Analysis and Prediction of Crimes by Clustering and Classification

Rasoul Kiani
Department of Computer Engineering, Fars Science and Research Branch, Islamic Azad University, Marvdasht, Iran

Siamak Mahdavi
Department of Computer Engineering, Fars Science and Research Branch, Islamic Azad University, Marvdasht, Iran

Amin Keshavarzi
Department of Computer Engineering, Marvdasht Branch, Islamic Azad University, Marvdasht, Iran

Abstract—**Crimes will somehow influence organizations and institutions when occurred frequently in a society. Thus, it seems necessary to study reasons, factors and relations between occurrence of different crimes and finding the most appropriate ways to control and avoid more crimes. The main objective of this paper is to classify clustered crimes based on occurrence frequency during different years. Data mining is used extensively in terms of analysis, investigation and discovery of patterns for occurrence of different crimes. We applied a theoretical model based on data mining techniques such as clustering and classification to real crime dataset recorded by police in England and Wales within 1990 to 2011. We assigned weights to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm (GA) is used for optimizing of Outlier Detection operator parameters using RapidMiner tool.**

Keywords—crime; clustering; classification; genetic algorithm; weighting; rapidminer

I. INTRODUCTION

A. Crime Analysis

Today, collection and analysis of crime-related data are imperative to security agencies. The use of a coherent method to classify these data based on the rate and location of occurrence, detection of the hidden pattern among the committed crimes at different times, and prediction of their future relationship are the most important aspects that have to be addressed.

In this regard, the use of real datasets and presentation of a suitable framework that does not be affected by outliers should be considered. Preprocessing is an important phase in data mining in which the results are significantly affected by outliers. Thus, the outlier data should be detected and eliminated though a suitable method. Optimization of Outlier Detection operator parameters through the GA and definition of a Fitness function are both based on Accuracy and Classification error. The weighting method was used to eliminate low-value features because such data reduce the quality of data clustering and classification and, consequently, reduce the prediction accuracy and increase the classification error.

The main purposes of crime analysis are mentioned below [1]:

- Extraction of crime patterns by crime analysis and based on available criminal information,
- Prediction of crimes based on spatial distribution of existing data and prediction of crime frequency using various data mining techniques,
- Crime recognition.

B. Clustering

Division of a set of data or objects to a number of clusters is called clustering. Thereby, a cluster is composed of a set of similar data which behave same as a group. It can be said that the clustering is equal to the classification, with only difference that the classes are not defined and determined in advance, and grouping of the data is done without supervision [2].

C. Clustering by K-means Algorithm

K-means is the simplest and most commonly used partitioning algorithm among the clustering algorithms in scientific and industrial software [3] [4] [5]. Acceptance of the K-means is mainly due to its being simple. This algorithm is also suitable for clustering of the large datasets since it has much less computational complexity, though this complexity grows linearly by increasing of the data points [5]. Beside simplicity of this technique, it however suffers from some disadvantages such as determination of the number of clusters by user, affectability from outlier data, high-dimensional data, and sensitivity toward centers for initial clusters and thus possibility of being trapped into local minimum may reduce efficiency of the K-means algorithm [6].

D. Classification

Classification is one of the important features of data mining as a technique for modeling of forecasts. In other words, classification is the process of dividing the data to some groups that can act either dependently or independently [7]. Classification is used to make some examples of hidden and future decisions on the basis of the previous decision makings [8]. Decision tree learning, neural network, nearest neighborhood, Nave Bayes method and support vector machine are different algorithms which are used for the purpose of classification [9].

E. Genetic Algorithm

In [10] S. Sindhiya and S. Gunasundari, have discussed about Genetic Algorithm (GA) as an evolutionary algorithm. “GA starts with an initial population called a candidate solution. After a sequence of iterations it achieves the optimal solution. The fitness is used to estimate the quality of each candidate solution. The chromosome, which has the highest fitness is to be kept in the next iteration. The crossover and mutation are the two basic operators of GA. The crossover is the procedure of taking above one parent solutions and generating a child solution. The mutation is used to preserve the genetic diversity from one iteration to the next iteration. And again the fitness function and the genetic operators are used to generate successive generations of individuals and are repeated several times until a suitable solution is found. The performance of GA depends on a number of issues such as crossover, mutation, fitness function and the various user determined parameters such as population size, probability of genetic operators.”

