Crime Analysis using k-means Clustering

[1] Anant Joshi, [2] A. Sai Sabitha, [3] Tanupriya Choudhury
[1] Amity University, Uttar Pradesh, [2] HoD IT, Amity University, Uttar Pradesh [3] Assistant Professor, Amity University, Uttar Pradesh
[1]anant.joshi@live.com, [2] assabitha@amity.edu, [3] tchoudhury@amity.edu

Abstract: **Analysis of crime is essential for providing safety and security to the civilian population. Using data mining, we can discover critical information which can help local authorities detect crime and areas of importance. The main purpose of this paper is to analyze the crime which entails theft, homicide and various drug offences which also include suspicious activities, noise complaints and burglar alarm by using qualitative and quantitative approach. Using K-means clustering data mining approach on a crime dataset from New South Wales region of Australia, crime rates of each type of crimes and cities with high crime rates have been found.**

Keywords-Crime analysis, statistical techniques, Data mining, Types of crimes, K-means clustering

I. Introduction

Crime Investigation is the use of facts to analyze a crime. It involves systematically studying the crime scene and the evidence gathered. There are many social, temporal, spatial and demographic factors which assist the police in the evaluation of crime. A lot of data is collected when a crime investigation is finished and this data can be used to find useful patterns. It can allows us to find geographical locations where a particular type of crime happens frequently and allow police agencies to better respond such crimes [6].

II. Theoretical Background

Analysis of crime is useful in order to recognize the occasionally contrastive trends over time in crime activity. It entails the use of statistical techniques, analytical methods and even the application of scientific social data collection to encapsulate the occurrence of any crime. There are many different types of crimes which are investigated in this manner such as theft, drug related offences, homicide, prostitution offences, robbery, blackmail and extortion. This paper aims to examine the process of crime analysis using qualitative and quantitative methods. This can give people foresight in to the type of crimes in their area and be aware of it before it happens to them [5]. Identification of the crime by observing characteristics of locations and analysis of content is called the qualitative analysis of a crime. Whereas the quantitative analysis means investigation of crime using statistics and numerical data such as rates, frequency etc.

III. Data Mining

Data Mining is an operation that analyzes data from various perspectives and summarizes or generalizes it into useful information or relationships. It is also characterized as the process of correlation discovery or discovery of patterns among fields in relational databases. Data mining as defined by Larose as an "interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization." Data mining is a very illustrious and imperative method to determine any crime. Data mining is made with the addition of two words 'data' and 'mining' in which mining includes the relation between the values of data of historical and current stipulation. Data mining is being used in many different fields to better understand and visualize data while also finding patterns and trends to enhance current information.

Crime data mining is being studied by federal, state, local, commercial and academic agencies. It aims to find patterns and trends using data collected by local law enforcement [8]. One such system was theorized in the "Crime Data Mining: A General Framework & Some Examples" paper. In this paper they discussed the framework that they have implemented. "The framework shows relationships between data mining techniques applied in criminal and intelligence analysis and the crime types listed." They applied several data mining techniques on criminal data using their framework, such as "Entity extraction, Clustering, Association, Sequential pattern mining and others" in order to achieve their goal [4][7]. According to their paper, different types of data mining techniques were used for different crime types in order to achieve the optimal result. Each technique is employed for a specific crime. They conducted three different data mining tasks on the COPLINK project dataset [1]. There are also other systems that are implementing criminal data mining tasks not only for law enforcement, but also for public use such as real estate agents, home owners etc.

