Using Linear Regression to Forecast Future Trends in Crime of Bangladesh

Md. Abdul Awal, Jakaria Rabbi, Sk. Imran Hossain, and M. M. A. Hashem
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
E-mail: awal.kuet@yahoo.com, jakaria.rabbi@yahoo.com, imranhrana@gmail.com, and mma.hashem@outlook.com

***Abstract*— Crime is basically unpredictable and a social disturbance. With the increase in population of Bangladesh, the rate of crime is also increasing and affecting our society fatally in various ways. So it has become significant to analysis crime data for better understanding of future crime trends. In this case, machine learning and data mining techniques can play a significant role to discover future trends and patterns of crime. In this paper, linear regression model is used to forecast future crime trends of Bangladesh. The real dataset of crime is collected from the website of Bangladesh police. Then the linear regression model is trained on this dataset. After training the model, crime forecasting is done for dacoit, robbery, murder, women & child repression, kidnapping, burglary, theft and others for different region of Bangladesh. This work may be helpful for Bangladesh police and law enforcement agencies to forecast, prevent or solve future crime of Bangladesh.**

Keywords— Data Mining, Crime Forecasting, Linear Regression, Gradient Descent.

I. INTRODUCTION

Historically crime is a well known social problem. Crime is an action that is deemed injurious to the public welfare and is legally prohibited. It is also an offense against the society which is punishable by the law. It disrupts not only the normal way of life but also the socio economic development of a society. Therefore, it is indispensable to analyze crime data to inform the police department and law enforcement agencies about specific and general trends and patterns of crime regularly.

Due to the rapid development in computerization and digitalization techniques, a large amount of data are now available in science, business, medicine, education, marketing, banking, airlines and many other areas. Fast computing systems give a scope to study this huge amount of data in various ways which were not convenient in earlier time. Such data may provide a great resource for knowledge discovery and decision support system. In order to understand, analyze and gradually make sense of this huge amount of data, a multidisciplinary approach, data mining is introduced to accept the challenge. Data mining techniques can play an important role to analyze this data and discover knowledge from them [1]. One of this data are crime data that highly affect the society. Therefore, to focus on the scientific study of crime and criminal behavior this crime data should be analyzed. The main objective of crime data analysis is to identify the crime characteristics and their relationships with the criminals [2,23]. The high volume of crime dataset and also the complexity of relationships between this kinds of data make the analysis procedure difficult [3,24[. At present manual investigation of crime data by analysts and law enforcement agencies is limited. In order to facilitate the analysis procedure, data mining techniques have become the important tools. Data mining techniques accelerate crime analytics, provide better analysis and generate effective solutions to save important resources and time [4]. The tasks of data mining are classification, clustering, evaluation, association, prediction and trend analysis [5]. By the application of data mining techniques, the knowledge that is discovered from crime data analysis may assist the police department and law enforcement agencies to predict, prevent or solve future crime of Bangladesh.

At present, different types of crime are responsible for causing a lot of problems all over the world. For that reason, crime analysts are expending time studying in crime and criminal behaviors in order to explore the characteristics and patterns of crime. It is acquainted that criminals follow repetitive patterns, so analyzing their behaviors can assist to find out relations among events from past crimes [6].

There are several characteristics that have great influence on crime data analysis. They includes: different kinsmen in a society, poverty, income groups, age groups, family structure (single, divorced, married), level of education, the locality where people live, number of police officers allocated to a locality, number of employed and unemployed people among others [6,7,22].

The Uniform Crime Reporting (UCR) program of Federal Bureau of Investigation (FBI) categorizes crime into four categories: murder, forcible rape, robbery, and aggravated assault [8]. In this paper, different types of crime such as dacoit, robbery, murder, women & child repression, kidnapping, burglary, theft and others crime are considered. For the purpose of this research, the dataset is collected from the website of Bangladesh police [1,2]. The dataset contains aggregated counts of different types of crime categorized by the police department and are considered to be forecasted. In this paper, the linear regression model is trained on the dataset of Bangladesh police. After training the model, basically crime forecasting is done for different types of crime for metropolitan and divisional region of Bangladesh.

The rest of the paper is organized as follows: section II describes the existing works, section III describes the research methodology and section IV presents the results of the experimental analysis. Finally, section V concludes the paper.