The rest of the paper is organized as follows. Section 2 describes the existing systems for analyzing crimes. The New framework and experimental results are presented in section 3. Section 4 contains the conclusion. Finally, section 5 discusses the future scope of this paper.

II. Letrature Review

J. Agarwal, R. Nagpal and R. Sehgal in [1] have analyzed crime and considered homicide crime taking into account the corresponding year and that the trend is descending from 1990 to 2011. They have used the k-means clustering technique for extracting useful information from the crime dataset using RapidMiner tool because it is solid and complete package with flexible support options. Figure1 shows the proposed system architecture.

Priyanka Gera and Dr. Rajan Vohra in [11] have used a linear regression for prediction the occurrence of crimes in Delhi (India). They review a dataset of the last 59 years to predict occurrence of some crimes including murder, burglary, robbery and etc. Their work will be helpful for the local police stations in decision making and crime supervision.

“After training systems will predict data values for next coming fifteen years. The system is trained by applying linear regression over previous year data. This will produce a formula and squared correlation(r^2).

The formula is used to predict values for comong future years. The coefficent of determination, r^2, is useful because is gives the proportion of variance of one variable that is predictable from other variable.” Figure 2 shows the proposed system architecture.

In [12] an integrated system called PrepSearch have proposed by L. Ding et al. It has been combined using two separate categories of visualization tools: providing the geographic view of crimes and visualization ability for social networks. “It will take a given description of a crime, including its location, type, and the physical description of suspects (personal characteristics) as input.

To detect suspects, the system will process these inputs through four integrated components: geographic profiling, social network analysis, crime patterns and physical matching.” Figure 3 shows the system design and process of PrepSearch.

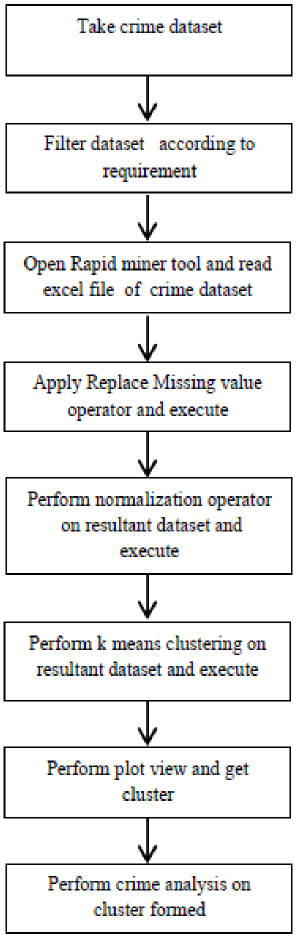
Take crime dataset
Filter dataset according to requirement
Open Rapid miner tool and read excel file of crime dataset
Apply Replace Missing value operator and execute
Perform normalization operator on resultant dataset and execute
Perform k means clustering on resultant dataset and execute
Perform plot view and get cluster
Perform crime analysis on cluster formed

Fig. 1. Flow chart of crime analysis [1]

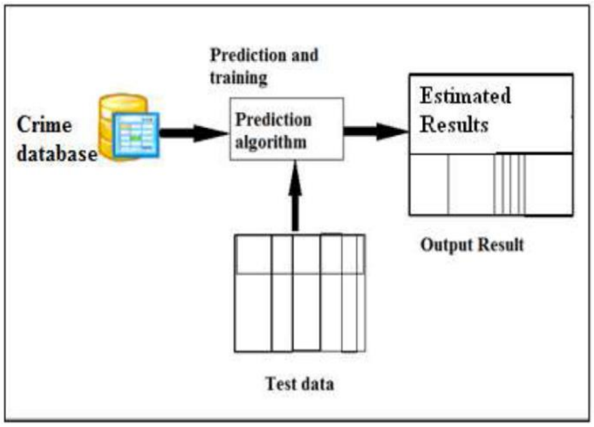

Fig. 2. Predicting future crime trends [11]

