Crime Prediction Using Fuzzy c-means Algorithm

Shuchita Mishra[1],Tanvi Pradhan[2],Priyanka Parmar[3],Suvarna Maji[4],Ekta Sarda[5]

[1,2,3,4]Student [5]Assistant Professor
[1,2,3,4,5]Department of Computer Engineering, Ramrao Adik Institute of Technology, Mumbai University,India

---***---

Abstract - *Committing crimes and becoming a victim of such brutal crimes has become unfortunately too easy in today's world and protection from such crimes has become a necessity. Our project aims to curb these high crime rates. We introduce data mining and clustering algorithms to predict the occurrence of crimes. Fuzzy C-Means is a useful technique to cluster offenders, identify various crime patterns, and analyze crime data. Data mining algorithms extract relevant and unique information and patterns from crime records. Clustering is done based on location of crime, gangs/offenders who are involved in the said crime and the date and time of the committed crime.*

Key Words: **Data Mining, Fuzzy c-means, Clustering, Prediction**

1. INTRODUCTION

Crime is the prime concern of this paper. Crime can take place at any time and in any region of the country. Most of the law enforcement agencies are focused on creating a tool through which future crime location can be detected. These tools are based on the large collections of data. Crime records are first designated based on the type of crime, then analyzed after which important crime hot-spots are identified and lastly, pre-diction and prevention of future crimes is performed. These steps are very needful in crime analysis.

2. LITERATURE SURVEY

Existing papers and techniques essentially talk about the various tools which can be used for data mining [3]. They give an idea about fuzzy set theory and fuzzy clustering techniques. Many papers have reviewed several clustering techniques that have been studied for Criminal Profiling [1] and have come up with the result that Fuzzy C clustering is a better technique to predict and prevent criminal activity.

A Fuzzy C-Means based clustering perspective is proposed to find similar sub sets from the crime data [2]. When crime records are handled manually, sometimes the charges may be ignored by the police, resulting in inaccurate data collection. Because of the large number of serious crimes, minor complaints may be ignored.

Our proposal is to develop a system of improved and enhanced features. In our system, we perform data clustering using Fuzzy C-Means (FCM) algorithm, performing location based clustering and predicting the crime details. The proposed system aims at overcoming the limitations of the existing system. The system aims to provide proper security and reduce manual work.

3. PROPOSED SYSTEM

Data clustering can be done in various ways. In this paper, we have compared two most widely used clustering algorithms: K-means and Fuzzy c-means algorithms.

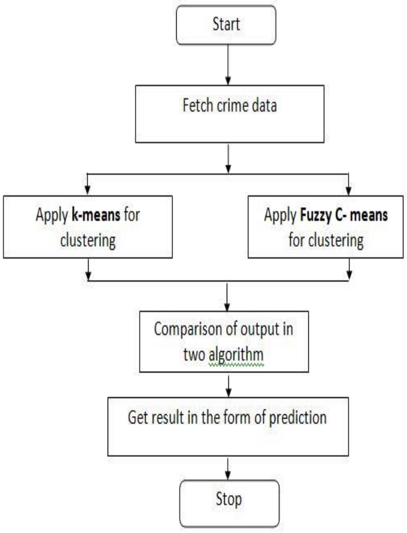

Start
Fetch crime data
Apply k-means for clustering
Apply Fuzzy C- means for clustering
Comparison of output in two algorithm
Get result in the form of prediction
Stop

Fig 1. Flowchart of proposed system

3.1 Proposed Methodologies

There are various methods and systems in place to collect and analyze data. One such method is the Fuzzy system which has been used in this paper.

This is one of the most commonly used tools for the data collection and filtering. The filtering mechanism will be accomplished by the use of rules of fuzzy. These rules are based upon the IF-THEN rules [3]. The conditions are specified within the IF condition. The conditions if satisfied then result will be obtained. If the condition is false then the result will not be obtained. Fuzzy sets are created in this case. Membership functions are also needed to be defined. The membership of the given data will be done by determining whether membership function result in 0 or 1. Fuzzy system consists of input stage, processing stage and then output stage. The fuzzy rules will be represented as If temperature is "low" THEN heater is "High"

Various truth values are used within the fuzzy system. These truth values are used in combination with AND, OR etc. if we require that both the conditions must be true then AND will be used. If either of the conditions can be true to produce the result then OR can be used.

