i) = \frac{b(i) - c(i)}{\max\{b(i), c(i)\}} \quad (1)$$

$$\mathbf{Y\,(i)\ = mx + c} \quad (2)$$

$$\mathbf{P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)}} \quad (3)$$

4. Conclusions:

Table 1 shows the accuracy of DBSCAN clustering and K-means clustering using Silhouette coefficient. The

clustering methods are implemented and their performance is tested based on accuracy. On comparing their performance the DBSCAN clustering has high accuracy for the given dataset and forms effective clusters. Table 2 shows Linear, Ridge and Naïve Bayes regressions and their corresponding R- squared error value. On comparing the different r^2 values and considering the accuracy of our model, naïve Bayes regression show better results as the values are closer to 1. Thus, this system will help law enforcing agencies, police officials and general public in enforcing laws and providing necessary protection in areas that are vulnerable to crime.

In future, this work can be extended to have improved regression algorithms along with Neuro-linguistic programming (nlp) which helps *in* natural language understanding and recognition to analyze and predict criminals more efficiently.

References:

[1] S.Sivaranjani, Dr.S.Sivakumari, Aasha.M. Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches

[2] Nelson Baloian, Col. Enrique Bassaletti , Mario Fernánde , Lt. Col. Oscar Figueroa, Pablo Fuentes, Raúl, Manasevich, Marcos Orchard, SergioPeñafiel1, joseA.Pino1, Mario Vergara Crime Prediction using Patterns and Context.

[3] Xiafenwang, Mattew S. Gerber and Donald E.Brown. Automatic Crime Prediction using Events Extracted from Twitter Posts.

[4] Nurul Hazwan Mohammad shamsuddin, Nor Azizah Ali, Razana Alwee. An Overview on crime prediction Methods.

[5] Mohammad A, Tayebi, UweGlasser, Patrica L. Brantingham. Learning where to inspect: Location learning for crime prediction

[6] Coral Featherstone. The relevance of social media as it applies in South Africa to crime prediction

[7] https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

[8] https://www.datascience.com/blog/k-means-clustering

[9] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[10] http://www.statsoft.com/Textbook/Naive-Bayes-Classifier

[11] https://onlinecourses.science.psu.edu/stat857/node/155

[12] https://study.com/academy/lesson/what-is-criminology-definition-history-theories.html

[13] https://www.simplilearn.com/introduction-to-mapreduce-quick-tutorial-video?utm_source=google&utm_medium=cpc&utm_content=dsa&utm_term=&utm_campaign=search-dsa-generic-in-adgroup-all-webpages&gclid=EAIaIQobChMI7bW05Iiv2gIVkyQrCh1vgAwcEAAYASAAEgImg_D_BwE

[14] http://bigdata-madesimple.com/what-is-clustering-in-data-mining/

[15] http://www.comp.dit.ie/btierney/Oracle11gDoc/datamine.111/b28129/regress.htm

[16] http://www.tnpolice.gov.in

[17] G pavai T.V geetha new crossover operators usng dominancea and co-dominance priciples for faster convergance and genetic algorithm