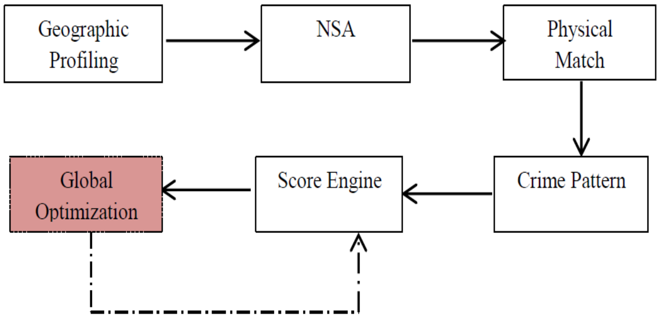

Fig. 3. System design and process of PrepSearch [12]

In [13] researches have introduced intelligent criminal identification system called ICIS which can potentially distinguish a criminal in accordance with the observations collected from the crime location for a certain class of crimes.

The system uses existing evidences in situations for identifying a criminal by clustering mechanism to segment crime data in to subsets, and the Nave Bayesian classification has used for identifying possible suspect of crime incidents. ICIS has been used the communication power of multi agent system for increasing the efficiency in identifying possible suspects. In order to describe the system ICIS is divided to user interface, managed bean, multi agent system and database. Oracle Database is used for implementing of database, and identification of crime patterns has been implemented using Java platform.

In [14] an improved method of classification algorithms for crime prediction has proposed by A. Babakura, N. Sulaiman and M. Yusuf. They have compared Naïve Bayesian and Back Propagation (BP) classification algorithms for predicting crime category for distinctive state in USA. In the first step phase, the model is built on the training and in the second phase the model is applied. The performance measurements such as Accuracy, Precision and Recall are used for comparing of the classification algorithms. The precision and recall remain the same when BP is used as a classifier.

In [15] researches have introduced crime analysis and prediction using data mining. They have proposed an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Also they have focused on causes of crime occurrence like criminal background of offender, political, enmity and crime factors of each day. Their method steps are data collection, classification, pattern identification, prediction and visualization.

III. New Framework

In this section a new framework is introduced for clustering and prediction of cluster members to analyze crimes. A dataset of crimes recorded by police in England and Wales[1] within 1990 to 2011 has been used, and RapidMiner will be used for the purpose of implementation.

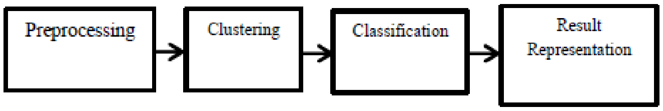

Fig. 4. New framework

A. Preprocessing Phase

1) Read the dataset of crime using Read Excel operator: The dataset selectd by the operator is read in RapidMiner tool.

2) Filter dataset according to requirement: Since there may be data in the read dataset that would not be used according to our method, the unnecessary data have to be filtered.

3) Apply Replace Missing Value operator: This operator replaces the dataset missing values with a new value, and adds it to our previous dataset. This can be done by one of the “Minimum”, “Maximum”, “Average” and “None” functions which is determined by the Default parameter. If “None” were selected, it will be leaded to no replacement. To achieve this purpose there is an accessible wizard using Column parameter.

Each one of these features chooses one of the functions through Column parameter. If one feature name were be shown as a key in the list, function name will be used as the key value. If the feature name does not exist in the list, a selected function with default parameters can be used for it. For nominal features the Mod function, nominal value that has been occurred the most in the dataset, can be replaced with the Average function as an example. For the nominal features which their replacement parameter has been assigned Zero for them, the first nominal defined value for the feature will be replaced with the missing values. We can also use the Replenishment Value parameter to specify the replacement values.

4) Outlier detection using Outlier Detection(Distance) operator: Outlier detection goals:

- Improving the quality of clusters in clustering phase,
- Increasing the accuracy of Decision tree in classification phase,
- Decreasing the classification error in classification phase.

