KLE Society's
KLE Technological University

Data Mining and Analysis Course Project Report

On

The Mercury Challenge

submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering in

Computer Science and Engineering

Submitted By

UNNATI BABRUWAD	**01FE16BCS219**
VAISHNAVI J	**01FE17BCS221**
VINAYAK MADHURKAR	**01FE16BCS228**
VINEET KAVISHETTY	**01FE16BCS230**

Under the guidance of

Ms. Sunita Hiremath

.

SCHOOL OF COMPUTER SCIENCE & ENGINEERING,
HUBLI – 580031 (India).

Academic year 2018-19

Contents

KLE Society's
KLE Technological University

2018 - 2019

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Course Project for the course *Data Mining and Analysis* , entitled THE MERCURY CHALLENGE is a bonafied work carried out by the student team Ms.Unnati Babruwad-01FE16BCS219, Ms.Vaishnavi J – 01FE16BCS221, Mr.Vinayak Madhurkar – 01FE16BCS228, Mr.Vineet Kavishetty – 01FE16BCS230, in fulfillment of completion of Fifth semester B. E. in Computer science and Engineering during the year 2018 – 2019. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above said programme.

Guide
Ms. Sunita Hiremath

External Viva:

Name of the Examiners	**Signature with date**
1.	
2.	

ABSTRACT

This project focuses on to predict the various activities happening in the Middle East. The activities are mainly Civil Unrest, Military activity and an Infectious disease – MERS (Middle East Respiratory Syndrome). IARPA throws a challenge to its patrons to predict these events well before the actual event occurs. In this course we mainly focus on the prediction of events related to Military Activity and MERS disease. The predictions are to be made based on the previous year's observations, i.e. from 2015 to 2018. The data analysis offers us an opportunity to develop new methods in the field of prediction of such kind of events. The project uses different methods such as clustering, decision tree classification, Apriori and classifier chains, the details of which are described in subsequent segments.

INTRODUCTION

Surprise events such as the fall of the Berlin Wall, Iraq's invasion of Kuwait, the civil unrest that gave rise to the Arab Spring, and Russian incursions into Ukraine, forced the U.S. government to respond rapidly, often in an absence of data related to the underlying causes of these events.

In an effort to provide early warning capabilities, the Department of Defense Integrated Crisis Early Warning System (ICEWS) and IARPA's Open Source Indicators (OSI) programs leveraged novel statistical and machine learning techniques using publicly available data sources to forecast societal events such as civil unrest and disease outbreaks with a high degree of accuracy. The IARPA Mercury Challenge is looking for novel and advanced methods to provide early warning for the U.S. Government of such events.

The Mercury Challenge seeks innovative solutions and methods for the automated generation of event forecasts in the Middle East.

The specific event classes of interest are:

- Military Activity (MA) in Egypt, Saudi Arabia, Iraq, Syria, Qatar, Lebanon, Jordan, and Bahrain:
    - Conflict – Incident where police, military, or other state/government security forces take action in some way; and
    - Force posture – A newsworthy action of police, military, or other state/government security forces that does not involve the use of deadly force.
- Infectious disease in Saudi Arabia: Weekly Middle East Respiratory Syndrome (MERS) count.

PROBLEM STATEMENT

The Mercury Challenge seeks methods for the automated generation of event forecasts in the Middle East.

- Military Activity
- Non-Violent Civil Unrest
- Infectious Disease

RELATED WORKS

- 'Beating the News' with EMBERS: Forecasting civil unrest using open source indicators.
    - The cascades model specifically designed to track activity on social media, especially recruitment of individuals to causes through the use of targeted campaigns, or the popularization of causes through adaptation of hash tags (Twitter).
- Predicting the international spread of Middle Eastern Respiratory Syndrome (MERS).
    - They used openly accessible data including the airline transportation network to parameterize a hazard based risk prediction model. The hazard was assumed to follow an inverse function of the effective distance (i.e. the minimum effective length of a path from origin to destination), which was calculated from the airline transportation data, from Saudi Arabia to each country. Both country specific religion and the incidence data of MERS in Saudi Arabia were used to improve model prediction.

METHODOLOGY & RESULTS

The approach towards the solution for prediction of military activity was as follows:

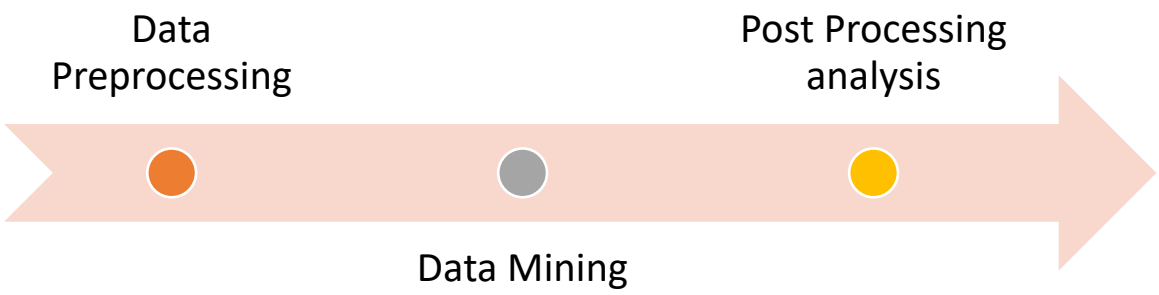

Figure 1: Data Mining Overview

MILITARY ACTIVITY

The First step towards the solution is to pre-process the data. The pre-processing of data involves:

1. Combining the data of every month into one big dataset.
2. Filling in the NULL values in every attribute.
3. Checking for Outliers in the dataset.
4. Removal of less relevant attributes.
5. Checking the relationship between different attributes of the dataset by plotting .
6. Conversion of data to appropriate format by Label encoding or One-Hot encoding.

The Second step is the Data mining .This involves:

1. Analysis of data
2. Building of a prediction model
3. Prediction

The third step involves analysis of the outputs and checking the correctness of the predictions.

DATA PREPROCESSING :

Step -1 : The data set for military activity was distributed on the basis of months .
The dataset contains military events recorded in particular months. These datasets need to be joined and create a large dataset.

Step – 2: Certain attributes contained NULL values, these values were replaced with relevant data. For example, the "State" attribute in the data-set contained null values. These were replaced by analyzing their latitude and longitude, and filling in the state name. We have used "Geopy".

Step – 3 : Relationship between different attributes is analyzed by plotting the graphs.

Step – 4 : From the graphs plot in the previous step , we keep the attributes that would contribute to the prediction.

Step – 5 : The dataset now has only relevant data , which acts as input to the model. These attributes are brought into such a form that is acceptable by the model being trained. This can be done applying One-hot encoding or Label Encoding, as per requirement.

GETTING THE MODEL READY:

Iteration -1 : Clustering

In this iteration we tried to cluster the latitude and longitude, by using K-means clustering algorithm, where-in we tried to find out the places which were highly prone to have an attack in the near future and figure out which actor was active in certain locations.