II. EXISTING WORKS

Crime data mining is taking onward attention to explore hidden patterns in crime data. Based on existing research, it has been observed that data mining techniques assist the

978-1-5090-1269-5/16/$31.00 ©2016 IEEE

procedure of crime patterns detection. To analyze the crime data, classification and machine learning algorithms are used. Yu et al. [9], employ an ensemble of data mining classification techniques to perform the crime forecasting. A variety of classification methods such as: One Nearest Neighbor (1NN), Decision Tree (J48), Support Vector Machine (SVM), Neural Network (Neural) with 2 layer network, and Naïve Bayesian (Bayes) were used to predict the crime "hotspot". Finally the best forecasting approach was proposed to achieve the most stable outcomes.

In [4], fuzzy association rule mining approach was used for community crime pattern discovery. The discovered rules were presented and discussed at regional and national levels for crime pattern investigation. Levine [2], developed a spatial statistical program for the analysis of crime incidents or other point locations. It was designed to operate with large crime – incident dataset collected by metropolitan police departments. In [11], different hotspot mapping techniques such as: point mapping, thematic mapping of geographic areas (e.g. Census areas), spatial ellipses, grid thematic mapping and kernel density estimation (KDE) were used for identifying hotspot of crimes. Finally, hotspot mapping accuracy was compared in relation to the mapping technique that was used to identify concentrations of crime events and by crime type – four crime types were compared (burglary, street crime, theft from vehicles and theft of vehicles).

In [1], a comparative study was conducted between some free available data mining and knowledge discovery tools and software packages. The result showed that, the performance of the tools for the classification task is affected by the kind of dataset used and by the way the classification algorithms were implemented within the toolkits. In [3], a clustering based model was used to anticipate crime trends. Performance of clustering technique was analyzed in forming accurate clusters, speed of creating clusters, efficiency in identifying crime trend, identifying crime zones, crime density of a state and efficiency of a state in controlling crime rate. In [6], an experiment was conducted to obtain better supervised classification learning algorithms (Naïve Bayesian (0.898), k Nearest Neighbor (k-NN) (0.895) and Neural Networks (0.892), Decision Tree (J48) (0.727), and Support Vector Machine (SVM) (0.678)) to predict crime status. Two different feature selection methods were tested on real dataset for prediction. Chi-square feature selection technique was used to improve the performance of mining results.

Malathi et al. [7], applied anomalies detection and clustering algorithms to predict the crime patterns and speed up the process of solving crime. MV algorithm, DBScan and PAM outlier detection algorithm were used to assist in the process of filling the missing value and investigation of crime patterns. In [8], a comparative study was conducted between the violent crime patterns from the communities and crime unnormalized dataset and actual crime statistical data for the state of Mississippi. Linear regression, additive regression, and decision stump algorithms were implemented on the communities and crime dataset. The linear regression algorithm performed the best among the three selected algorithms. By considering the geographical approach, Nath et al. [10], applied a combination of k-means and weighting algorithm which show regional crimes on a map and cluster crimes according to their types.

Bruin et al. [12], created digital profiles for all offenders by extracting the factors: crime nature, frequency, duration and severity from the crime database. Comparison of all individuals were performed on these profiles by a new distance measure and cluster them accordingly. Thongtae et al. [13], delivered a comprehensive survey of efficient and effective methods on data mining for crime data analysis. They explored the illegal activities of professional identity fraudsters based on knowledge discovered from their own histories. Bagui et al. [14], used WEKA for mining association rules, developing a decision tree, and clustering to retrieve meaningful information about crime from a U.S. state database.

Abraham et al. [15], employed the association rule data mining technique to generate profiles from log data to realize the criminal's behavior. In [16], a density tracing based approach was used to discover patterns among datasets by finding those crime and spatial features that exhibit similar spatial distributions by measuring the dissimilarity of their density traces.

III. METHODOLOGY

Data mining is a part of the interdisciplinary field of knowledge discovery in database [17]. Data mining consists of collecting raw data and, extracting information that can be applied to make predictions in many real world situations such as the stock market or tracking spending habits at the local Wal-Mart [8]. In this paper, data mining is used to forecast future trends in crime of Bangladesh. The methodology is divided into two parts, these are: 1) dataset description and 2) model selection

A. Dataset Description:

To conduct this research, the real dataset of crime of previous years is collected from the website of Bangladesh police. The dataset contains aggregated counts of different types of crime categorized by the police department of Bangladesh. The dataset is divided into two groups according to the region of Bangladesh: metropolitan region data and divisional region data. The dataset consists of *840* instances or crimes that had been collected from across the country. The dataset contains three predictive features and one goal feature. The three predictive features are: region, month and year. The goal feature is the predicted value of different types of crime. The data in each instance belong to different regions of Bangladesh. The regions are represented in the form of a number, every number represents its perspective region of Bangladesh.

B. Model Selection:

In order to quantitatively forecast the status of crime, different data mining techniques can be used. The associated task for the dataset used in this paper is regression. Therefore, liner regression model is used to forecast the status of crime. The linear regression model is simple and provides enough description of how the input affects the output. It predicts a variable Y (target variable) as a liner function of another variable X (input variable/features), given m training examples of the form $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$, where $x_i \varepsilon X$ and $y_i \varepsilon Y$.

The form of hypothesis of linear regression can be expressed as

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$
$$h_\theta(x) = \theta^T x$$

Where $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$ are regression parameters. To find the regression parameters the following cost function is used:

$$J(\theta_0, \theta_1, \theta_2, \ldots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

The objective of linear regression is to minimize the cost function $J(\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$ by finding $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$, so that $h_\theta(x)$ is close to y for every training examples (x,y). One way to minimize the cost function is to use the batch gradient descent algorithm. The algorithm is below:

Repeat until convergence {

$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$ *(for every j)*

}

Where α is the learning rate. With each step of gradient descent, the parameters $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$ come closer to the optimal values that will achieve the lowest cost $J(\theta_0, \theta_1, \theta_2, \ldots, \theta_n)$. To speed up the minimization process using gradient descent, every feature is normalized in the range $[-1,1]$. To normalize the feature the following equation [18] is used:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation.

IV. Results and Discussions

All of the results from the implementation of linear regression is provided in this section. The algorithm was run to forecast each of the crime: dacoit, robbery, murder, women & child repression, kidnapping, burglary, theft and others. Here dacoit, robbery, burglary and theft is considered into a single group to forecast. The result is divided into two parts. One is metropolitan region result and other is divisional region result. Table 1 shows that, how accurate the liner regression is to forecast future crime trends of Bangladesh. For training purpose, learning parameter α is set to the value 0.01 and gradient descent is iterated 400 times. Fig 1. Shows the convergence of gradient descent with an appropriate learning rate. After training linear regression, the regression parameters are: $\theta_0 = 1870.38$, $\theta_1 = 2422.64$ and $\theta_2 = 143.27$. These parameters are used to forecast murder in 2014 as follows:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$
$$h_\theta(x) = 1870.38 + 2422.64 * region + (-143.27) * year$$
$$h_\theta(x) = 1870.38 + 2422.64 * 10 + (-143.27) * 2014$$
$$h_\theta(x) = 255$$

The above calculation is done for Dhaka metropolitan police (DMP) region. Before calculation all the features are normalized. Different numeric values are used for different regions which was discussed in the subsection: dataset description. Here numeric value *10* is used for DMP region.

Crime forecasting at metropolitan region: In order to forecast different crimes in different months of 2016 at metropolitan region the following regression parameters are found after training linear regression model: $\theta_0 = -212.69$, $\theta_1 = -3.48$, $\theta_2 = 0.24$ and $\theta_3 = 0.11$

$$h_\theta(x) = -212.69 + (-3.48) * region + 0.24 * month + 0.11 * year$$
$$h_\theta(x) = -212.69 + (-3.48) * 10 + 0.24 * 1 + 0.11 * 2016 = 15$$

TABLE 1: TOTAL NO OF MURDERS IN 2014 AT METROPOLITAN AND DIVISIONAL REGION OF BANGLADESH.

Region	Actual no. of Murder	Predicted no. of Murder
DMP	262	255
CMP	120	107
KMP	22	39
RMP	22	37
BMP	15	28
SMP	44	46
Dhaka Range	1395	1243
Chittagong Range	792	757
Sylhet Range	277	366
Khulna Range	520	486
Barisal Range	209	256
Rajshahi Range	463	406
Rangpur Range	349	416

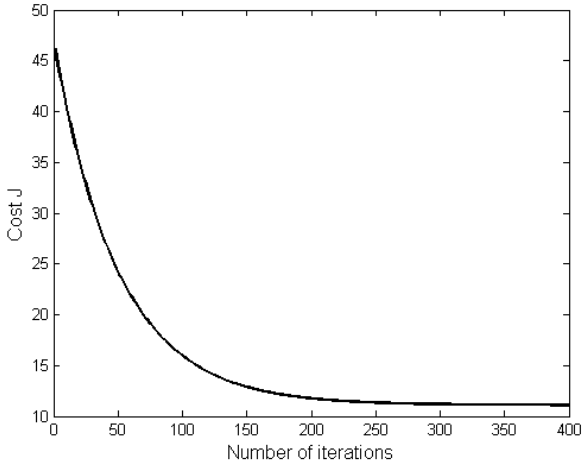

Figure 1. Convergence of gradient descent with an appropriate learning rate.

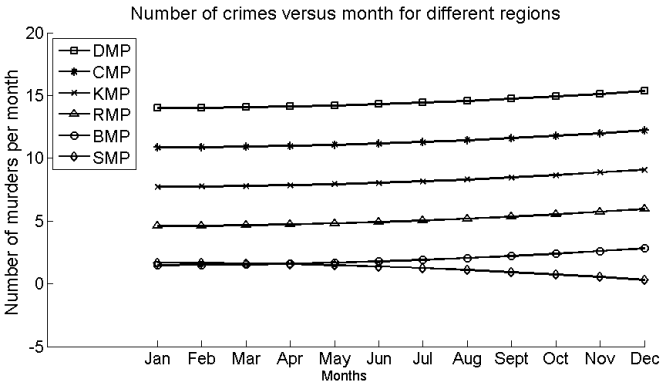

Figure 2. Murder forecasting of 2016 at metropolitan region.

The above calculation is done for crime forecasting of murder in January of 2016 at DMP region. Similar calculation is done for other types of crime in different months of 2016 at metropolitan region. All the results of crime forecasting for metropolitan regions are depicted through the fig. 2 to 6.

These figures show the change of different types of crime such as: dacoit, robbery, murder, women & child repression, kidnapping, burglary, theft and others with respect to month.

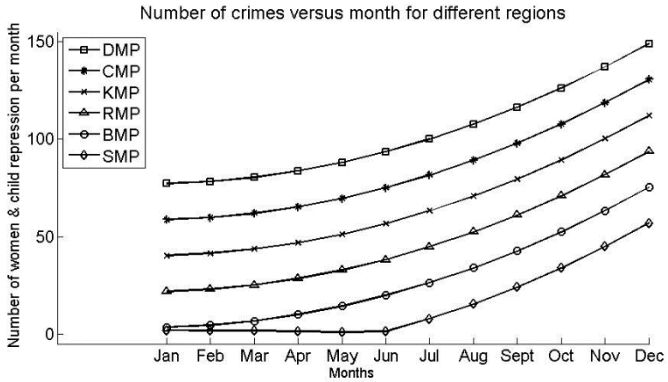

Figure 3. Women & child repression forecasting of 2016 at metropolitan region.

Crime forecasting at divisional region: In order to forecast different crimes in different months of 2016 at divisional region the following regression parameters are found after training linear regression model:

$$\theta_0 = 2396.3,\ \theta_1 = -14.46 \text{ and } \theta_2 = -0.67 \text{ and } \theta_3 = -1.18.$$

$$h_\theta(x) = 2396.3 + (-14.46) * \text{region} + (-0.67) * month + (-1.18) * year$$

$$h_\theta(x) = 2396.3 + (-14.46) * 10 + (-0.67) * 1 + (-1.18) * 2016 = 73$$

The above calculation is done for crime forecasting of murder in January of 2016 at Dhaka range. Similar calculation is done for other types of crime in different months of 2016 at divisional region. All the results of crime forecasting for divisional regions are depicted through the fig. 7 to 11. These figures show the change of different types of crime such as: dacoit, robbery, murder, women & child repression, kidnapping, burglary, theft and others with respect to month.

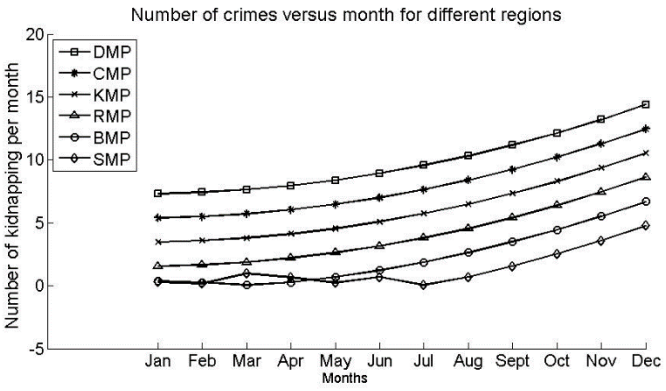

Figure 4. Kidnapping forecasting of 2016 at metropolitan region.

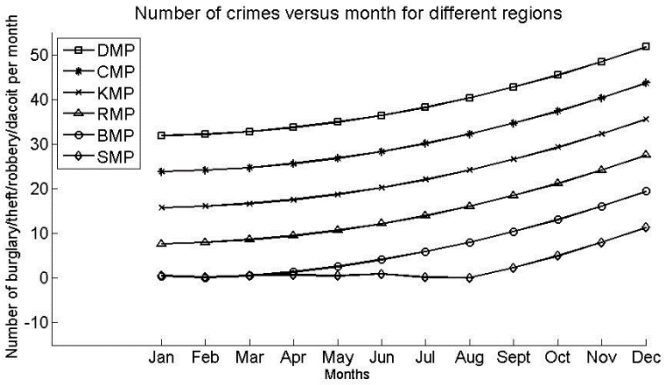

Figure 5. Burglary/theft/dacoit/robbery forecasting of 2016 at metropolitan region.

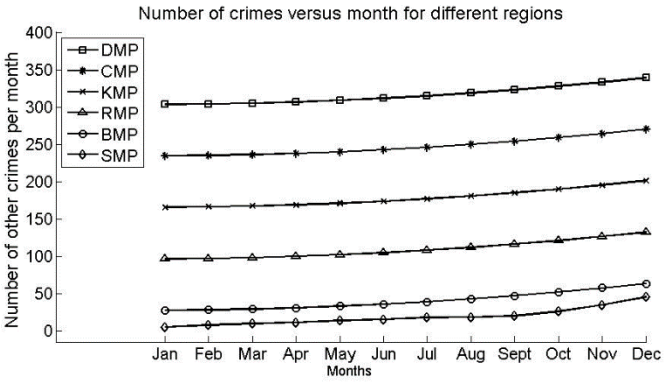

Figure 6. Other crimes forecasting of 2016 at metropolitan region.

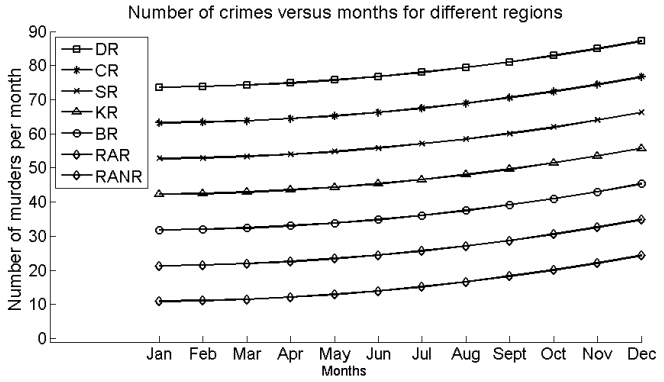

Figure 7. Murder forecasting of 2016 at divisional region.

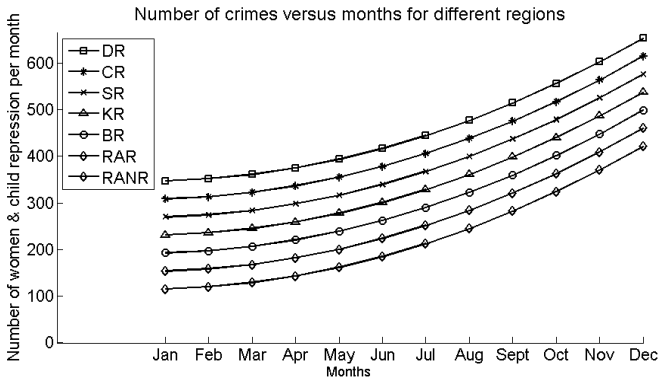

Figure 8. Women & child repression forecasting of 2016 at divisional region.

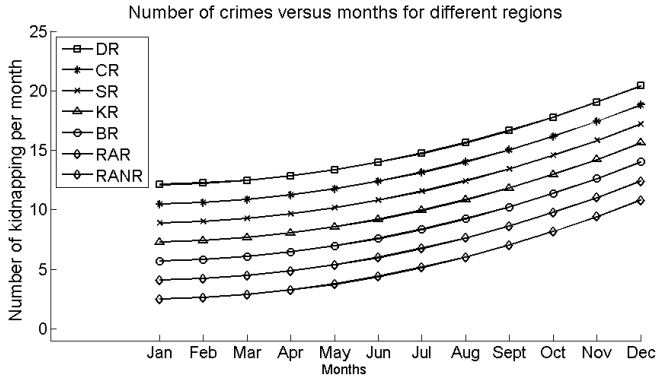

Figure 9. Kidnapping forecasting of 2016 at divisional region.

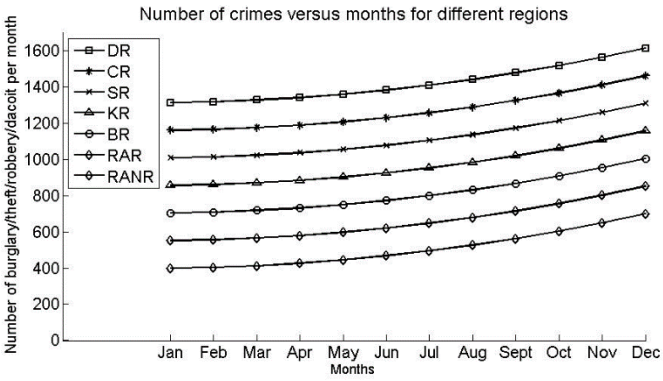

Figure 10. Burglary/theft/dacoit/robbery forecasting of 2016 at divisional region.

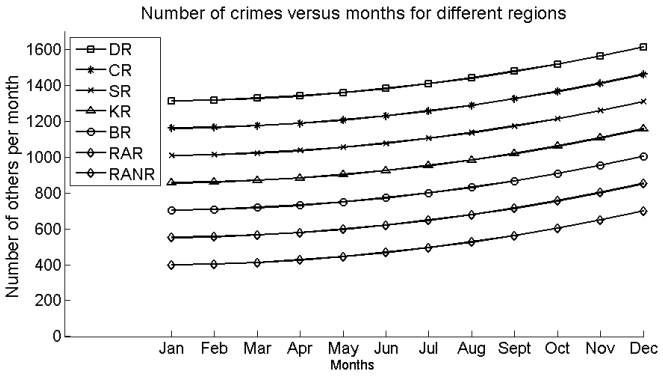

Figure 11. Other crimes forecasting of 2016 at divisional region.

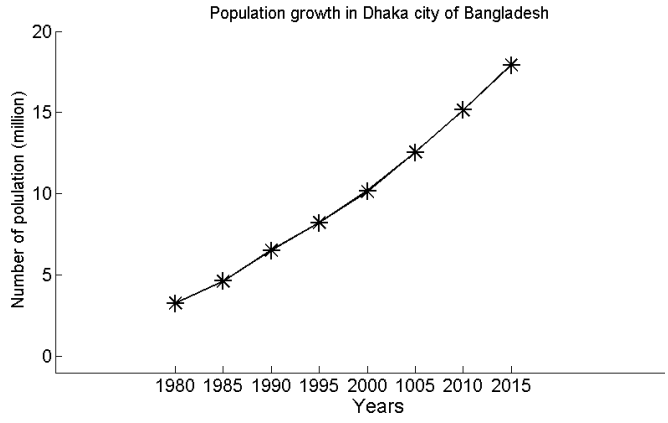

Figure 12. Population growth in Dhaka city of Bangladesh.

As depicted in fig. 12, the rate of population is increasing day by day. The comparison between fig. 2 and 12 shows that, most of the cases the crime rate is also increasing with the growth of population. Similar comparison can be shown between fig. 12 and others. Most of the cases, it shows the increase of crime with the growth of population.

V. CONCLUSIONS

At present, data mining is playing an important role in crime control and criminal suppression in many countries. In this paper, data mining technique is used to forecast future crime trends of Bangladesh. For this purpose, linear regression model is trained by crime data of previous years. After training linear regression, different types of crime are forecasted for the year of 2016. All the results are depicted in the fig. 2 to fig. 11. Table 1 shows that, how accurate the liner regression is to forecast future crime trends of Bangladesh. From the experimental result it is also observed that, most of the crimes are increasing with the growth of population. Thus the knowledge discovered from crime data analysis may assist police department and various law enforcement agencies to forecast, prevent or solve the future crime trends of Bangladesh. A future plan is to forecast the location of crime occurrence, so that prior actions can be taken to prevent crime.

REFERENCES

[1] http://www.dmp.gov.bd/application/index/page/crime-data

[2] http://www.police.gov.bd/Crime-Statistics-yearly.php?id=337

[3] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, E. M. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods", International Journal of Advanced Computer Science and Applications, The SAI Organization, Special Issue on Artificial Intelligence, pp. 18-26, 2011.

[4] N. Levine, "CrimeStat: A Spatial Statistic Program for the Analysis of Crime Incident Locations (v 2.0)", Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC, May 2002.

[5] S. Yamuna, N. S. Bhuvaneswari, "Data mining Techniques to Analyze and Predict Crimes", The International Journal of Engineering and Science, Vol.1, Issue.2, pp.243-247.

[6] A. L. Buczak, C. M. Gifford, "Fuzzy Association Rule Mining for Community Crime Pattern Discovery", In ACM SIGKDD Workshop on Intelligence and Security Informatics (ISIKDD' 10), 2012.

[7] J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", vol. 5. Morgan Kaufmann Publishers, USA, 2012.

[8] S. Shojaee, A. Mustapha, F. Sidi, M. A. Jabar, "A study on classification learning algorithms to predict crime status", In: International Journal of Digital Content Technology and its Applications (JDCTA) 7.9 (May 2013), pp. 361–369, issn: 1975-9339.

[9] A. Malathi, S. S. Baboo, "Enhanced Algorithms to Identify Change in Crime Patterns", International Journal of Combinatorial Optimization Problems and

Informatics, Aztec Dragon Academic Publishing, vol. 2, no.3, pp. 32-38, 2011.

[10] L. McClendon, N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data", Machine Learning and Applications: An International Journal (MLAIJ) vol. 2, no. 1, March 2015.

[11] C. H. Yu, M. W. Ward, M. Morabito, W. Ding, "Crime Forecasting Using Data Mining Techniques", In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW' 11), pp. 779-786, 2011.

[12] S. V. Nath, "Crime pattern detection using data mining", in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 41-44, 2006.

[13] S. Chainey, L. Tompson, S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime", Security Journal, 21, 4–28, [104].

[14] J. S. D. Bruin, T. K. Cocx, W. A. Kosters, J. Laros, J. N. Kok, "Data mining approaches to criminal career analysis", in Proceedings of the Sixth International Conference on Data Mining (ICDM'06), pp. 171-177, 2006.

[15] P. Thongtae, S. Srisuk, "An Analysis of Data Mining Applications in Crime Domain", In Proceedings of the IEEE International Conference on Computer and Information Technology Workshops, pp. 122-126, 2006.

[16] S. Bagui, "An Approach to Mining Crime Patterns", International Journal of Data Warehousing and Mining, 2, 1, pp. 50–80, 2006.

[17] T. Abraham, O. D. Vel, "Investigative Profiling with Computer Forensic Log Data and Association Rules", In Proceedings of the IEEE International Conference on Data Mining (ICDM'02), pp. 11 – 18, 2002.

[18] P. Phillips, I. Lee, "Crime Analysis through Spatial Areal Aggregated Density Patterns", GeoInformatica, Springer, vol. 15, no. 1, pp. 49-74, 2011.

[19] S. M. Nirkhi, R. V. Dharaskar, V. M. Thakre. "Data Mining : A Prospective Approach for Digital Forensics", International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 6 pp. 41-48, 2012.

[20] https://en.wikipedia.org/wiki/Feature_scaling.

[21] R. Wortley, L. Mazerolle, "Environmental Criminology and Crime Analysis", Willan Publishing, UK, 2008.

[22] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: Towards crime prediction from demographics and mobile data", In Proceedings of the ACM ICMI, to appear, (2014).

[23] E. Ferrara, P. D. Meo, S. Catanese, and G. Fiumara, "Detecting criminal organizations in mobile phone networks", Expert Systems with Applications 41, 5733–5750 (2014).

[24] J. R. Zipkin, M. B. Short, and A. L. Bertozzi, "Cops on the dots in a mathematical model of urban crime and police response", Disc Cont Dyn Syst, B 19 (2014).