3.2 K-means clustering

K-means is a technique of clustering records into a determined number of unique clusters. The"K" refers to the number of clusters specified. There are various distance measures which determine the cluster to which an observation is to be assigned. This algorithm strives to minimize the measure between the centroid of the cluster and the given observation by iteratively assigning an observation to a cluster and terminate the loop when the smallest distance measure is obtained.

Following is the overview of the algorithm:

1. The sample space is initially partitioned into K clusters and the data elements are randomly assigned to the clusters.

2. Then for each sample:

A. The distance from the element to the centroid of the cluster is evaluated.

B. IF the sample is closest to its own cluster THEN leave it ELSE select another cluster.

3. Repeat steps 1 and 2 until no elements are moved from one cluster to another. When step 3 terminates the clusters are stable and each sample is assigned a cluster which results in the lowest possible distance to the centroid of the cluster.

Common distance measures include the Euclidean distance, the Euclidean squared distance and the Manhattan or City distance. We use the Euclidean distance to measure the distance between an element and the centroid.

The squared Euclidean measure corresponds to the shortest geometric distance between two points.

A faster way of determining the distance is by use of the squared Euclidean distance which calculates the above distance squared.

3.3 Fuzzy C means clustering

In FCM, each data point is assigned a membership value that denotes the degree of its belongingness to that particular cluster. The addition of all the membership values for each data point should be one.

With fuzzy c-means, the centroid of cluster is computed by:

- Choose a number of clusters.

- Assign coefficients randomly to each point to form clusters.

- The algorithm needs to be repeated until merged (that is, the change in coefficients between two iterations should not be more than the sensitivity threshold value):

- Centroid computation for each cluster.

- For each data points compute the coefficient of it being in that cluster.

The overall procedure consists of three main steps:

1. Clustering the raw data.

2. Producing the membership functions from the data.

3. Creating the fuzzy inference system.

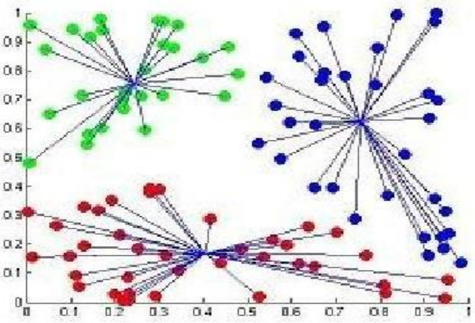


Fig. 2 Fuzzy c means clustering

Let X = {x 1 , x 2 , x 3 ..., x n } be the set of data points and V = {v 1 ,

v 2 , v 3 ..., v c } be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the fuzzy membership using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)} \qquad (1)$$

3) Compute the fuzzy centers using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c \qquad (2)\ 4)$$

Repeat step 2) and 3) until the minimum value is achieved or ||U (k+1) - U (k) || < β.

where,

'k' is the iteration step.
'β' is the termination criterion between [0, 1].
'U = (μ ij) n*c ' is the fuzzy membership matrix. 'J' is the objective function.

3.4 Naïve Bayes classifier

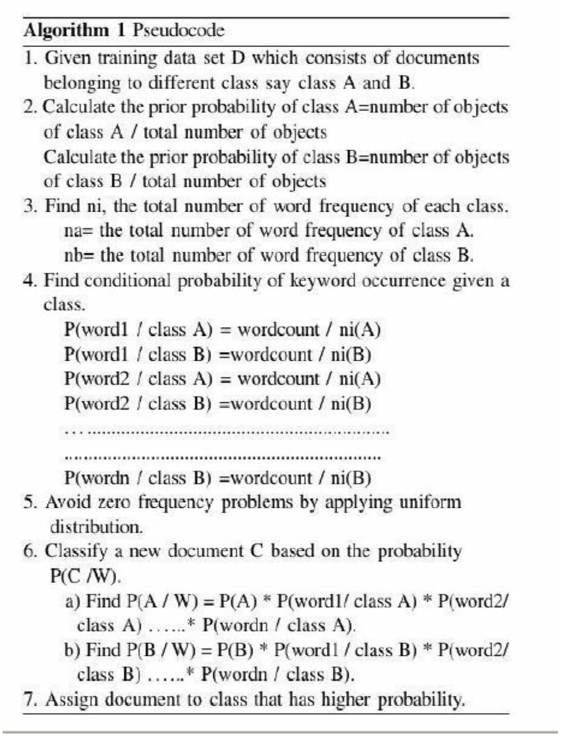
There are various predictive methods that help police departments determine when and where crime will happen before it actually happens. In this analysis, we will focus on a simple predictive task of predicting which gang has committed the crime. For this, we are using the Naïve Bayes [5] approach.

In classification, posterior probability can be interpreted as: "the probability that a particular element belongs to class i_1 given its observed values". It can be expressed in simple words as follows:

Posterior probability = (conditional probability * prior probability) / evidence

Let

- $x_i$xi be the feature vector of sample i,i{1,2,...,n}i,i{1,2,...,n} ,
- ω_jωj be the notation of class j,j{1,2,...,m}j,j{1,2,...,m},
- and P(xi|ω_j)P(xi|ωj) be the probability of observing sample $x_i$xi given that it belongs to class ω_jωj .



```
Algorithm 1 Pseudocode
1. Given training data set D which consists of documents
   belonging to different class say class A and B.
2. Calculate the prior probability of class A=number of objects
   of class A / total number of objects
   Calculate the prior probability of class B=number of objects
   of class B / total number of objects
3. Find ni, the total number of word frequency of each class.
     na= the total number of word frequency of class A.
     nb= the total number of word frequency of class B.
4. Find conditional probability of keyword occurrence given a
   class.
     P(word1 / class A) = wordcount / ni(A)
     P(word1 / class B) =wordcount / ni(B)
     P(word2 / class A) = wordcount / ni(A)
     P(word2 / class B) =wordcount / ni(B)
     ..............................................................
     ..............................................................
     P(wordn / class B) =wordcount / ni(B)
5. Avoid zero frequency problems by applying uniform
   distribution.
6. Classify a new document C based on the probability
   P(C /W).
     a) Find P(A / W) = P(A) * P(word1/ class A) * P(word2/
        class A) .....* P(wordn / class A).
     b) Find P(B / W) = P(B) * P(word1 / class B) * P(word2/
        class B) .....* P(wordn / class B).
7. Assign document to class that has higher probability.
```

Fig 3. Crime prediction using Naïve Bayes algorithm

The general expression of the posterior probability can be denoted as:

$$P(wj|xi) = P(xi|wj)\frac{P(wj)}{P(xi)} \quad (3)$$

Under the naive assumption, the class-conditional probabilities or (likelihoods) of the samples can be directly approximated from the training dataset instead of evaluating all possible values of x. Thus, the class conditional probability of a d-dimensional vector x can be evaluated as:

$$P(x|j) = P(x1|j)P(x2|j)P(xd|j) = k \quad (4)$$

4. COMPARISON OF CLUSTERING ALGORITHMS

From the above analysis we found the following differences in the K-means and Fuzzy C-means techniques:

1. Fuzzy c handles uncertainty condition of the data.

2. Unlike k-means, fuzzy c can be applied on both, numeric as well as categorical data.

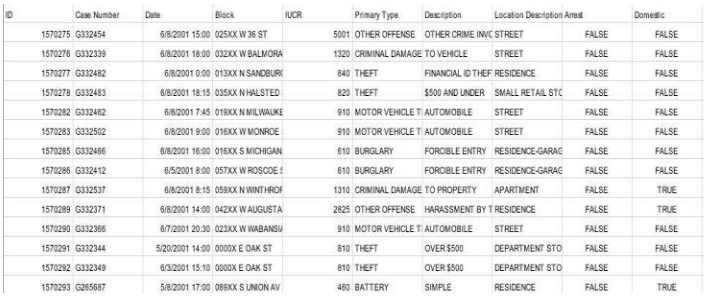
3. In fuzzy c an object can have a membership degree of 0 to 1 whereas in k means clustering, an object can either have a membership degree of 0 or 1.

