International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072

COMPREHENSIVE COMPARATIVE ANALYSIS OF METHODS FOR CRIME RATE PREDICTION

Omkar Vaidya[#1], Sayak Mitra[#2], Raj Kumbhar[#3], Suraj Chavan[#4], Rohini Patil [#5]

[1,2,3,4,5] *Department Of Computer Engineering, Terna Engineering College, Mumbai University, India*

Abstract - *Crime analysis and prediction is a systematic approach for analyzing and identifying different patterns, relationships and trends in crime. The system can predict regions which have high probability for occurrence of crime and indicate crime prone areas. It will be beneficial for the law enforcement agencies to speed up the process of solving crimes with the increasing advent of computerized systems and with the help of crime data analysts. The previously unknown but useful information from an unstructured data can be extracted by using the concept of data mining and using data clustering algorithms. Here we have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Analysis of police data provides insight that lets officers track criminal activities, predict the likelihood of incidents, effectively deploy resources and solve cases faster.*

***Key Words*: Crime-patterns, data clustering, data mining, k-means, fuzzy c, law-enforcement.**

1. INTRODUCTION

Crime rate is increasing now-a-days in many countries considerably. Crime cannot be predicted since it is not systematic. Despite of the modern technologies and hi-tech methods used to curb the crime, the criminals are successful in achieving their misdeeds. Although we cannot predict the victims of the crime but the place and probability of the occurrence of the crime can very well be predicted.

The process of solving crimes has been the prerogative of the criminal justice and law enforcement specialists. The increase in the use of computerized systems to track crimes, the law enforcement officers are helped by computer data analysts to speed up the process of solving crimes. We develop a data mining paradigm to help in solving crimes at a faster rate. More specifically, we will use clustering models to help in identification of crime patterns.[2]

For the purposes of our modeling, we will not need to get into the depths of criminal justice but will confine ourselves to the main kinds of crimes. Cluster (of crime) has a special meaning and refers to a geographical group of crime, i.e. a lot of crimes in a given geographical region. Such clusters can be visually represented using a geo-spatial plot of the crime overlaid on the map of the police jurisdiction. The densely populated group of crime is used to visually locate the 'hot-spots' of crime. However, when we talk of clustering from a data-mining standpoint, we refer to similar kinds of crime in the given geography of interest. Such clusters are useful in identifying a crime pattern.

The results obtained by prediction will certainly not be 100% accurate but the results will help in reducing crime rate to a certain extent by providing different types of security measures in crime sensitive areas. So in order to develop a crime analysis tool we have to collect different types of crime records and evaluate it [1].

The crime data available is in the structured form (criminal data) as well as unstructured (audio and image data from communications and surveillance). By bringing together the data available, the volume of the data can contribute to the understanding of what happened in the past and the based on the patterns what is likely to happen in future. With the help of such prediction, it can help the law enforcement agencies and take actions by understanding the following:

- Identify areas typically frequented by violent criminals.
- Match trends in regional or national gang activity with local incidents.
- Profile crimes to identify similarities and match the crimes to known offenders.
- Identify the conditions most likely to trigger violent crime, and predict when and where these crimes may occur in the future.

2. LITERATURE SURVEY

We have seen that in crime terminology a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in data mining terminology a cluster is group of similar data points – a possible crime pattern. Thus appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. Thus clustering algorithms in data mining are equivalent to the task of identifying groups of records that are similar between themselves but different from the rest of the data. Thus, in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

The different methods in data clustering for crime rate prediction are explained below.

2.1 K-Means

K-means clustering is one of the methods of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

1. Initially, the number of clusters must be known let it be k

2. The initial step is to choose a set of K instances as centre's of the clusters.

3. Next, the algorithm considers each instance and assigns it to the cluster which is closest.

4. The cluster centroids are recalculated either after whole cycle of re-assignment or each instance assignment.

5. This process is iterated.

K means algorithm complexity is O(tkn), where n is instances, c is clusters, and t is iterations and relatively efficient . It often terminates at a local optimum. Its disadvantage is applicable only when mean is defined and need to specify c, the number of clusters, in advance. It is unable to handle noisy data and outliers and is not suitable to discover clusters with non-convex shapes.

Step #1. If k=4, we select 4 random points and assume them to be cluster centers for the clusters to be created.

Step #2. We take up a random data point from the space and find out its distance from all the 4 clusters centers. If the data point is closest to the green cluster center, it is colored green and similarly all the points are categorized among the 4 clusters.

Step #3. Now we calculate the centroid of all the green points and assign that point as the cluster center for that cluster.

Similarly, we calculate centroids for all the 4 colored(clustered) points and assign the new centroids as the cluster centers.

Step #4. Step-2 and step-3 are run iteratively, unless the cluster centers converge at a point and no longer move.

2.2 Fuzzy-C

Fuzzy C is also called as soft clustering. In fuzzy clustering we make a fuzzy partition of the data set. But to accommodate the introduction of fuzzy partitioning, the membership matrix (U) is initialized randomly according to equation (i). Each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. In partition data set membership function is used. This function is called membership function and its value is between 0 and 1 .It is an iterative algorithm.

