Crime detection using k-means and facial recognition

[1]Dr. Pradeep Mohan Kumar, [2]S Athul Sriram, [3]Arnold John Regis

***ABSTRACT**—The influx of massive amount of data being readily available in the modern world enables us to provide an alternate method of crime detection using data mining. Data mining is the practice of examining pre-existing databases in order to generate new information. Many large corporations and businesses are implementing data mining algorithms in view of detecting potential intrusion, fraud or even crime. Even in the year 2020, many organizations still identify intruders by their outer look or by other sensitive attributes like gender, race or religion. Making decisions for criminal activity based on such sensitive attributes will not help to actually detect possible criminals and in fact encourages various kinds of discrimination. A possible alternate system could detect the objective misbehavior of a potential criminal, rather than using irrelevant information like gender, race or religion. Legal data stored in crime databases will be used instead of sensitive attributes. Legal data includes behavior of past crimes done by the individual. If the system identifies any suspected intruder, system will mine the data in database. If the person is detected in the database, their data record will be examined to determine whether it is a harmless person or a potential criminal.*

Keywords—*crime, data mining, recidivism, k-means, clustering, facial.*

I. INTRODUCTION

There is no real consistency to the occurrence of crime. However, it is not entirely random in terms of the criminals themselves. The existing systems to deal with crime have failed to deliver results in the consistency expected from them. There thus exists a poor condition in such a dangerous and important field. Data mining is a powerful tool with great ability to help criminal investigators improve their efficiency, if properly used on the most relevant parts of crime data. The choice of technique has far reaching consequences in the result of data mining. As the use of technology permeates into the field of crime detection, computer data analysts have started working with law enforcement officers and detectives to hasten the crime solving process. A suspect is an individual that is believed to be responsible for committing a crime. The suspect may be an identified or an unidentified individual. The suspect is not considered to be a convict until proved guilty, which is a vital part of the criminal justice system. A victim is one who suffers due to the crime. Generally, victims are also the ones who report the crime. The crime may also have some witnesses. Recidivism is the tendency of a convicted criminal to reoffend, and is at the heart of the use of previous crime records as a primary component in the data mining process. As a data miner working with crime data, an analyst must deal with issues regarding the private nature of sensitive crime based data, so that

[1] *Department of Computer Science,SRM Institute of Science and Technology,Kattankulathur, India, pradeepk@srmist.edu.in.*

[2] *Department of Computer Science,SRM Institute of Science and Technology,Kattankulathur, India ,athulsri@gmail.com*

[3] *Department of Computer Science, SRM Institute of Science and Technology, Kattankulathur, India,ajregis98@gmail.com.*

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

data mining modeling process does not interfere with the legal boundaries with respect to crime data. In crime terminology, a group of crimes in a geographical location is termed as a cluster, while in data mining terminology, a group of data points with similar vale and characteristics is called a cluster. Clustering algorithms in data mining work on the basis of of identifying groups of records that are similar among the members of that cluster, but different from the rest of the data. The issues with crime patterns are concerned with finding and predicting possible crimes. Crime rate is inconsistent, and the crime patterns are always changing. Due to this, the behaviors in crime are difficult to be explained and predicted.

II. STATE OF THE ART

A. Crime Analysis Using K-Means Clustering

The use of k- means clustering for a qualitative and quantitative view of various crimes is illustrated. Crime rates are calculated based on each type of crime as well as location.

B. Crime Pattern detection using Data Mining

Data mining has the scope for wide use in the field of crime detection. This paper uses certain clustering techniques to cluster crime data to find the presence of possible patterns in crime

.

C. A Survey of Data Mining Techniques for Analysing Crime Patterns

To increase efficiency of crime detection, it is important to choose the appropriate data mining algorithm with respect to the task required. This paper provides a comparison between various data mining techniques.

D. An Analysis of Data Mining Applications in Crime Domain

A large number of different, yet efficient methods for data mining in crime data analysis are presented. These techniques involve finding the illegal activities of professional fraudsters using knowledge obtained through the use of their histories and records. Detecting crime through the analysis of data can be difficult because a large amount of data is generated by the activities of criminals, which exceeds the data which can be efficiently processed. In addition, the quality of data analysis depends greatly on background knowledge of analysis.

E. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences

The automatic analysis of the human face with regard to expressions is an interesting problem with a wide variety of applications. Majority of the existing systems for facial expression analysis use emotional cues in the process of determining the overall expression in the face as happy or sad, or some other relative expression.

III. PROPOSED WORK

The major facets of the system each involves the use of a single modules. These modules are integrated together to form the overall system architecture and provide the functioning required for a crime detection system.

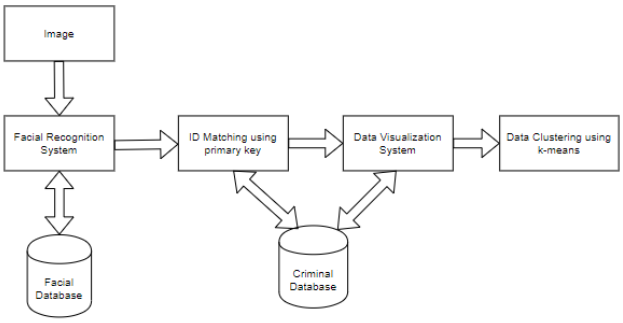
Image
Facial Recognition System
ID Matching using primary key
Data Visualization System
Data Clustering using k-means
Facial Database
Criminal Database

Figure 1: Architecture of a data based crime system

F. Facial Recognition

A facial recognition system is a system capable of identifying an individual, based on the facial details alone from an image or a frame taken from a video. The exact working of a facial recognition system is generally based around the comparison of certain facial features in the captured image to the images available in the database. A face has many different and distinguishable features. In facial recognition, these features are defined as nodal points. Each human face has roughly 80 nodal points. Some of these features which are measured by the software are the

- Width of the nose.
- The distance between the eyes.
- The depth of the eye socket.
- The shape of the cheekbones.
- The length of the jaw line.

G. Data Representation

The information available in the dataset has a variety of fields, each providing a different kind of information. For an easier understanding of this data, we can use graphical representations. The graphical representation of data allows easy understanding by any individuals, regardless of prior knowledge and experience. It also allows a quick and easy analysis and comparison of a particular data value in a column against the other values.

H. Data clustering using K-means

Data Mining can involve the clustering of data to detect interesting patterns. There are a number of clustering algorithms. K-means clustering is a simple and popular unsupervised algorithm used in machine learning . Data clustering comprises the grouping of similar data points together and the discovery of underlying patterns. A cluster is a collection of similar data points aggregated together due to their relative closeness in characteristics when compared to data points outside the cluster. A target value k is chosen, which refers to the number of centroids required in the data set. A centroid is a unique location representing the center of each cluster.

The working of the K-means algorithm requires the identification of k number of centroids, which are then used to allocate every data point to a cluster based on the distance from the centroid, while minimizing the value of the centroids. The means in the K-means refers to finding the average of the data, which is the centroid.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

IV. IMPLEMENTATION

I. Overview

The proposed system uses several different concepts tied in with the field of data mining and machine learning, with the implementation and result of each individual module noted.

J. Feature Definition

The current design uses data from Kaggle.com, and the data set has been chosen from the resource available through use of the mentioned web application.

The data set chosen has 8 attributes upon which the various algorithms are tested and trained. These attributes include front facing images, each with an unique identification value. This value is used to map the image to the data set containing data such as height, weight, crime committed, race, eye colour, hair colour and sex.

K. Data Clustering

According to data clustering, data items which have similar properties and values are considered to be overall more similar to each other than with respect to values found further away in term of the distribution of values. A set of similar values grouped together is termed as a data cluster. Data Clustering is a technique used to detect these clusters in the given data set. There are many algorithms which can be used to accomplish the objective of data clustering, such as the K-means algorithm.

L. K- means Algorithm

The K-means Algorithm is a data clustering algoithm which works on the principle of selecting of k values to serve as the centre of independent clusters, and working around these values.These k values are termed as centroids, and is the value representing the centre of the cluster. In the use of the K-means algorithm, the centroids are initially chosen, and the values are then assigned to clusters based on the proximity of the chosen value to the values of the centroid. The calculation of the distance from each centroid is accomplished by calculating the Euclidean distance between two points in space.Euclidean Distance(d)

$$d= [(x_2-x_1)^2+(y_2-y_1)^2+\ldots+(z_2-z_1)^2]^{1/2}$$

Where x_1,y_1,z_1 represent coordinates of the first point,

x_2,y_2,z_2 represent coordinates of the second point.