978-1-5386-2529-3/17 $31.00 © 2017 IEEE
DOI 10.1109/CINE.2017.23

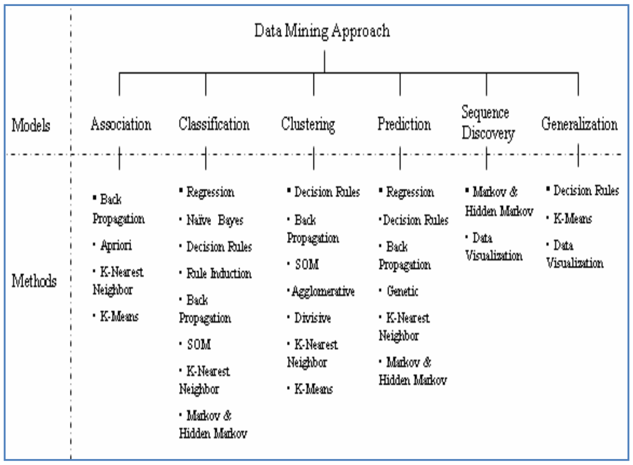


Figure 1: Data Mining Approaches

A. Classification: Classification, as the name suggests, is the process of placing or assigning the categorical variables into predefined classes. An algorithm needs to be selected to place these data in categories. The decision rules are based on training data and then used to locate these data in pre-determined groups. The rules are further validated by the validation dataset.
B. Clustering: In clustering, the data items are clustered according to their logical relationships or natural groupings and a structure as a whole is generated. There are no pre-defined groups, thus, clustering comes in the group of undirected Data Mining techniques. Each cluster is collection of homogeneous elements, which may be exclusive to that group, but are similar to each other [2]. K-means clustering is a simple clustering method that has been used in similar research [4].
C. Prediction: Predicting the future can be possible if we have enough data to work with. Especially, using data mining. Data can be mined to foretell trends, patterns and behaviors. To form a foundation for the prediction model, the previously generated decision rules which are obtained using classification or clustering are used. This is the main idea behind prediction. Therefore, all previously discussed techniques have, to some extent, the capability of prediction.
D. Sequence Discovery: This technique determines sequential patterns. Datasets can often contain many sequential patterns which are usually in the form of associations between various fields in the dataset, but these patterns are crucially based on time and often follow a particular queue. An assumption is made on the distinctness of data values.
E. Generalization: This technique forms subsets by extracting data into descriptive subsets. Entire portions of the mined data may be retrieved to create these subsets. It is also often called Summarization.
F. Association Rule mining: Datasets often have repeating data. These data form a pattern that can prove to be useful if it repeats periodically or frequently. By discovering frequently prevailing item sets in a dataset, it can extrapolate rules from patterns.

IV. Methodology

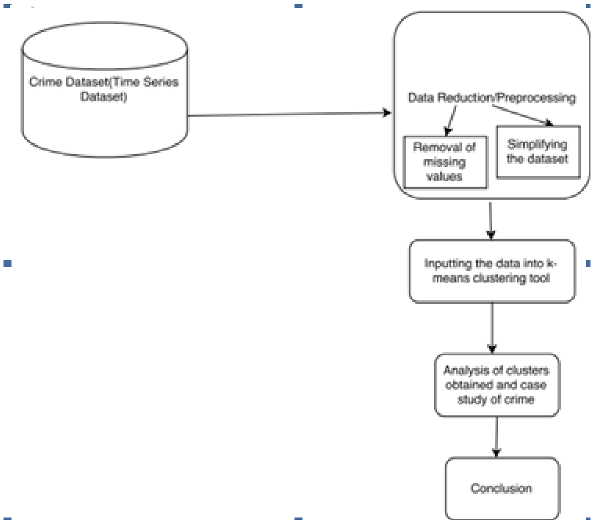

Crime Dataset(Time Series Dataset)
Data Reduction/Preprocessing
Removal of missing values
Simplifying the dataset
Inputting the data into k-means clustering tool
Analysis of clusters obtained and case study of crime
Conclusion

Figure 2: Methodology flowchart

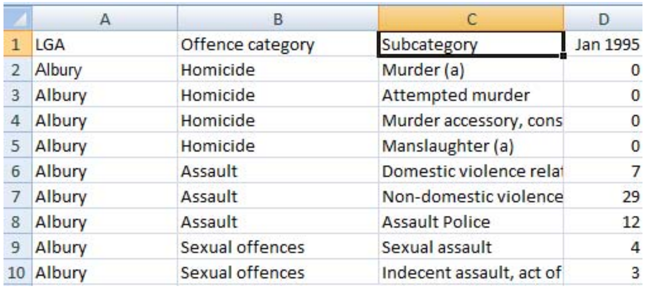
The above flowchart depicts the methodology followed for the analysis of time series crime data set.

1. Collection of dataset from the Bureau of Crime Statistics and Research of New South Wales Government's Justice Department website.
2. Preprocessing of dataset (data cleaning, data selection, data transformation)
3. Analysis of k-means using clustering tool
   a. Identification of k using silhouette measure.
   b. Inputting the data into k-means clustering tool.
4. Cluster 0 to cluster 4 obtained using k-means
5. Analysis of clusters obtained using k-means and case study of crime at various locations.

V. Experimental Setup

5.1 Data set

The dataset was collected from the Bureau of Crime Statistics and Research of New South Wales Government's Justice Department website [10]. It contains over 9000 records of crime that took place in Australia. Specifically, this dataset focuses on the New South Wales region which is located in South East Australia. This dataset further consists of attributes like Location, Sub Location, Crime Type, Crime Subtype and Year.



	A	B	C	D
1	LGA	Offence category	Subcategory	Jan 1995
2	Albury	Homicide	Murder (a)	0
3	Albury	Homicide	Attempted murder	0
4	Albury	Homicide	Murder accessory, cons	0
5	Albury	Homicide	Manslaughter (a)	0
6	Albury	Assault	Domestic violence relat	7
7	Albury	Assault	Non-domestic violence	29
8	Albury	Assault	Assault Police	12
9	Albury	Sexual offences	Sexual assault	4
10	Albury	Sexual offences	Indecent assault, act of	3

Figure 3: Crime Dataset

5.2 Meta Data

This dataset further consists of attributes like Location, Sub Location, Crime Type and year.

5.3 Tool Used

Rapid miner is a software tool that provides an integrated environment for machine learning, information mining, text mining, predictive examination and business investigation. It is used for information mining process including results representations, validations and optimizations. Rapid Miner was used to create and analysis the clusters in this paper and all results were taken from RapidMiner.

5.4 Preprocessing

1. Remove the missing value
2. Selection of relevant attributes
3. Attribute reduction
4. Addition of new attribute "total".

The data was loaded in the clustering tool (k-means). (Figure 5)

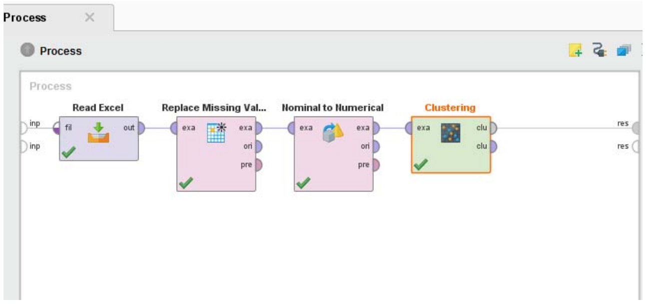


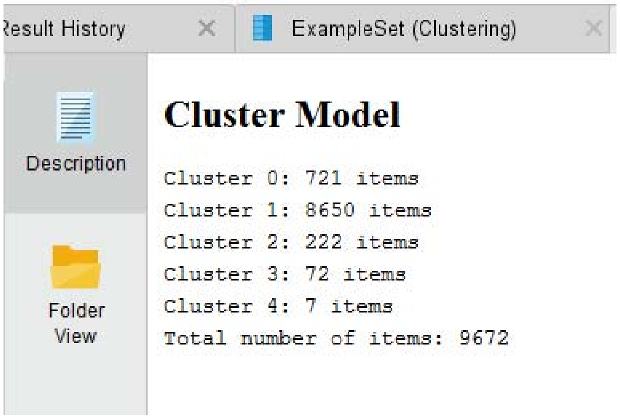
Figure 4: Process Model

5.5 Validity Measure

The value of k is 5 with max runs 10 .The measure used was mixed measures as the time series dataset considered is numerical and categorical data.

VI. Analysis and Results

Case Study 1: Cluster 1 has maximum data with 8650 records and cluster 4 has minimum data with 7 items as shown in the figure.



Cluster Model

```
Cluster 0: 721 items
Cluster 1: 8650 items
Cluster 2: 222 items
Cluster 3: 72 items
Cluster 4: 7 items
Total number of items: 9672
```

Figure 5: Cluster Model I

Crime analysis with respect to cluster_2, cluster_3 and cluster_4 are shown in in the figure.

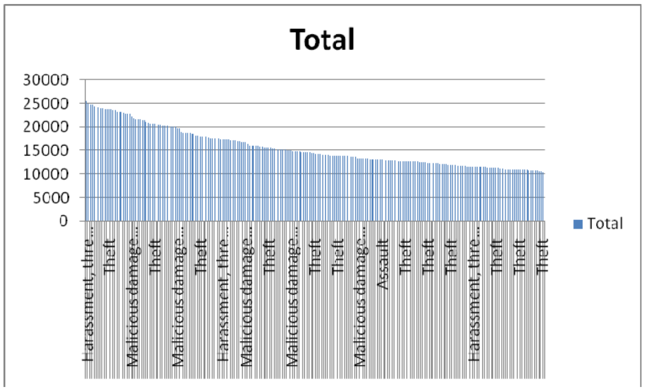


Figure 6: Offences versus total with respect to LGA (Cluster_1)

The highest criminal activities in cluster_2 are "Harassment, Threatening Behavior and Private Nuisance" of city Blacktown.

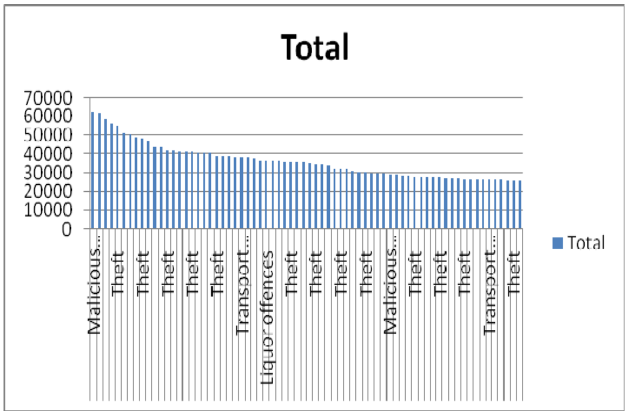


Figure 7: Offences versus total with respect to LGA (Cluster_3)

The highest crime in cluster_3 is "Malicious damage to property" in the city of Campbelltown.

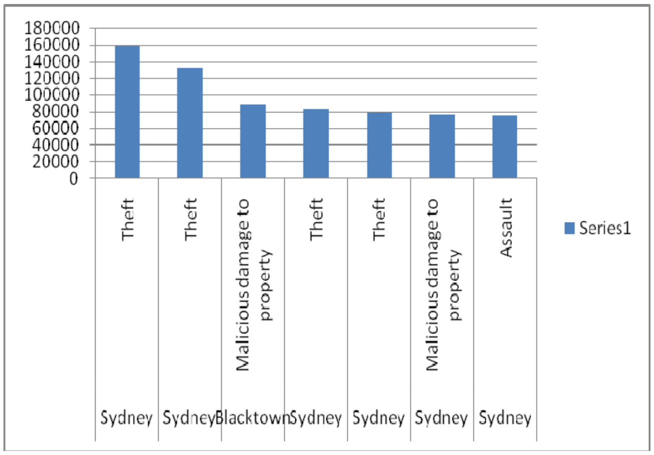


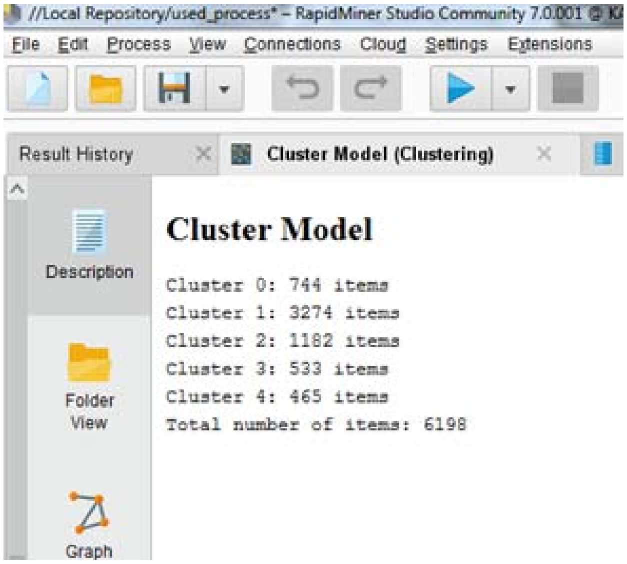
Figure 8: Offences versus total with respect to LGA (Cluster_4)

The highest crime in cluster_4 is "Theft" in Sydney, the capital of New South Wales.

Table 1: Summary of Clusters

Clusters	City Name	Type of crime	Rate
Cluster_2	Blacktown	Harassment, threatening behavior and private nuisance	Low
Cluster_3	Campbelltown	Malicious damage to property	Medium
Cluster_4	Sydney	Theft	High

Case Study 2: The cluster_1 has 8350 records. These records were further fed into the clustering algorithm to get the sub clusters. Simple analysis of these clusters shows us that Cluster 1 has maximum data with 3274 records and cluster 4 has minimum data with 465 items as shown in Figure 9.


//Local Repository/used_process* – RapidMiner Studio Community 7.0.001 @ K
File Edit Process View Connections Cloud Settings Extensions
Result History
Cluster Model (Clustering)
Description
Folder View
Graph

Cluster Model

```
Cluster 0: 744 items
Cluster 1: 3274 items
Cluster 2: 1182 items
Cluster 3: 533 items
Cluster 4: 465 items
Total number of items: 6198
```

Figure 9: Cluster Model II

Crime analysis with respect to Cluster1_0 cluster1_2, cluster 1_3 and cluster1_4 are shown in the figure below.

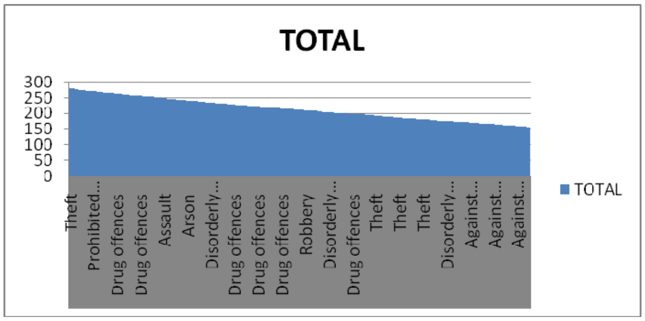


Figure 10: Offences versus total with respect to LGA (cluster1_0)

The highest crime in cluster1_0 is "Theft" in the city of Warren.

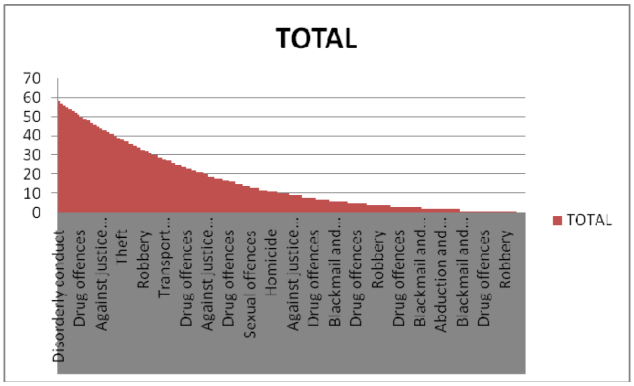


Figure 11: Offences versus total with respect to LGA (cluster1_1)

The highest crime in cluster1_1 is "Disorderly Conduct" in the Unincorporated Far West region of New South Wales.

,

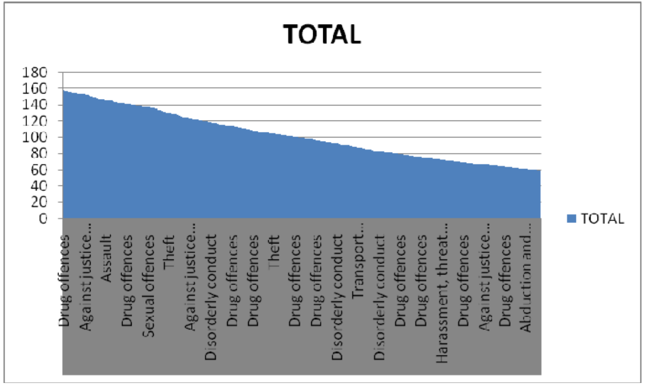


Figure 12: Offences versus total with respect to LGA (cluster1_2)

The highest crime in cluster1_2 is "Drug Offence" in the city of Orange.

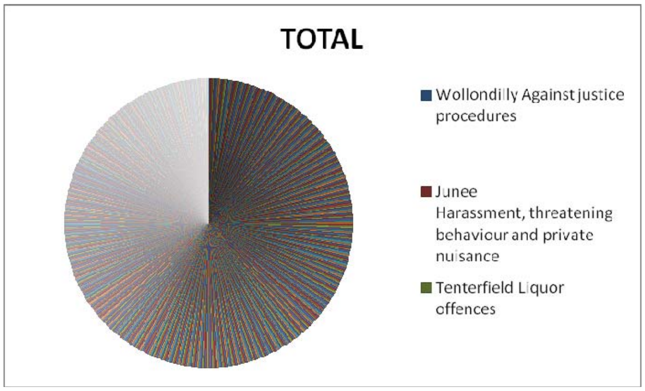


Figure 13: Offences versus total with respect to LGA (cluster1_3)

The highest crime in cluster1_3 is "Against justice procedures" in the city of Wollondilly.

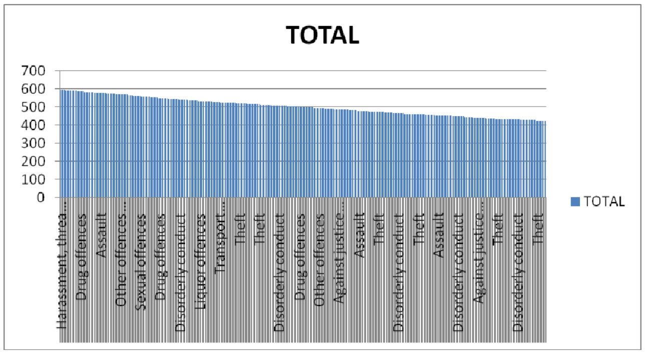


Figure 14: Offences versus total with respect to LGA (cluster1_4)

The highest crimes in cluster1_4 are "Harassment, Threatening Behavior and Private Nuisance" in the city of Lachlan.

