Crime Forecasting for City of Portland

Deepak Shanmugam, Nitesh Srivatsav, Meghana Pochiraju
Department of Computer Science
University of Texas at Dallas
Richardson, USA

Suneesha Kudipudi, Vignesh Vijayakumar, Vinayaka Raju Gopal
Department of Computer Science
University of Texas at Dallas
Richardson, USA

***Abstract*— Over the course of the last decade, Crime has seen a steady increase across all major cities in America, and this creates an uneasiness among the public. With the help of cutting edge computing technologies, we now have access to a tool that uses machine learning techniques to warn us of the number of crimes which may possibly occur for a specific instance of time. In this project, we develop an unsupervised machine learning technique with an external knowledge base, as our input to the Forecasting system which provides us with a specific number for a given location. We use the Spark MLIB for clustering the dataset for creating a structured input for our forecasting system. The forecasting system puts the ARIMA model to use.**

Keywords*— *Machine Learning, Autoregressive moving average model, Big Data, Apache Spark, Forecasting, Pyplot

I. INTRODUCTION

Crimes are usually committed in counties where the average per capita income is lower than the average for the entire state. These counties demonstrate a higher tendency for the number of crimes for a given instance of time. Our idea is to forecast a number for every county in the city of Portland so that it would benefit the local Police Department in allocating members in advance thereby preventing or minimizing damage. Human forecasting has always been a norm for several ages but it uses obsolete parameters in judging the state of affairs in each county and involves a lot of difficulty. Due to the advance in technologies, we can leave the task of predicting these numbers to computers which put some of the cutting-edge algorithms like ARIMA to effective use, thereby saving time and money.

We have a framework which implements the idea of this project. After getting the uncleaned data from the internet, we transform and manipulate the data to match our requirements. We then make use of the Spark MLIB to generate clusters. We cluster the data based on the co-ordinates available in the dataset using the K-means algorithm. We consider clustering to be necessary since it eliminates redundancy by grouping points closer to one another in the same cluster. We store these results on the local system. Finally, we make use of the ARIMA model for forecasting the clustered data. The implementation is available in the Spark-TS library. Since the dataset is of considerable size, we need a Big Data technology to handle it, and we have gone ahead with Spark, since it is superior to Hadoop, Hbase, Mahout considering performance parameters.

To give, a summary of the rest of the report, Section 2 gives us a detailed overview of the different tools used in the project. Sections 3 and 4 includes the experiment details and results followed by Conclusion and future works. Section 6 consists of bibliography.

II. BACKGROUND

Data Transformer is one of the entities of the framework which is responsible for grouping data based on location parameters to prepare the data for clustering.

Apache Spark is an open-source distributed framework for data analytics. It avoids the I/O bottleneck of the conventional two-stage MapReduce programs by providing in-memory cluster computing. Also, it supports both batch processing and streaming data

Spark MLib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction

ARIMA- In the statistical analysis of time series, autoregressive–moving-average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average.

MatLab PyPlot- Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. ... For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with Python.

Kibana- Kibana is an open source data visualization plugin for Elasticsearch. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster. Users can create bar, line and scatter plots, or pie charts and maps on top of large volumes of data.

III. EXPERIMENT DETAILS

A. Data Transformation and Manipulation

To cluster the data, we need to do some pre-processing. To prepare the data, we have written a script in Python/Scala which groups the data based on the location and time

parameters available in the dataset. We are doing this step to calculate the sum of the crimes occurring per location at any instance of time.

B. K-Means Clustering

After proper pre-processing, we have the data ready to be clustered. As explained previously, we cluster the data in order

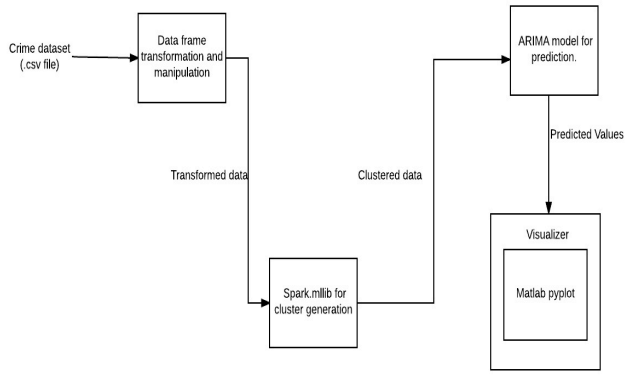

Fig. 1: Framework

to reduce the redundancy. We have imported the Spark Mlib library to do K-Means clustering. To give an overview of K-means, it is basically an algorithm used for clustering. The input to K-Means is a set of points (observations), and an integer K. The goal is to partition the input points into K distinct sets (clusters). The first step is to initialize the algorithm by choosing K initial cluster centroid locations. These steps are repeated until the algorithm "converges".