M. Data Visualization

Visualization of data is an analytical skill exercised in machine learning and data science. It provides a platform to generate a graphical representation of the information provided in the data set. This enables a quick and easy method for even a layman to understand the variance in the data being used for processing. The relationship between various values in a specific entity of the data set can be highlighted using data visualization. Numeric or categorical data can be represented using data visualization. The choice ofe the data type does not influence the efficiency with which it can be represented, but rather instructs the manner in which visualization occurs, and the nature of the values for which data is represented under some specific field or category.

The representation of two of the columns of data is shown below.

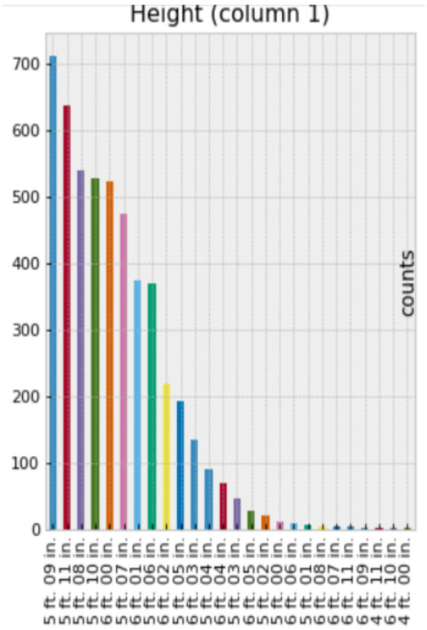

Figure 2: A graphical layout showing variation in height based on numeric values

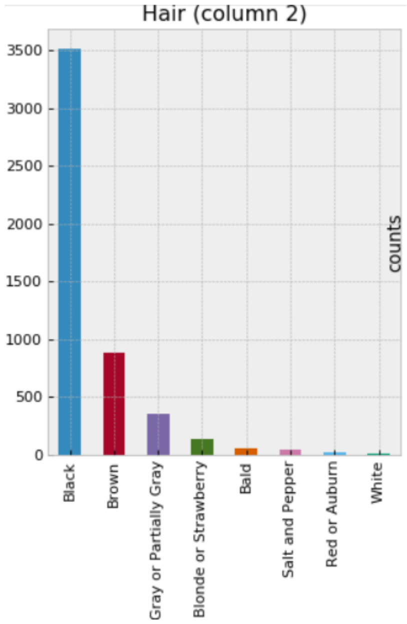

Figure 3: A graphical representation showing variation in hair colour of the criminals

V. FUTURE WORKS

There exists a wide variety of possible improvements in the use of technology and specifically data science in the field of crime. The facial recognition system can initially be optimised to ensure the delivery of quick results, regardless of the size of the data set being analysed. A major field of growth in this field is the reworking of the manner in which data is being made available to the public. While the field of crime undoubtedly has a large aomunt of sensitive information which cannot be divulged, it is also exceedingly difficult to work on improving the system to deal with crime if there is an absence of accurate data to work with. Research into the use of newer algorithms to tackle the task of analyzing the crime records may help improve the productivity of the crime system as well.

VI. CONCLUSION

There is a vital role available for the data science in the field of crime, but we must ensure it is not compromised by the absence of accurate crime data to be used for designing the right system to tackle the major issue of reducing crime.

REFERENCES

1. Crime Analysis Using K-Means Clustering -Anant Joshi ,A. Sai Sabitha & Tanupriya Choudhury.
2. Crime Pattern Detection Using Data Mining - Shyam Varan Nath
3. A Survey of Data Mining Techniques for Analyzing Crime Patterns - Ubon Thongsatapornwatana.
4. An Analysis of Data Mining Applications in Crime Domain - P. Thongtae & S. Srisuk
5. A review : Crime analysis using Data mining techniques and Algorithm - Chhaya Chauhan & Smriti Sehgal
6. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences - M. Pantic & I. Patras

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020