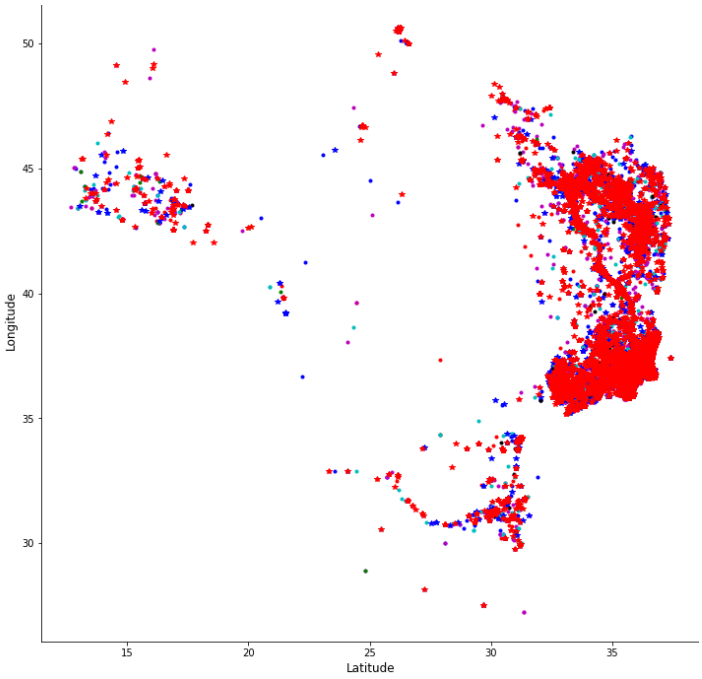

Figure 2: K-Means Clustering

Iteration -2 : Finding the patterns using Apriori algorithm

In this iteration we tried to use Apriori rules such as Support, Confidence, Lift , etc to find patterns between actors and cities, i.e. which actor was highly associated with specific cities.
The same was used to formulate the rules between Actor and Event Subtype, to find which actor was associated with either Force posture or Conflict.

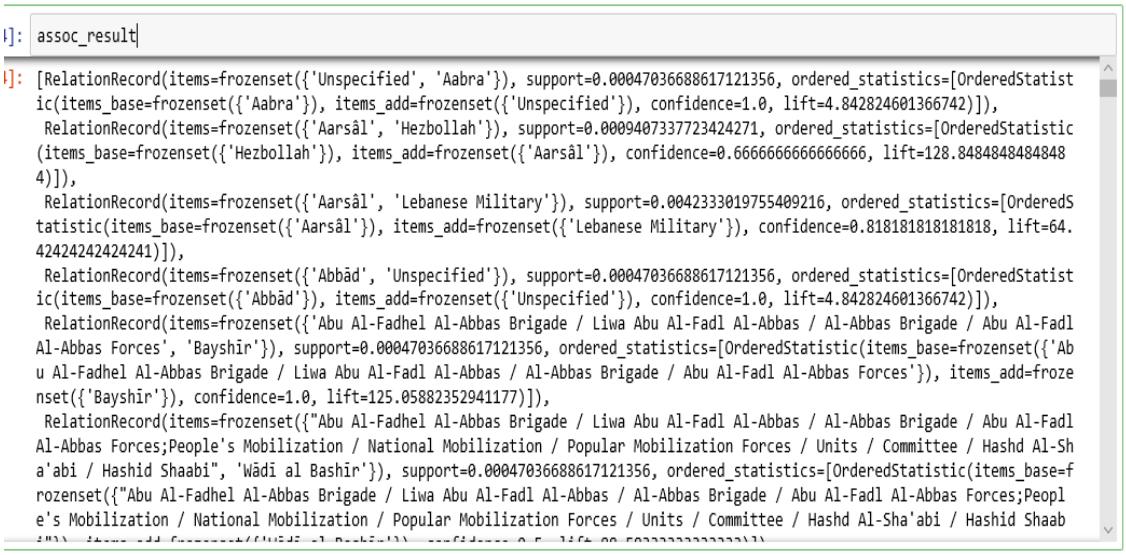

Figure 3: Association Rules [Actor , City]

Iteration -3 : Decision Tree

This iteration focuses on feeding a few attributes as input to a decision tree and predict one attribute, which collective act as inputs (inputs and output of the previous decision tree) to a next decision tree which further predicts a different attribute.

Initially, the attributes - ‘Country’, ‘Month’ and ‘Year’ are fed as input to a decision tree, and the Actor is predicted at first. Then, to the second decision tree ‘Country’, ‘Month’, ‘Year’ and ‘Actor’ are passed as inputs and ‘City’ is taken as output. On having the ‘City’ predicted, this paves way for attributes like ‘State’, ‘Latitude’ and ‘Longitude’. Lastly, all the attributes with us are given as input to another decision tree, which predicts the ‘Event_Subtype’.

