PREDICTING THE RATE OF CRIME USING KNN ALGORITHM

DEWASHISH KUMAR1, ABHINASH KUMAR2, DR. S. GOWRI3, DR. A. SIVASANGARI4, MS. J. S. VIMALI5, DR. G. ROSLINENESA KUMARI6
1, 2, 3, 4 Sathyabama Institute of Science and Technology, Chennai
5 Shadan women's college of Engineering and Technology, Hyderabad

ABSTRACT

Crime is a common societal problem that has an effect on people's personal happiness and economic growth. It is regarded as a vital factor in deciding whether or not people should migrate to another city and what places should be avoided while travelling. With the rise of violations, law enforcement agencies are requesting more advanced regional data systems and modern information mining methods to help them better investigate corruption and protect their organisations. Despite the fact that violations can occur anywhere, it is essential that when criminals are looking for opportunities to commit wrongdoing, they search in the most common places. We hope to raise people's awareness of dangerous places in different eras by providing an information mining method for determining the most criminal hotspots and discovering the type, place, and time of committed wrongdoings. As a result, our proposed arrangement could allow people to avoid certain areas at particular times while also saving lives. Getting this kind of knowledge will also assist people in making better choices on where they want to live. Police authorities, on the other hand, may use this solution to increase the degree of wrongdoing prediction and aversion. It will be beneficial for the allocation of police assets. We want to make it possible by making the bulk of this data accessible.

I. INTRODUCTION

Crimes are usually motivated by a societal grievance that has an effect on society and, as a result, on the culture's profit-making process. Criminality is a major danger to society. There are several crimes that occur at regular intervals. It's possible that it's growing and spreading at a rapid and broad scale. Crime can occur anywhere, from a small village to a large city. Robbery, murder, assault, battery, false arrest, abduction, and homicide are examples of various types of crimes. Since crime is on the increase, it is important to settle crimes even more quickly. The rate of rise in illegal activity has accelerated, and it is the sole responsibility of the police department to monitor as well as minimise it. Since there is such a large volume of crime reports, the police department's biggest issues are crime prediction and criminal identification. There is a need for technology that will allow for quicker case resolution.

Computer-based intelligence is a subfield in programming design, but it's often referred to as perceptive analysis or informative demonstrating on occasion. Its goal is to build new or potentially impact existing counts to take from data in order to construct generalised models which offer precise desires, or to find methodology, especially with the new and obscured large set of data. A dataset is usually divided into a few subsets at the start of an AI adventure. The preparation and test datasets are the fundamental subsets, with an optional third endorsement dataset added on top. Whenever these types of data subsets which are generally are generated from the basic dataset a judicious model or classifier which is readied using the proper planning of the data, and in a short span of time later the model's insightful exactness is settled using the test data.

II. LITERATURE REVIEW:

Surprisingly, the most meticulous attempts at wrongdoing mapping can be found all the way back to the roots of criminology by own. Various investigations into the circulation of corruption in France and England were conducted in the mid-nineteenth century. Branting ham and Branting ham (1991) provided an insight of some of the finding made during the fundamental investigations at the time being. Guerry and Quetelettogether mapped abuses at the division level in France and discovered that the wrongdoing were not spread uniformly through departments. They also discovered that in the long run, there was soundness in the two areas with high
wrongdoing and regions with low wrongdoing. Plint, Glyde, and Mayhew conducted investigations in England in response to these discoveries. In the United States, Shaw and McKay (1942) used wrongdoing maps widely in their fundamental investigation of teenage wrongdoing in Chicago. Shaw and McKay used the biological model of Park and Burgess (1924) to divide the region into five unique areas. They found that the zone adjacent to the focal business district, the zone of growth, consistently faced the highest rates of adolescent misconduct as well as other social problems, regardless of the ethnic group that occupied the zone at the time. This study aided in the advancement of the social disruption hypothesis and sparked a number of comparative mapping projects in various cities as Chicago, Philadelphia, Richmond, Virginia, Cleveland, and various cities, which were highly boring and only offered a clear understanding of wrongdoing designs. The most significant advancement in the area of wrongdoing mapping has been the use of GIS programmes for mapping them. There are few important advantages of using virtual maps instead of real maps. PC’s have greatly reduced the amount of time and effort which is needed to construct wrongdoing maps. Given the relatively low effort and the ease of use of a large number of these product programmes, engaging in wrongdoing mapping would never again entail a substantial investment. Second, these GIS systems reduce the risk of making a mistake while giving wrongdoing events regional direction. Thirdly, the virtual maps are much more adaptable than physical maps, which allows experts and wrongdoing investigators to compare between the spatial distribution of violations to various characteristics of the area under investigation (e.g., registration department data, city arranging and zoning the maps, and maps delivered by different organizations). Finally, GIS and the other spatial investigation programming provide useful observable instruments for analysing and identifying the instances of criminal activity which cannot be detected and identified by a simple visual inspection.

III. SYSTEM ARCHITECTURE & METHODOLOGY

The methods used in crime rate prediction are:

- Data Collection
- Data Pre-processing
- Using KNN Algorithm
- Result

3.1 DATA COLLECTION:

We have chosen a set of dataset for predicting crime rates, and we'll use the necessary attributes for training our model on it. We obtained the data from various online sources. The dataset's form is 878049×9. The different columns in our dataset depict the specific date on which the crime was reported, the types of crimes, the summary...
or significance of the crime, the days on which the crime occurred, the district, and the address where the crime occurred.

