TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

Serial Commercial Crime and Population Pattern Mining Using K-Means Clustering and K-Star Classification

Nik Nur Aisyah Nik Ghazali[1], Siti Norul Huda Sheikh Abdullah[*2], Siti Zaharah Abd. Rahman[3], Muhammad Ariff Abdullah[4], MdNawawi Junoh[5], Zainal Abidin Kasim[6]

[1,2,4,5]Digital Forensic Research Group, Center for Cyber Security
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia
*snhsabdullah@ukm.edu.my
[5,6]SAC at Royal Malaysia Police (RMP), Inspector General Secretariat,
R&D Royal Malaysia Police Headquarters,
50560 Bukit Aman, Kuala Lumpur, Malaysia

Article Info
Volume 81
Page Number: 4786 - 4799
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *23 December 2019*

Abstract:

Crime pattern mining has taken a center stage recently that given many governments a push to increase its power to reduce the number of crime and to prevent another case and also apprehend the criminal. In Malaysia, there is yet any research that combines both population and commercial serial crime information to retrieve inquisitive patterns. In data mining, the use of machine learning has played a crucial part not only in obtaining patterns but also understanding the hidden knowledge behind the data. The purpose of this study is to introduce the use of unsupervisedK-Means algorithm to find groups in serialcommercialcrime obtained from Royal Malaysia Police (RMP) and demographic databasesreceived from National Department of Statistics, Malaysia and a set of supervised machine learnings to obtain the best accuracy of crime pattern based on types of crime.This study is based on datasets from Selangor and Kuala Lumpur with 15857 instances ranging from January 2012 until June 2014. The results show that K-means clustering able to repopulate the crime pattern according to its criminal's demography with Cluster=11 and 17441.0within cluster sum of squared errors whereas in supervised learning, Kstar gained the highest accuracy rate upto 75.34%.

Keywords: *Crime Pattern Mining; Machine Learning; K-Means; Clustering, K-Star, Classification*

1. Introduction

Serial Commercial Crime is a growing concern for Malaysians. The increasing number of crime in Malaysia had not only taken a toll to the citizens but also foreigners. Criminal is getting more gruesome nowadays, not only the crime causes the loss of one valuables but it also involved the loss of life. Several high profile murder cases in Malaysia have made to the international news headlines such as the gruesome murders of Mongolian model [1], Myanmar nationals [2] and French tourist in Tioman Island [3].

Surveys conducted by the Performance Management and Delivery Unit (PEMANDU) showed that crime was the second largest area of concern for the people after the economy [4]. Therefore, the Malaysian Government has introduced and implemented the Reducing Crime NKRA.The Government of Ma-

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

laysia has taken several initiatives to combat crime. This could be seen by several agenda including the one highlighted in our recently Government Transformation Program (GTP). Reducing crime has become one of the National Key Result Areas (NKRAs) under Malaysian Government Transformation Program. Reducing crime has taken number one priority after fighting corruption. Under NKRA Reducing Crime, 55 initiatives have been endorsed by government to achieve two simultaneous goals: to reduce crime in the country and to restore public confidence in their own personal safety [4].

One of the most memorable events of the year 2000 for the data mining community was the "glamorous" crime buster news appeared in London's Sunday Telegraph on November 12, 2000. The news informed the world that data mining techniques can be applied in operational crime fighting. It tells a story about how a young crime analyst at Scotland Yard, from his office was able to predict by searching data for patterns and later foiled the daring robbers attempt to steal diamonds from the Millenium Dome.

Pattern mining is one of the most popular task in data mining. Initially it has been used as a market basket problem [5], as it has the ability to analyze datasets for patterns that could represent the whole dataset. Currently, as the dataset evolves into a more complex form and the number of data has increases, the pattern also changes. These new type of patterns is called sophisticated patterns [6]. Due to its complexity, these patterns need to be carefully mined. Pattern mining has successfully been applied to a various application, ranging from frequent itemset mining in transaction databases to frequent pattern based clustering and spatiotemporal data mining [7].