Table 2: Table of clusters

Clusters	City Name/Region	Type of crime	Rate
Cluster1_0	Warren	Theft	Medium
Cluster1_1	Unincorporated Far West	Disorderly conduct	Low
Cluster1_2	Orange	Drug Offence	Medium
Cluster1_3	Wollondilly	Against justice procedures	High
Cluster1_4	Lachlan	Harassment, threatening behavior and private nuisance	Very High

VII. CONCLUSION

The summary of the case studies:

1. Cities with the highest crime rates are:
   1.1 Sydney
   1.2 Lachlan
   1.3 Wollondilly
2. These crimes were mostly located in the central south west region of New South Wales and in the city of Sydney.
3. As shown in the map in Figure 15, these places fall under central south west region of New South Wales.
4. As shown in Figure 16, the places of New South Wales lie in Sydney or are near Melbourne which are some of the most dangerous areas in Australia [9].
5. The most common crimes in the area are "Theft", "Drug Offences", "Harrasment, threatening behavior and private nuisance", and "Against Justice Procedures".

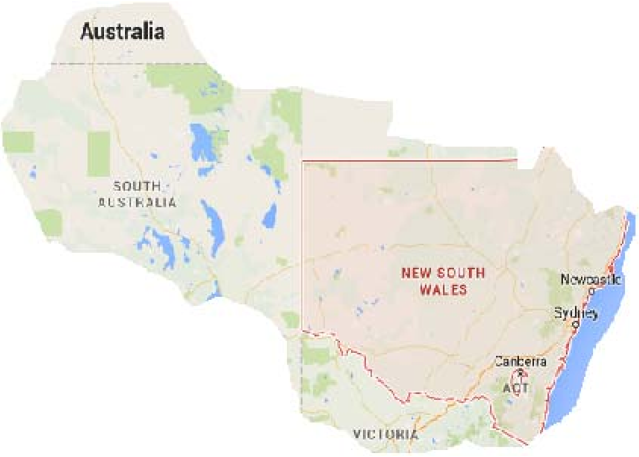


Figure 15: Map of New South Wales

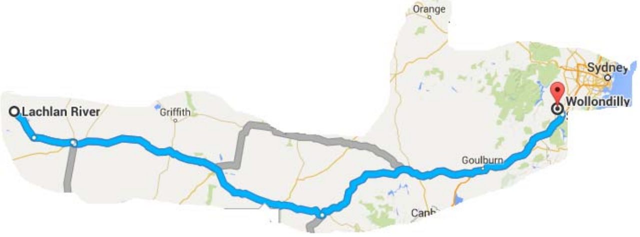


Figure 16: Map showing regions with high crime in NSW

REFERENCES

[1] Chen., H., et al., mining: an overview and case studies, in Proceedings ofthe 2003 annual national conference on Digital government research.2003, Digital Government Society of North America: Boston, MA.

[2] Fan, C., Xiao, K., Xiu, B., & Lv, G. (2014, August). A fuzzy clustering algorithm to detect criminals without prior information. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 238-243). IEEE.

[3] Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2013, June). Detecting Patterns of Crime with Series Finder. In *AAAI (Late-Breaking Developments)*.

[4] Kiani, R., Mahdavi, S., & Keshavarzi, A. (2015). Analysis and Prediction of Crimes by Clustering and Classification. *Analysis*, *4*(8).

[5] Malathi, A., & Baboo, S. S. (2011). An enhanced algorithm to predict a future crime using data mining.

[6] Corcoran, J. J., Wilson, I. D., & Ware, J. A. (2003). Predicting the geo-temporal variations of crime and disorder. *International Journal of Forecasting*, *19*(4), 623-634.

[7] Lin, S., & Brown, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Decision Support Systems*, *41*(3), 604-615.

[8] Nath, S. V. (2006, December). Crime pattern detection using data mining. In *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006*

IEEE/WIC/ACM International Conference on (pp. 41-44). IEEE.

[9] Tina Sani (2014, August 26). Top 10 Most Dangerous Cities In Australia 2014. Retrieved from http://www.onlinenewspoint.com/top-10-most-dangerous-cities-in-australia-2014/

[10] New South Wales Government, Bureau of Crime Research and Statistics crime dataset. Retrieved from http://www.bocsar.nsw.gov.au/Pages/bocsar_crime_stats/bocsar_detailedspreadsheets.aspx