Mostly the aim of Fuzzy c means algorithm is to minimize a dissimilarity function

Using cluster centers (centroid) and minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point.

The overall procedure consists of three main steps:

1. The raw data clustering.

2. From data membership functions should be extracted.

3. Create the fuzzy inference system.

Algorithm:

1. Choose a number of clusters.

2. For being in the clusters randomly assign to each point coefficients.

3. Repeat until the algorithm has converged (that is, the coefficients change between two iterations is no more than threshold values, the given sensitivity threshold).

4. The centroid must be calculated for each cluster.

5. Also [5] Using dissimilarity functions to calculate the dissimilarities between centroid and data points.

6. For each point, compute its coefficients of being in the clusters.

7. The sum of square error which measures how well a clustering fits the data.

8. Calculate the threshold value ,if we find the correct threshold value then stop.

2.3 Hierarchical Clustering

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:[1]

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

A set of N items to be clustered, and an N*N distance matrix, the basic process of hierarchical clustering is this:

Begin by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances between the clusters the same as the distances between the items they contain.

Find the closest pair of clusters and merge them into a single cluster, so that now you have one cluster less.

Compute distances between the new cluster and each of the old clusters.

Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Algorithmic steps for Agglomerative Hierarchical clustering: [6]

Let X = {x1, x2, x3... xn} be the set of data points.

1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = min d[(i),(j)] where the minimum is over all pairs of clusters in the current clustering.

3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].

4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way: d[(k), (r,s)] = min (d[(k),(r)], d[(k),(s)]).

5) If all the data points are in one cluster then stop, else repeat from step 2).

Divisive Hierarchical clustering - It is just the reverse of Agglomerative Hierarchical approach.

2.4 Self Organizing Map

A self-organizing map consists of components called nodes or neurons. Associated with each node are a weight vector of the same dimension as the input data vectors, and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The self-organizing map describes a mapping from a higher-dimensional input space to a lower-dimensional map space. The procedure for placing a vector from data space on to the map is to find the node with the -closest (smallest distance metric) weight vector to the data space vector. While it is typical to consider this type of network structure as related to feed forward networks where the nodes are visualized as being attached, this type of architecture is fundamentally different in arrangement and motivation. It has been shown that while self-organizing maps with a small number of nodes behave in a way that is similar to K-means, larger self-organizing maps rearrange data in a way that is fundamentally topological in character. It is also common to use the U-Matrix. The U-Matrix value of a particular node is the average distance between the node's weight vector and that of its closest neighbors. In a square grid, for instance, we might consider the closest 4 or 8 nodes (the Von Neumann and More neighborhoods, respectively), or six nodes in a hexagonal grid.

Algorithm:

1-Randomize the map's nodes' weight vectors.

2-Grab an input vector.

3-Traverse each node in the map. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector. Track the node that produces the smallest distance (this node is the best matching unit, BMU).

4-Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector.

5-Increase s and repeat from step 2.

Table -1: Comparison of Methods

	Performance	Data type-ideal and random dataset	Data type-ideal and random dataset Quality
K-MEANS	Less accuracy	Gives better result when using ideal dataset.	Good quality when using huge dataset.
FUZZY C	Better than k-means.	Gives better result when using ideal dataset.	Good for large dataset.
HC	More accuracy with less number of clusters.	Gives better result when using random dataset.	Show good result when using small dataset.
SOM	As the number of cluster increases, accuracy increases.	Gives better result when using random dataset.	Show good result when using small dataset.

3. CONCLUSION

The increase in the rate of crime in today's world is a very important thing that needs to be curbed for a safer environment. In this paper, we have analyzed the use of data mining techniques for crime rate prediction. The crime rate

can be predicted by providing the appropriate input and also the changing crime patterns. Crime patterns cannot be static since patterns change over time, so different clustering techniques like K-Means; Fuzzy C etc. are used to handle the changing crime patterns. More in depth inputs will provide better and accurate results.

REFERENCES

[1] Shiju Sathyadevan, Devan M.S and Surya Gangadharan. S: Crime Analysis and Prediction Using Data Mining, 2014 First International Conference on Networks & Soft Computing.

[2] Shyam Varan Nath : Crime Pattern Detection Using Data Mining.

[3] Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal : Crime Analysis using K-Means Clustering, International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, December 2013.

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

[5] Adeyiga J.A: A Review of Different Clustering Techniques in Criminal Profiling, International Journal of Advanced Research in Computer Science and Software Engineering -Volume 6, Issue 4, April 2016.

[6] C.M. FUNG KE WANG AND MARTIN ESTER: - Hierarchical Clustering Algorithms.

[7] ANDREW MOORE: - K MEANS AND Hierarchical Clustering Algorithms.

[8] BERND FIRTZKE :- Self Organizing method with constant neighborhood range and adaptation strength Volume 2, Issue 5, September 1995.

[9] THOMAS VILLMANN,RALF DER:-Transactions of neural network, Volume 08, No 2, March 1997.