This operator tries to detect outliers in the dataset according to their distance with their neighbors. This operator discovers outliers by a kind of search that can be known as a statistical search. This method starts the search according to the distance of K-th nearest neighbors, and then it tries to sort the search result by their local position. The ones which are farther than their K-th nearest neighbor will be specified as the outlier in the dataset more probably. Theory says dispersion and distance of outliers is more than the average of dataset. Then according to the distance of each data with its K-th nearest neighbor, all data will be ranked and then we can say the higher ranking shows the outlier of the dataset.

[1] www.gov.uk/government/publications/offences

Parameters:

- Number of neighbors: Specifies the value for the k-th nearest neighbors,
- Number of outliers: The number of top-n outliers need to be looked for.

5) Outlier Distance operator parameters optimization using GA: At this point, as shown in Figure 5, given that the process of optimizing the Outlier Detection operator parameters must improve the results of the predicted cluster members in the clusters, parameters of Accuracy and Classification error are used to define the fitness function

Optimize Parameters (Evolutionary) operator values:

- Max generation= 50
- Population size= 5
- Crossover prob= 0.9

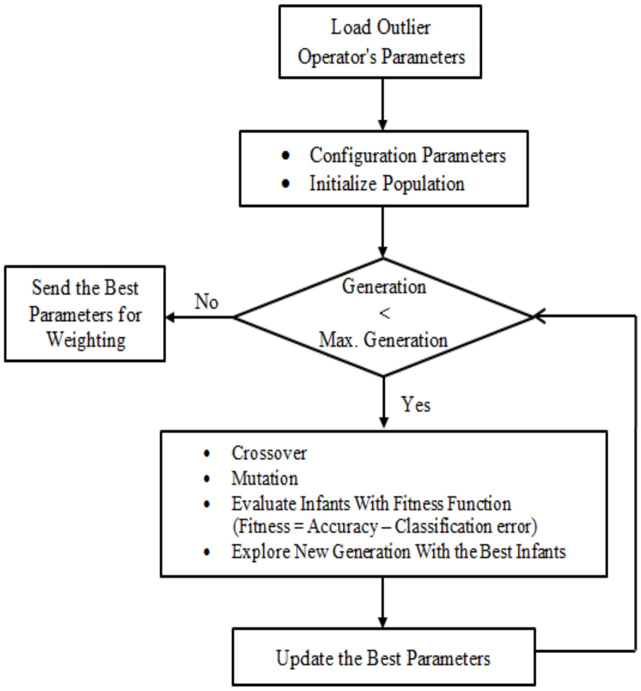
Load Outlier Operator's Parameters
Configuration Parameters
Initialize Population
Generation < Max. Generation
No
Send the Best Parameters for Weighting
Yes
Crossover
Mutation
Evaluate Infants With Fitness Function (Fitness = Accuracy – Classification error)
Explore New Generation With the Best Infants
Update the Best Parameters

Fig. 5. Genetic algorithm for optimizing

6) Store a new dataset: After applying the changes, the dataset is stored for further use.

B. Clustering Phase

1) Apply Weight by Deviation operator: One of the clustering algorithm challenges is high-dimensional data, so to deal with this challenge is using weighted features and removing low-value featuers. A suitable operator to apply our idea is Weight by Deviation. It creates weights from the standard deviations of all featuers. The values can be normalized by the “Average”, “Minimum”, or “Maximum” of the featuers.

Parameters:

- Normalize weights: Activates the normalization of all weights,
- Normalize: Indicates that the standard deviation should be divided by the minimum, maximum, or average of the featuers.

2) Thershold selection: There is not specific critria for selecting the threshold. It is selected based on the Trial and error method for removing low-value featuers. The threshold is determined, and all the featuers, that their values are equal or less than it, will be removed from the dataset.

3) Store a new dataset: After applying the changes, the dataset is stored for further use.

4) Performe K-means algorithm on result dataset: At this step, the clustering process is carried out on the dataset using K-means operator.