We have implemented the above algorithm on our dataset and have assigned the initial value of K to be 3. We have represented the above number since we wanted to classify the locations into regions with high, medium and low crime. After running the algorithm, we get the clustered locations and we take the centroid of the respective clusters and convert it into latitude and longitude for plotting the data in Kibana which will be explained later.

C. ARIMA

We use ARIMA model to forecast the number of crimes per location and instance of time. The ARIMA model is used widely in research and it provides a comprehensive model in the domain of time series analysis. Time series predictions are based on changes over time in historical data sets and they can produce mathematical models by using statistical data that can be extrapolated.

A variety of factors affect outside the analytics domain affects the crime rate of a given county and this leads to difficulties when forecasting using regression methods. This feature is the main advantage of time series analysis for predicting the crime rate because time series analysis can consider the effects of various factors.

In our framework, it happens to be the same case as there is no steady increase or decrease in the crime rate in our dataset over the years and hence using regression would not give an accurate forecast of the next variable. Hence, we have implemented forecasting using ARIMA where the input to the model happens to be in a format, time series format to be specific so that it can analyze the trends in the dataset over the years and then give us a forecast for the next week to ten days depending upon the statistical model generated.

D. VISUALIZATION

One of our motives is to point out forecasted data with differing levels of crimes on a plot using an open source tool namely MatLab PyPlot. As we had previously explained about Pyplot in general, to plot the forecasted data which, we had received after the clustering the locations, we need to send this data and their corresponding crime values to a Python script which can then be accessed by MatLab.

MatLab pyplot provides an interface to visualize data which is used by the python script we have written to view the data.

IV. EXPERIMENT RESULTS

A. Clustering

We had decided to go with 3 clusters initially after properly verifying using the Elbow method. Since the line curves at K=3 , we have chosen the number to be 3.

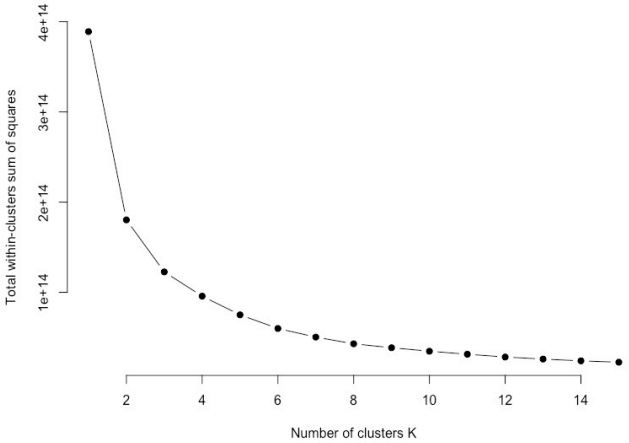

Fig. 2: Elbow Method for deriving Cluster No

After clustering, we had divided the grouped data from pre-processing into 3 clusters represented in the table below.

Clusters	Records Included
HIGH	345561
MEDIUM	344089
LOW	139734

Table 1

B. Forecasting Results

After using the ARIMA model, we can predict the outcome for a new set of data (Test data) for the month of Oct 2016.

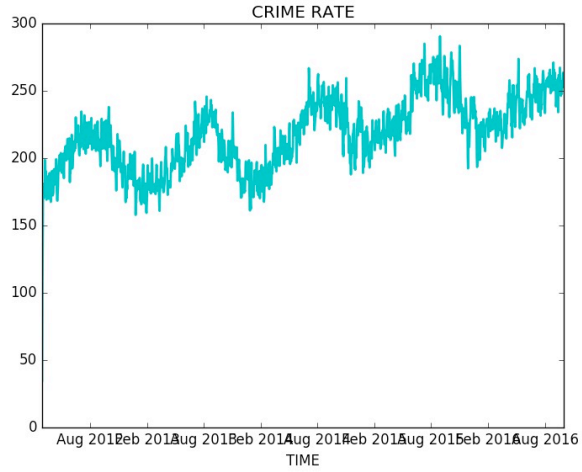

Fig. 3: ARIMA model results for High Cluster (2012-16)

Accuracy: When the forecasted value from ARIMA and the original value for the initial data (2012-16) is compared with a minimum buffer, the accuracy was found to be similar with accuracy in the range of 72%. Thus, we can conclude the model will have a similar forecasting prediction accuracy even for future values.

C. Data Visualization

The tile maps pointing out the regions with differing crime densities are pointed out in the plot.

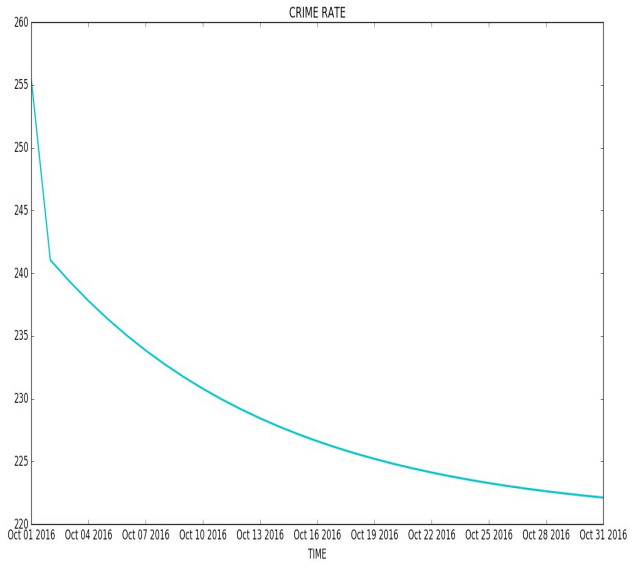

Fig. 4: Forecast for October 2016 from ARIMA

As we can see from the original data in Fig. 3, the graph summits near Aug 2016 and hence by comparison with the previous cycles, the predicted value for Oct 2016 decreases in Fig. 4 which is in accordance to the values obtained in the previous years. Hence, we can assume our method to predict the crime rate using ARIMA model to be true.

We have implemented the visualization part using Kibana to have a detailed overview of the forecasted results. The graphs in Pyplot and Kibana can then be compared to validate one another.

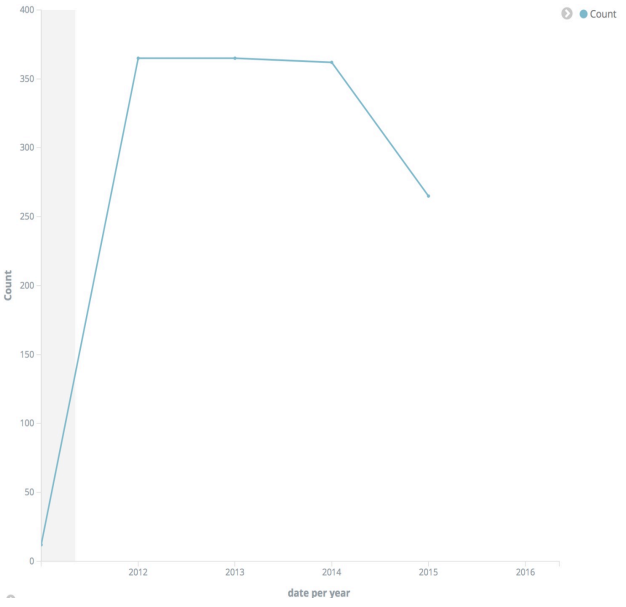

Fig. 5: Initial Data forecasted by ARIMA (Kibana)

FUTURE WORK AND CONCLUSION

We have received a forecast from the ARIMA model which we have visualized through Pyplot. This data can be invaluable to the police for staffing clusters with appropriate numbers.

In accordance to the results we have obtained using this framework, we plan to make it real time by implementing the same framework using a real-time dataset. We would also like to tweak the algorithm used for forecasting to obtain results with more accuracy.

REFERENCES

- Spark MLib
- ARIMA model
- Matlab and Pyplot

[1] Andrew Mccallum, Kamal Nigam, and Lyle. H Ungar, "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching".

[2] Srini Penchikala Big Data Processing with Apache Spark - Part 4: Spark Machine Learning https://www.infoq.com/articles/apache-spark-machine-learning

[3] Bill Haffey Predictive Analytics and Machine Learning: An Overview https://www-01.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs

[4] Spark MLlib http://spark.apache.org/mllib/

[5] https://nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx

[6] https://www.elastic.co/products/kibana

[7] USE OF ARIMA TIME SERIES AND REGRESSORS to forecast the sale of Electricity Beatrice Ugiliweneza, University of Louisville, Louisville, KY

[8] Weighting in the regression analysis of survey data with a cross-national application by Chris Skinner∗ Ben Mason.