Iteration – 4 : Classifier Chains

Classifier chains is a machine learning method for problem transformation in multi-label classification. The classifier is applied to input with first class label and the next classifier takes the previous classifier as input.

Classifier chains accept date and country as inputs, where in the date is feature extracted and split into day, month and year, this is used to predict other attributes.

```
pred=pd.DataFrame(predictions,columns=['News_Source','Actor_encoded','City_encoded','State_encoded','EventSubtype_encoded'])
pred
```

	News_Source	Actor_encoded	City_encoded	State_encoded	EventSubtype_encoded
0	17	189	2821	88	1
1	17	3	7834	23	1
2	26	131	1178	12	0
3	26	105	4308	12	1
4	26	320	6636	47	0
5	26	320	7025	47	0
6	32	90	5818	12	0
7	26	316	3333	38	0
8	26	320	2047	38	0
9	26	30	7364	12	1
10	17	285	5694	28	1
11	26	314	3300	50	0
12	26	126	1178	12	0

Figure 4: Classifier Chain

INFECTIOUS DISEASE - MERS

The approach towards the solution for prediction of Disease was as follows:

The First step towards the solution is to pre-process the data. The pre-processing of data involves:

1. Removal of less relevant attributes.
2. Checking the relationship between different attributes of the dataset by plotting.
3. Conversion of data to appropriate format by Label encoding or One-Hot encoding.

The Second step, is the Data mining .This involves:

4. Analysis of data
5. Building of a prediction model
6. Prediction

The third step involves, analysis of the outputs and checking the correctness of the predictions.

DATA PREPROCESSING :

Step - 1: There were no Attributes containing null values.

Step - 2: There were no Outliers detected.

Step - 3: Relationship between different attributes is analyzed by plotting the graphs.

Step - 4: From the graphs plot in the previous step , we keep the attributes that would contribute to the prediction.

Step - 5: Data Transformation to convert the categorical data into numerical.
Label Encoding and One-Hot Encoding are the methods. Label Encoding is used.

GETTING THE MODEL READY :

Iteration -1 : Time Series - Simple Exponential Smoothing

Simple Exponential Smoothing was applied to the testing data. It yielded the following results:

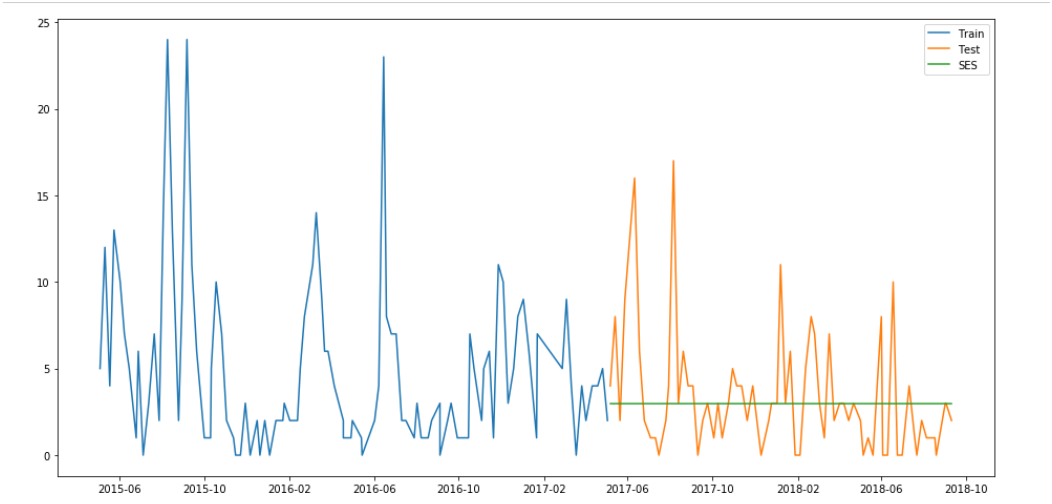

Figure 5: Simple Exponential Smoothing

The root mean square error (rms) :3.44

Iteration -2 :Time Series : Holt-Winter's Method

We observed that there is a Seasonal Trend in the Data. Certain months have high case counts. Holt- Winter's Method uses this seasonality and predicts the data accordingly. The following are the results:

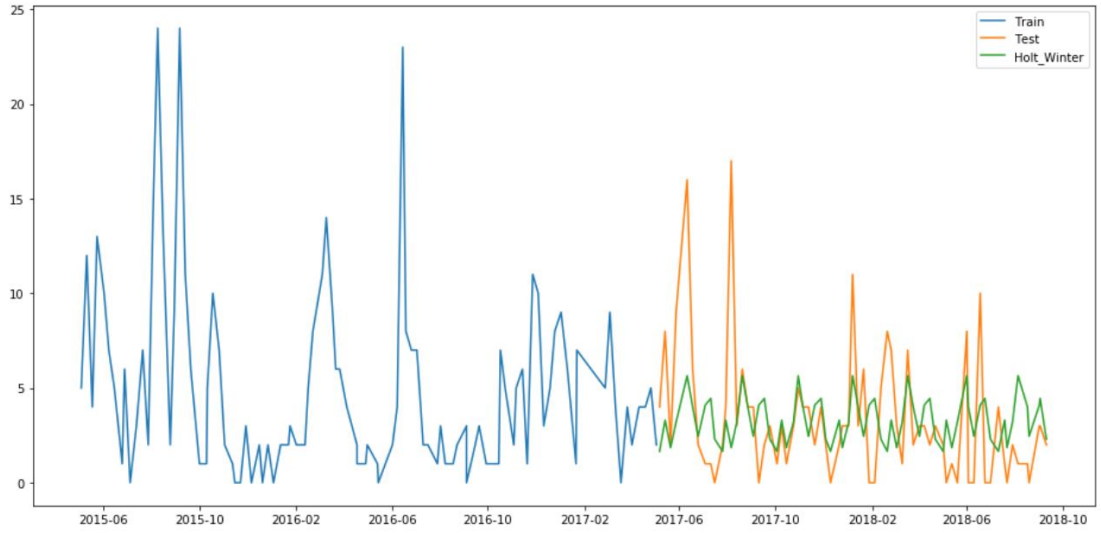

Figure 6: Results of Holts' Winter method

The RMS error value for Holt Winters method = 3.37

CONCLUSIONS

- Military Activity

Various methodologies like Clustering, Apriori was applied for the forecasting of Military Events in Middle East. Further Classification algorithms like single label classification such as Decision Tree with a layered architecture of predicting one attribute at a time and Multi-Label Classification such as Classifier Chains was applied where the output of one classifier is taken along with the input of previous classifier to predict the output of the next classifier and so on which leads to the generation of an entire event.

- Infectious Disease

The aim was to predict the Case Count of the occurrence of MERS Disease in Saudi Arabia.
As a part of Analysis we found a Seasonal Trend, where a particular time period in a year had noticeable change in the case count.
Therefore for a Seasonal Trend Data, Time Series Model such as Holt Winter's Model was applied along with other Time Series model as an Iterative Process.

REFERENCES

1. Beating the News' with EMBERS:
   Forecasting Civil Unrest using Open Source Indicators
   - Naren Ramakrishnan, Patrick Butler
2. Geopolitical Forecasting Skill in Strategic Intelligence
   - DAVID R. MANDEL

BIBLIOGRAPHY

- https://scikit-learn.org/stable/auto_examples/multioutput/plot_classifier_chain_yeast.html
- https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html
- https://medium.com/@kangeugine/hidden-markov-model-7681c22f5b9