4. Fuzzy c algorithm takes more time to execute as the number of calculations is extensive whereas k-means is much faster.

5. EXPECTED RESULTS

As per our expectations, Fuzzy clustering is proposed to address the issue in which a criminal can belong to more than one cluster at a time depending on the degree of membership or probability going by the nature of criminals who can belong to as many groups as possible.

The expected dataset shall look something of this kind on the basis of which we shall perform clustering and prediction.



ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic
1570275	G332454	6/8/2001 15:00	025XX W 36 ST	5001	OTHER OFFENSE	OTHER CRIME INVC	STREET	FALSE	FALSE
1570276	G332339	6/8/2001 18:00	032XX W BALMORA	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	FALSE	FALSE
1570277	G332482	6/8/2001 0:00	013XX N SANDBUR(	840	THEFT	FINANCIAL ID THEF	RESIDENCE	FALSE	FALSE
1570278	G332483	6/8/2001 18:15	035XX N HALSTED	820	THEFT	$500 AND UNDER	SMALL RETAIL STC	FALSE	FALSE
1570282	G332462	6/8/2001 7:45	019XX N MILWAUKE	910	MOTOR VEHICLE T	AUTOMOBILE	STREET	FALSE	FALSE
1570283	G332502	6/8/2001 9:00	016XX W MONROE	910	MOTOR VEHICLE T	AUTOMOBILE	STREET	FALSE	FALSE
1570285	G332466	6/8/2001 16:00	016XX S MICHIGAN	610	BURGLARY	FORCIBLE ENTRY	RESIDENCE-GARAG	FALSE	FALSE
1570286	G332412	6/5/2001 8:00	057XX W ROSCOE S	610	BURGLARY	FORCIBLE ENTRY	RESIDENCE-GARAG	FALSE	FALSE
1570287	G332537	6/8/2001 8:15	059XX N WINTHROF	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	FALSE	TRUE
1570289	G332371	6/8/2001 14:00	042XX W AUGUSTA	2825	OTHER OFFENSE	HARASSMENT BY T	RESIDENCE	FALSE	TRUE
1570290	G332366	6/7/2001 20:30	023XX W WABANSI/	910	MOTOR VEHICLE T	AUTOMOBILE	STREET	FALSE	FALSE
1570291	G332344	5/20/2001 14:00	0000X E OAK ST	810	THEFT	OVER $500	DEPARTMENT STO	FALSE	FALSE
1570292	G332349	6/3/2001 15:10	0000X E OAK ST	810	THEFT	OVER $500	DEPARTMENT STO	FALSE	FALSE
1570293	G265687	5/8/2001 17:00	089XX S UNION AV	460	BATTERY	SIMPLE	RESIDENCE	FALSE	TRUE

Fig 4. Dataset for crime prediction system

Based on the issues identified above, the proposed Fuzzy Clustering will examine how real life data can be integrated and analyzed to produce "profiles" of activity and behavior of criminals. The aim is to provide investigators with rich sources of intelligent information which can be used to predict and prevent criminal activity. Fig 5 below reveals the probability of an arrest for a category of crime. After the clustering of the crimes dataset, we can make a graph showing the probability of arrest which will help in predicting the possible perpetrators. It is shown that Narcotics offenses have the highest probability at around a 97 percent chance of a arrest given an offense. Others have lower probabilities of arrest such as Homicide with about a 50 percent chance of arrest. We think that the crimes that are easier to solve have a much higher probability of arrest than crimes that are harder to solve.