3.2 DATA PREPROCESSING:
The process of transforming raw data into the appropriate format is known as data preprocessing. Data cleaning, data reduction, and data integration are all examples of data preprocessing methods. The data is gathered from the different sources, each with its own collection of attributes and data. We transform the raw data into the appropriate format because some data is replicated and duplicated, contains some attributes that is not for our interest, and contains NULL values.

3.3 CLUSTERING ALGORITHM:
Because of the variety of applications, clustering algorithm has become and gained a rising research problem in the field of data processing. Many data and their meaning have been implemented by cluster algorithmic programmes in recent years, and they are now used in a broad range of applications such as image manipulating, processing micro-objects, visual converse, narcotics, and political economy, which contributes to the quality of the algorithms. The main disadvantage of cluster algorithms is that they cannot be ordered. Also, the algorithmic programme progress will produce spectacular result when used in conjunction with an information set, but it will not produce accurate results. Clustering algorithms are mostly unsupervised learning algorithms. It's more like data pre-processing. They are, in reality, updated on a daily basis. Only a few people are capable of working with data. It is extremely helpful at the very beginning of every module, also known as data pre-processing. Whenever there is any chance for clustering as many data items as possible, clustering in data mining is a very crucial algorithm as the result which is going to be obtained is in the form of scalar.

Steps to take when using the K-means algorithm are as follows:

Assume that X represents the data clusters as a vector and the V represents the points of centres as sets.

1) Clustering centres are denoted by the letter 'c.'
2) Measure the distance between cluster centres and each data point.
3) Assign the information objective to the cluster centre with the shortest distance from all other cluster centres.
4) Build a new cluster core using the following formula:
   \[ v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \]
   Where, 'ci' represents the amount of data in the ith cluster.
5) Measure the distance between each cluster centre and each data point.
6) Repeat from step 3, if no data points are assigned.

K-means is one of the most effective methods for clustering problems. The k clustering centroids that can be used to mark new data entries. This approach uses a specific clustering technique (assume k clusters) to classify a given data set. The first step is to designate k centres, each for a single cluster. Because of the absolute difference in position and the completely different outcome, the centres should be positioned correctly. The best choice is to position them as far apart as possible from each other clusters. The next step is to assign each aim happiness to a specific data set and link it to the nearest centre. The next move is taken until no intention remains unfulfilled. We tend to recalculate k new centroids at the moment . There is a loop formed. As a result, the k centres change the positions one by one until all of them have been changed. Subsequently, this formula focuses to minimise associate objective function which grasp as a square error function, which is given by:
\[ J(V) = \sum_{l=1}^{c} \sum_{j=1}^{c_l} \left( ||x_i - v_j|| \right)^2 \]

where, The Euclidean interval between Xi and Vj is given by (||Xi-Vj||).

‘c’ denotes the total number of cluster centres.

3.4 SYSTEM CLASSIFICATION:

Following is an outline of the acknowledged ML techniques –

- Regulated Learning
- Unaided Learning
- Semi-managed Learning
- Support Learning

Regulated Learning:

In different cases, we can prepare the calculation for these as the marked models as an information where the ideal outcome is already known in advance. Along with contrasting exact outcomes, a learning estimate can yield a lot of knowledge guidelines. If any type of error occurs, the genuine outcome is being contrasted by the learning calculation which conflict with exact result and also shows us the warning. Driven learning uses a variety of techniques, such as relapse, arrangement, angle boosting, and prediction, to proactively validate the estimation of a name on additional unlabeled type of data. This is the main reason why this approach is widely used in environments where prerequisite knowledge is used to forecast events which is most likely occur in the future. Consider predicting whether a credit card transaction will most likely be fraudulent, or which insurance clients will most likely file claims.

Unaided Learning:

This machine learning methodology discovers its use in places where knowledge is undocumented. The framework will not be provided with the "valid answer" in this case, and those of calculation must discern what is being shown. Thus, main goal here is to look over the data and thus find an example as well as structure within the available information indexes. For instance, this type of learning figures out client segments with most similar characteristics and identity and then helps the organisation to treat them equally in marketing efforts. Also, it can discern properties that differentiate client fragments from each other. In any case, it has to do with identifying a similar structure in the available informational set. Additionally, these calculations may figure out the exceptions which are available in informational indexes.

Semi-Managed Learning:

Semi-managed learning is used and implemented in almost similar circumstances where the supervised learning is required. Nonetheless, it's worth noting that this strategy prepares using both unlabeled and labelled data. In an ideal world, a small set of named data will be used alongside a large amount of unlabeled data, so getting unlabeled data requires less time, money, and effort. This form of AI is commonly used in combination with techniques such as relapse, characterization, and expectation. Organizations that believe it would be difficult to meet the substantial costs associated with a structured learning process opted for semi-directed learning.

IV. RESULT:

As a result, the model describes how the KNN clustering algorithm is helpful in detecting crimes and its types. The crime hotspot monitoring system makes it simple for police officers to investigate various crimes in a specific area. The modern form of live streaming generates a vast volume of data that must be clustered to be managed.

V. CONCLUSION:

The content given by the dataset for classification is the only basis for crime rate prediction. Many classification algorithms have been used to predict accuracy, but the KNN algorithm predicts more accurately than the others. As a result, this paper concludes that using the KNN algorithm to forecast crime rates is more reliable.
REFERENCE:


