Crime Pattern Analysis based on Machine Learning and Big Data using Apache Spark

Palash Sontakke, Chang-Soo Kim*

Department of IT Convergence and Application Engineering, Pukyong National University Busan, South Korea.

e-mail: palashsntkk65@gmail.com, cskim@pknu.ac.kr

Abstract

The global population is increasing rapidly because of increasing urbanization and such increasing urbanization directs the up-growing need of urban safety and preventions. This urbanization is also responsible for two things that is increased job opportunities and increased the crime rates. In this era technology has gone far more forward in a positive way. By making use of these technologies such as machine learning, artificial intelligence and big data we presented an approach through which crime pattern analysis is done. We have used apache spark (scala-programming) and machine learning algorithm for predictive crime pattern analysis. The data that we have used is a real-world data set based on Chicago city of United State of America. Our main goal of work is to define a predictive crime analysis which shows top crime patterns related to the top community areas of Chicago city.

Keywords: Big data, Apache spark, Machine learning, Crime analysis, Big Data Analytics.

1. Introduction

The rapid urbanization of world is already transforming the people’s lives with respect to economy and socialization. Such transformation brings more opportunities as well as concerns for management issues and urban safety. Increasing crime rates has been a major setback of the rapid urbanization and it is important to suppress the crimes rates to improvise the urban safety [1][2]. The rise in crime over the past 20 years is far more closely connected with the rise in mass unemployment. Many and new technologies are helping police departments to find and track the trends of crime. Other researches have shown that the crime rate falls out high if the cities are big and large [3]. Hence, the major challenges arrive when the urban areas are large because the existing law enforcement departments does not have the advanced tools and technologies to distinguish the meaningful ongoing crime patterns [4]. So, it is necessary to have an approach where local law enforcement departments can do the analysis at ease of efficiency in terms of crime pattern predictions.

As such technologies, we have made use of apache spark and machine learning models for analysis and crime pattern predictions. Apache spark is an open-source cluster-computing framework and provides variety of functionalities for big data processing and machine learning approaches [5]. Hence, our main approach of this study is towards demonstrating how eloquence apache spark can be in predictive crime analytics.

This paper presents the analysis of crime on actual timestamp basis and predictive crime pattern according to top 10 community area of Chicago city with the help of MLlib libraries provided by apache spark. The sections of the papers are classified in such a way that first we describe the properties of the crime data set, analysis using
apache spark. Then the prediction model, results of predictive analytics and finally the conclusion.

2. Crime Analysis and data

2.1 Data Description

For the study and experimental purpose, we used the data from Chicago police department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. This dataset reflects reported incidents of crimes that occurred in the city of Chicago from 2001 to 2018. The experimental data set have total 22 columns [6], from which we have used only following columns.

<table>
<thead>
<tr>
<th>Sr no</th>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Date</td>
<td>Occurred incident date.</td>
</tr>
<tr>
<td>2</td>
<td>IUCR</td>
<td>The Illinois Uniform Crime Reporting code. This is directly linked to the Primary type and description. See the list of IUCR codes at <a href="https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e">https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e</a></td>
</tr>
<tr>
<td>3</td>
<td>Primary Type</td>
<td>Crime description for IUCR code.</td>
</tr>
<tr>
<td>4</td>
<td>Description</td>
<td>The secondary description of the IUCR code, Subcategory for primary type.</td>
</tr>
<tr>
<td>5</td>
<td>Location Description</td>
<td>Location where incident occurred.</td>
</tr>
<tr>
<td>6</td>
<td>Arrest</td>
<td>Indicates whether arrest was made or not (Boolean).</td>
</tr>
<tr>
<td>7</td>
<td>Domestic</td>
<td>Indicates whether crime was domestic or not (Boolean).</td>
</tr>
<tr>
<td>8</td>
<td>Beat</td>
<td>It's some smallest police geographic area/indicates where incident occurred. The Chicago police department has 22 police districts.</td>
</tr>
<tr>
<td>9</td>
<td>District</td>
<td>Indicates the police district where the incident occurred.</td>
</tr>
<tr>
<td>10</td>
<td>Ward</td>
<td>The ward number (City Council district).</td>
</tr>
<tr>
<td>11</td>
<td>Community Area</td>
<td>Area where incident occurred (total 77 Community Area).</td>
</tr>
<tr>
<td>12</td>
<td>Year</td>
<td>Year of incident.</td>
</tr>
</tbody>
</table>

The data above describe also have more features such as Id number, case number, latitude, longitude and location coordinates etc but these features does not have strong correlation with respective selected features hence were not used for analysis purpose. Used data-set is publicly open and can be found on the web-based framework called plenar.io and can also be found on the public data repository managed by Kaggle. For more closure on the data-set you could refer the above open data repositories [7] [8].

2.2 Crime Analysis using Apache Spark

Apache spark is open-source framework which is highly scalable, fast and it is compatible with different programming languages such as java, Scala, python and R. We made use of such highly scalable apache spark architecture using Scala programming language [9] [10].

The crime analysis that we performed throughout the entire process is on yearly, monthly, weekly, daily and hourly basis. The main purpose of this analysis is to find out the criminal movement in top community areas of Chicago city. First analysis that we perform indicates top 10 crimes occurred throughout 18 years and the yearly count of incidents in which the arrest was made.
The performed analysis shows the increased number of crime and how the crime has been evolved throughout 18 years based on monthly, weekly, daily and hourly manner. Fig 2. Shows the time-series evolution of the total number of crimes.

To identify the crimes occurred on domestic level, we categorized the domestic attribute and found out that most of the crimes occurred are non-domestic. Fig 3 shows the domestic crime occurrence on monthly, daily and hourly basis.

As the frequency of crimes occurred at domestic level are less so, we performed an analysis to find out top four locations where crime occurred most of the time. Fig 3 indicates that maximum number crime occurred on specific location on monthly, weekly, daily and hourly basis throughout 18 years.
3. Predictive Crime analytics with spark ML libraries

Apache spark provides variety of libraries for big data processing. The relative ease of using Apache spark on large data-set makes it a great choice for big data endeavours. For predictive crime analytics we used the spark machine learning libraries [6] [10]. For the programming purposes we have used the apache spark version 2.3.0 and Scala version 2.11.

3.1 Logistic Regression

Most of the machine learning methods can be formulated as convex optimization problem. For our predictive crime analysis, we used the logistic regression model from the apache spark ml library. In general, the logistic regression apprehends a vector of variables and evaluates coefficients or weights for each input variable and then predicts the class of stated input [11] Our problem definition comes under the multinominal logistic regression where the output is probabilities of different possible outcomes of dependent variable on given set of independent variables [12] [13]. The multinominal logistic regression algorithm from apache spark ml library is based on the following mathematical formulation [10].

Where is the convex function, is the variable vector and is a loss function.

3.2 Prediction Model

Prediction model is created in such a way that can predict the most occurring crimes in respective community areas. The model is consisting of the features described in the table 1 Those features were transformed into string-indexed format for ease and smoothness of model. For tuning and increasing accuracy, the normalization is performed on the selected features [10]. Fig 4 shows the proposed architecture of prediction model.

![System Architecture](image_url)
Following is the pseudo algorithm to objectify the above system architecture:

**Pseudo Algorithm:**

**Input:** Crime data-set (training and testing)

**Output:** Positive and negative polarity [Accuracy and error rate],

 Crime pattern in top 10 community area.

**Step-1** Primary feature selection and feature grouping
a) Type casting of features (Boolean to string-type and double-type to integer-type)
b) Feature grouping for preparation of training and testing data set such that testing contains data related to top 10 community area and training contains rest of the data.
c) Split training and testing data into 70% training and 30% testing data.

**Step-2** Feature transformation and labeling process
a) Function StringIndexer(string-Type columns)
   Input -> setInputCol(column-name)
   Output -> setOutputCol(columnname + "indexed")  //feature column indexing
b) Function StringIndexer(string-Type columns)
   Input -> setInputCol(column-name)  //feature that we are going to predict
   Output -> setOutputCol(predictLabel)
c) Function VectorAssembler()
   Input -> setInputCol(all indexed features)
   Output -> setOutputCol(features)  //transforming selected features into vectors

**Step-3** Feature normalization
a) Function Normalizer()
   Input -> setInputCol(features)
   Output -> setOutputCol(norFeatures)  //normalizing each vector to have unit norm.

**Step-4** Logistic regression
a) Function logisticRegression()
   Input -> setLabelCol(predictLabel)  //output of step 2 (b) as input
   Output -> setOutputCol(predictedLabel)  //predicted labels

3.3 Results of predictive analysis

Our proposed model predicted the pattern of crime which have been occurred for maximum number of times with respective to the top community areas in the Chicago city with the prediction accuracy of 75.80% and 24.20% of evaluation error rate. The correlation among attributes is not high which is an obvious indication of obtaining a less accuracy score. Therefore, to make shift in obtained accuracy we converted the required string type feature into vector index and then we normalize each vector value to have a unit norm.

The prediction accuracy is obtained after the hyperparameter tuning of the logistic regression model. The process of hyperparameter tuning is also known as model selection. For model selection we used the cross-validation method where the number of folds is 10. *CrossValidator* begins by splitting the data a set of folds which are used as separate training and testing datasets. To evaluate a particular *ParamMap, CrossValidator* computes the average evaluation metric for 10 *ModelS* produced by fitting the Estimator on 10 different dataset pairs. [10] [14] Fig 6 shows the predictive analysis of the crime pattern occurred in different community area.
Fig. 6. Predictive crime pattern analysis.

Most of the predictive models based on crime analysis that we have studied as related study for our research study are related to finding the highest number of crimes irrespective of the community areas similarly, other crime analysis models are based on forecasting the numeric occurrence of crimes in upcoming years. On the other hand, we tried to predict the recursive crime pattern in top 10 community areas. Comparing the predictive models according to accuracy wise, due to imbalanced data our model performed to produce accuracy of 75.80% by using apache spark, the predicted values have a good agreement related to exact values, as the predicted error value is less.

4. Conclusion

In this paper we present an approach to perform crime analysis using apache spark and predictive crime pattern analytics using apache spark machine learning libraries based on logistic regression algorithm. We performed this experiment on the crime data obtained from Chicago police department’s CLEAR system which is available publicly and we successfully performed predictive crime pattern analysis with the accuracy of 75.80 %. Our predictive model allows us to analyse the increased crime trends over past 18 years in top 10 community area of Chicago city.

As an extended study of our work, in future we will try to propose a model with more accuracy. Our future scope towards this study is to build a time-series forecasting model using apache spark. Apache spark provides third party libraries for time-series analysis and forecasting. So, we will try to utilize those libraries for our future studies related to crime analysis.

References


[8] https://www.kaggle.com