5) Enable K-means Operator Parameters: The classification process is performed on the data after data clustering; therefore, the target class is defined in this step. The goal is that the cluster obtained in the previous step be defined as the target class. The k-means operator parameters are used for this purpose.

Parameters:

- Add cluster attribute: If enabled, a cluster id is generated as new special attribute directly in this operator. Othrwise this operator does not add an id attribute,
- Add as label: If true, the cluster id is stored in an attribute with the special role “label” instead of “cluster”.

C. Classification Phase

1) Training and Testing Data: In this phase production of training and testing data is done using Sample (Stratified) and Set Minus operators for increasing confidence in the response without replacement.

- Sample operator is used to reserve 10% of data,
- Set Minus operator is used to reduce training data from the dataset.

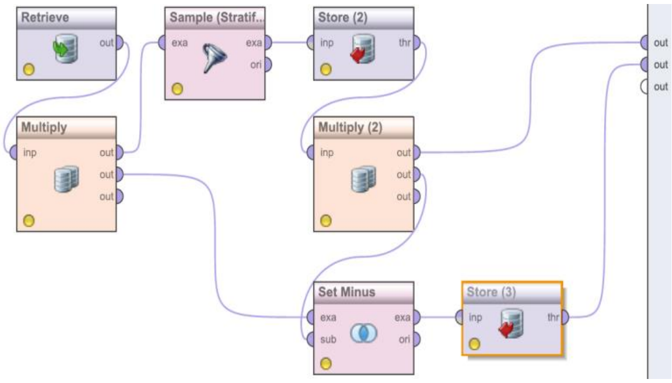

Fig. 6. Production of training and testing data

2) Decision Tree: We use Decision Tree operator to learn decision tree model, and the value of Criterion parameter is selected "gini_index".

3) Apply the model and test data: The model and test data have been produced in the previous steps are applied on Apply Model operator inputs to predict the cluster members.

D. Result Presentation Phase

At this phase, the following operators are used to show the results obtained from the presented framework.

- The model accuracy and classification error are calculated by Performance operators,
- Log operator is used to record and save performance report,
- A comparison of accuracy and classification error are used to evaluate the effect of optimization Outlier Detection operator parameters,
- Analysis of crimes based on the new framework.

Figure 7 shows the new framework scheme with details. Comparison between results in Table 1 shows that when the number of clusters is similar, after optimizing the parameters of the Outlier Detection operator, the classification accuracy increased and classification error decreased and, consequently, the obtained fitness function was optimized. Figure 8 shows the predicted occurrence rate for the crimes of buggery, homicide, and robbery. Also Figure 9 shows the implementation of the model by Rapid Miner tool.

TABLE I. RESULTS

Mode	Number of Cluster	Accuracy of Prediction	Classification Error	Fitness Function
Optimized	6	91.64%	8.36%	83.28
Non-Optimized	6	85.74%	13.26%	72.48