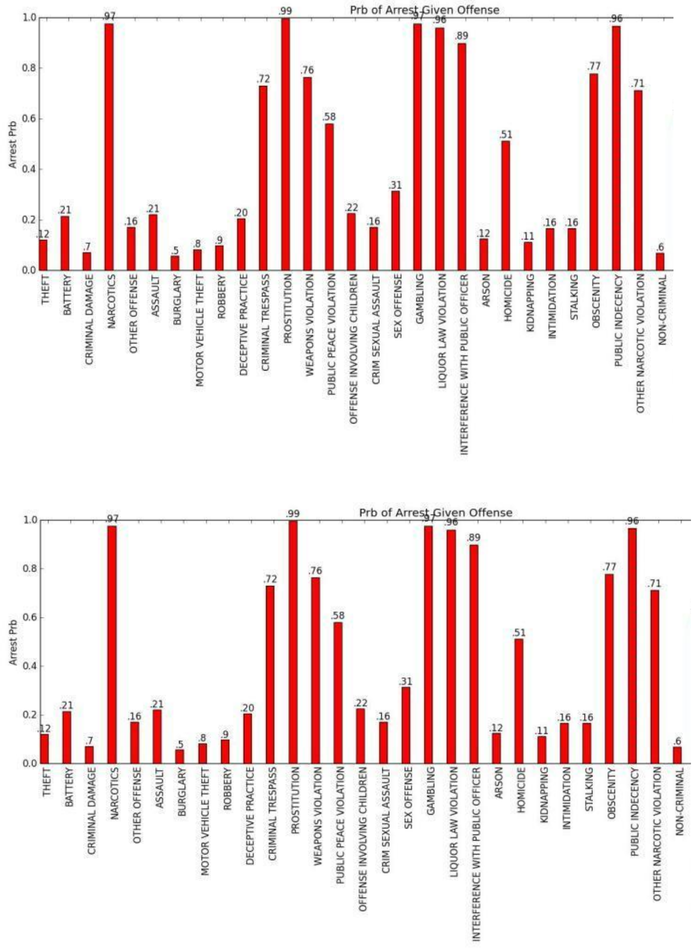


Fig. 5 Probability of arrest for given offense

Our first predictive task was to predict arrests for reported crimes. A prediction can be shown in the following format –

CASE NUMBER	ARREST MADE
HM776500	0

Fig. 6 Prediction result of Naïve Bayes

Where the first column is the case number and the second column is whether or not an arrest will be made for this case number.

6. CONCLUSION

Attribute weightage-based clustering algorithm was developed for finding similar sub sets from crime data. The existing technique was modified in three ways, such as i) new attribute weightage scheme, ii) suitability to mixed data, and iii) Using FCM-based clustering instead of k-means. The performance of the proposed clustering algorithm is analyzed based on clustering accuracy with the K-Means algorithm. From the experimental results it is observed that proposed FCM method has better accuracy than the K-means method. In further enhancement of the system, duplicate crime data detection can be performed as many times crime records might not be carefully organized causing distortions in clustering and prediction. The "Meat" of the record can be extracted as the whole record may be a collection of a lot of data which is irrelevant or something completely unrelated. It will be good for the system to tag the part of the record that is most likely related to the crime (actual record) and those who are unlikely to be related. Furthermore, the less domain knowledge required the better.

In our future scope there could be an implementation of the project on the datasets of Indian crimes and cities.

REFERENCES

[1] A Review of Different Clustering Techniques in Criminal Profiling, Volume 6, Issue 4, April 2016, International Journal of Advanced Research in Computer Science and Software Engineering

[2] An Attribute Weighted Fuzzy Clustering Algorithm for Mixed Crime Data, Vol 9(8), February 2016, Indian Journal of Science and Technology

[3] Data Mining Techniques used in Crime Analysis: - A Review, Volume: 03 Issue: 08 | Aug-2016, International Research Journal of Engineering and Technology (IRJET)

[4] CSE 190: Predictive Policing on Crime Data. Joshua Wheeler, Nathan Moreno, Anjali Kanak

[5] Crime Analysis and Prediction Using Data Mining. Sathyadevan, Shiju & S, Devan & S Gangadharan, Surya. (2014). . 10.1109/CNSC.2014.6906719.

[6] Datasets, Center for Machine learning and Intelligent Systems, Donald Bren School of Information and Computer Sciences, University of California, Irvine. Available from: http://cml.ics.uci.edu/. 25/05/2015.

[7] References from www.google.com