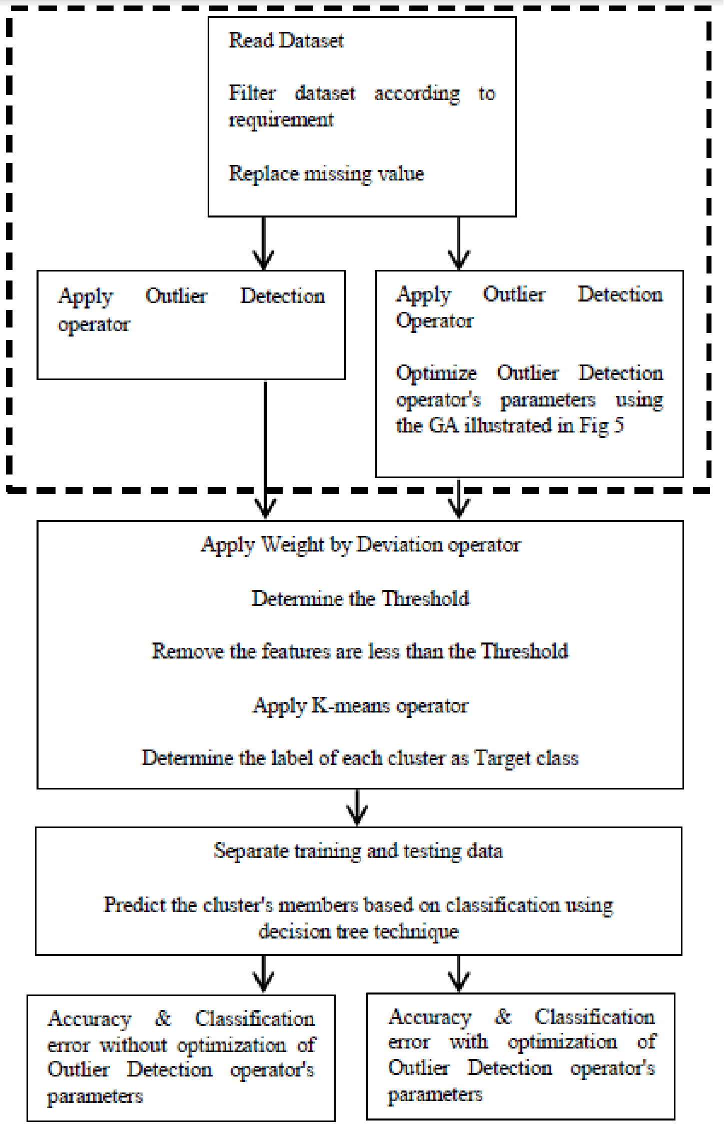
Read Dataset
Filter dataset according to requirement
Replace missing value
Apply Outlier Detection operator
Apply Outlier Detection Operator
Optimize Outlier Detection operator's parameters using the GA illustrated in Fig 5
Apply Weight by Deviation operator
Determine the Threshold
Remove the features are less than the Threshold
Apply K-means operator
Determine the label of each cluster as Target class
Separate training and testing data
Predict the cluster's members based on classification using decision tree technique
Accuracy & Classification error without optimization of Outlier Detection operator's parameters
Accuracy & Classification error with optimization of Outlier Detection operator's parameters

Fig. 7. The new framework scheme with details

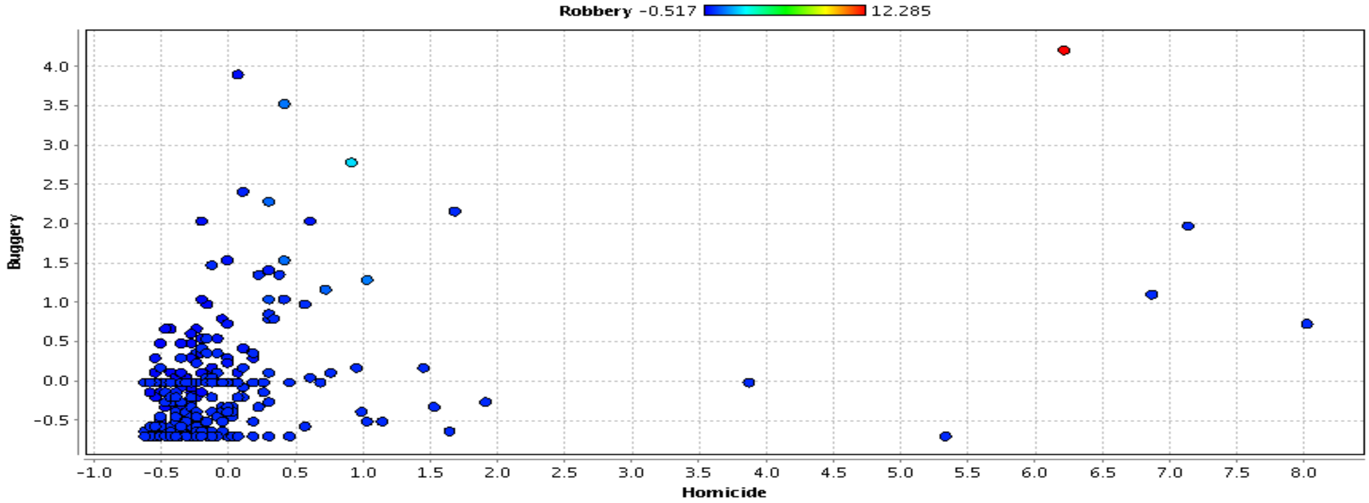

Fig. 8. Prediction of robbery, buggery and homicide

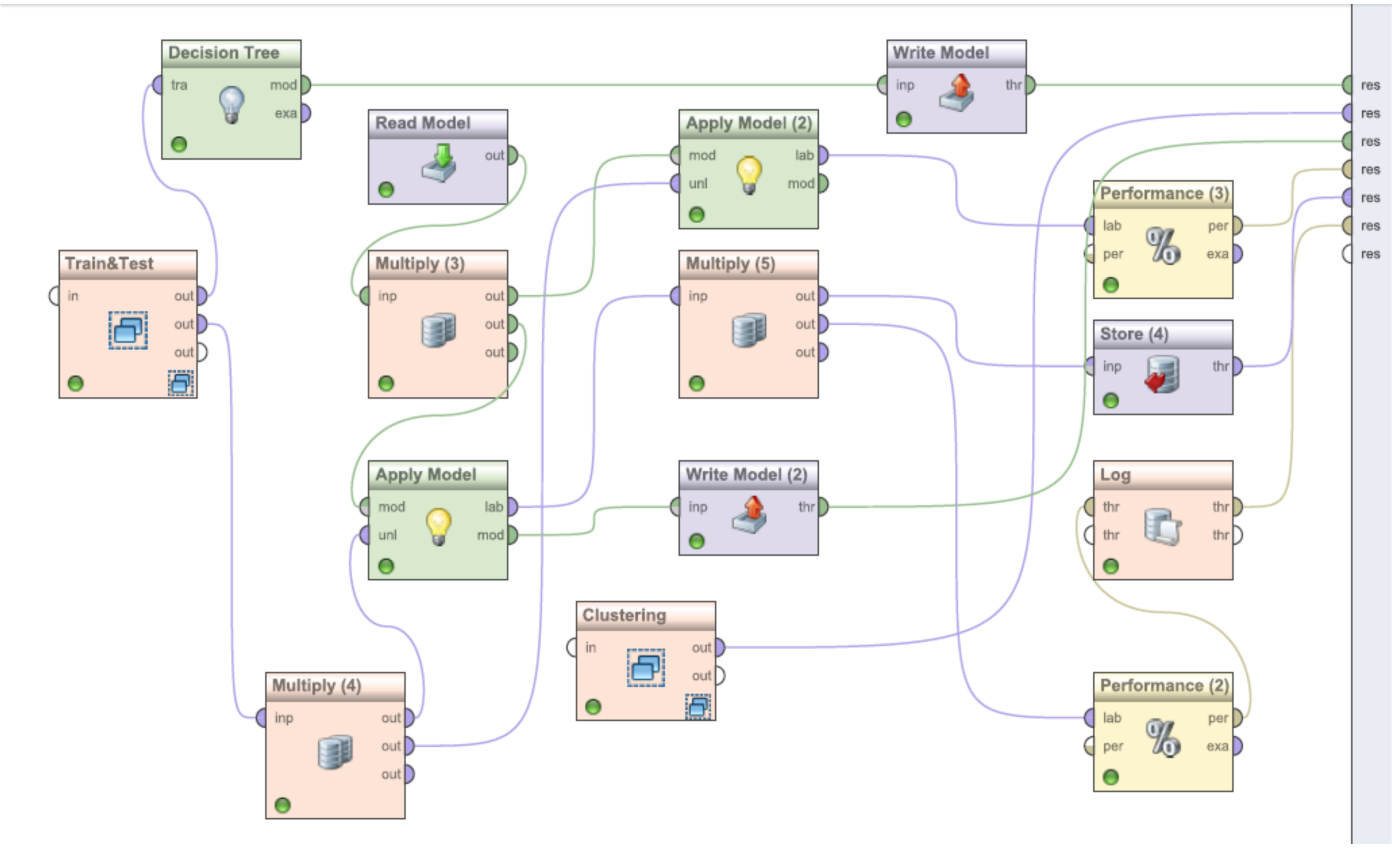

Fig. 9. Model for prediction of criems using decision tree

IV. Conclusion

This paper presents a new framework for clustering and predicting crimes based on real data. Examining the methods proposed for crime prediction shows that the parameters such as the effect of outliers in the data mining preprocessing, quality of the training and testing data, and the value of features have not been addressed before. In this framework, the GA was used to improve outlier detection in the preprocessing phase, and the fitness function was defined based on accuracy and classification error parameters. In order to improve the clustering process, the features were weighted, and the low-value features were deleted through selecting a suitable threshold. The proposed method was implemented, and the results of the optimized and non-optimized parameters were compared to determine their quality and effectiveness.

The main purposes of the new framework for clustering and classification of crimes are mentioned below:

- Generation of training and testing data,
- Removing low-value attributes using weighting technique to deal with high-dimensional data challenge,
- Optimization of Outlier operator parameters using GA.

V. Future Scope

One of the most important issues that should be addressed in the model presented in this paper to improve the clustering process and crime detection is the optimization of the number of clusters in the clustering process and the optimization of the technique used in the prediction phase of model development.

Acknowledgment

Hereby, we would like to kindly acknowledge sincere cooperation and valuable instructions of Mr. Mojtaba Montazeri and Mr. Mohammad Sarabi.

References

[1] J. Agarwal, R. Nagpal, and R. Sehgal, "Crime analysis using k-means clustering," International Journal of Computer Applications, Vol. 83 – No4, December 2013.

[2] J. Han, and M. Kamber, "Data mining: concepts and techniques," Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000.

[3] P. Berkhin, "Survey of clustering data mining techniques," In: Accrue Software, 2003.

[4] W. Li, "Modified k-means clustering algorithm," IEEE Congress on Image and Signal Processing, pp. 616- 621, 2006.

[5] D.T Pham, S. Otri, A. Afifty, M. Mahmuddin, and H. Al-Jabbouli, "Data clustering using the Bees algorithm," proceedings of 40th CRIP International Manufacturing Systems Seminar, 2006.

[6] J. Han, and M. Kamber, "Data mining: concepts and techniques," 2nd Edition, Morgan Kaufmann Publisher, 2001.

[7] S. Joshi, and B. Nigam, "Categorizing the document using multi class classification in data mining," International Conference on Computational Intelligence and Communication Systems, 2011.

[8] T. Phyu, "Survey of classification techniques in data mining," Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol. IIMECS 2009, March 18 - 20, 2009, Hong Kong.

[9] S.B. Kim, H.C. Rim, D.S. Yook, and H.S. Lim, "Effective Methods for Improving Naïve Bayes Text Classifiers," In Proceeding of the 7th Pacific Rim International Conference on Artificial Intelligence, Vol.2417, 2002.

[10] S. Sindhiya, and S. Gunasundari, "A survey on Genetic algorithm based feature selection for disease diagnosis system," IEEE International Conference on Computer Communication and Systems(ICCCS), Feb 20-21, 2014, Chermai, INDIA.

[11] P. Gera, and R. Vohra, "Predicting Future Trends in City Crime Using Linear Regression," IJCSMS (International Journal of Computer Science & Management Studies) Vol. 14, Issue 07Publishing Month: July 2014.

[12] L. Ding et al., "PerpSearch: an integrated crime detection system," 2009 IEEE 161-163 ISI 2009, June 8-11, 2009, Richardson, TX, USA.

[13] K. Bogahawatte, and S. Adikari, "Intelligent criminal identification system," IEEE 2013 The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka.

[14] A. Babakura, N. Sulaiman, and M. Yusuf, "Improved method of calssification algorithms for crime prediction," International Symposium on Biometrics and Security Technologies (ISBAST) IEEE 2014.

[15] S. Sathyadevan, and S. Gangadharan, "Crime analysis and prediction using data mining," IEEE 2014.