In the field of law and justice**,** pattern mining plays an important part especially in crime analysis process. Bruce et al.[8] had highlighted seven primary types of crimes. They are series, spree, hot prey, hot product, hot spot, hot place and hot setting. Each type is distinguished by a clear pattern that later helps police force in their tactical response.

This study is focusing on mining crime pattern and possible rules with the use of machine learning formodelling the type of crime and its relation with demographic and population of the place of crime. The paper is structured as follows: literature review in section II, materials and methods in section III, followed by experimental results in section IV and finally conclusion and future work is given in section V.

2. Literature Review

2.1.Crime and Pattern Mining

The availability of crime databases led researchers and crime analysts to study deeper into the crime behavior and thus helps investigation using data mining methods. In the past years, several researches had been undertaken in this field of study ranging from the used of data mining tools [9], [10], analysis of criminal history, social network analysis [11], geographical profiling [12] and combination of several methods [13].

Chen et al. [9] has applied clustering technique to recognize different types of crimes. This technique groups data into similar characteristics classes and finds boundary within different characteristics classes, for example, to identify suspects who conduct crimes with similar weapons or distinguish among groups belonging to different gangs. A study ongeographical profiling [13] to capture the criminal by looking at the history of the criminal activities and geography analysis on the crime scene. It applied modified decision tree to gives weightage to suspect and Genetic Algorithm to predict the criminal activities.

2.2.Current Method in Crime Prevention

Currently through a system called Police Reporting Systems (PRS), participating police stations will plot the time, location, victim of incidents upon receiving a report. Then these incidents reports are mapped into a centralized Geographical Information System (GIS) Database hence enabling the Royal Malaysian

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

Police (RMP) to visualize crime incidents nationwide and report on hot spots and trends of crime [14]. Fig. 1 shows the Standard Operating Procedure (SOP) for crime data capturing by enquiry officer (EO) and investigation officer (IO).

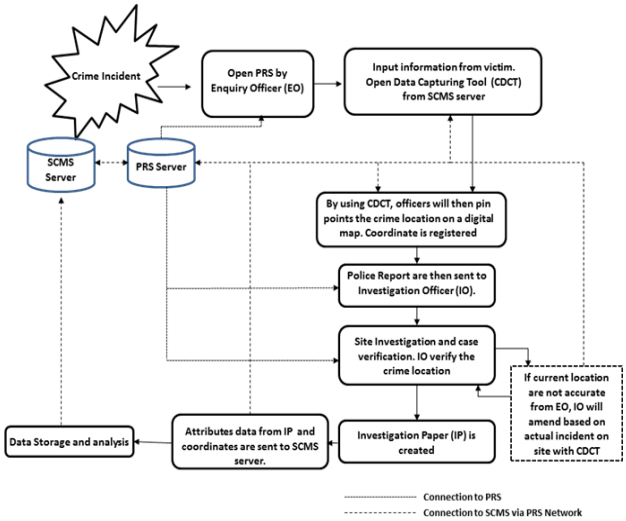

Fig. 1: Standard Operation Procedure (SOP) for crime data capturing and archiving by EO and IO.

2.3.Supervised and Unsupervised Learning

Crime pattern mining, predictions or forecasting can be developed through both qualitative and quantitative methods but it is notoriously difficult. In [15], Yu et. al. categorized crime forecasting techniques into three classes which are statistic mapping, mathematical modelling, and clustering. Clustering is an unsupervised machine learning task. Basically machine learning algorithms attempt to extract important generalized pattern from a given dataset that can be translated into important organization tasks or strategy. Learning is said to be the mixture between representation, evaluation and optimization [16].Supervised and unsupervised learning are categorized under machine learning topic. Both methods are able to overcome various problems and real world applications and had been widely used in speech recognition, computer vision, and robot control and accelerating empirical science [17-19]. Example of supervised learning is classification as conducted by [20] while clustering is a sample of unsupervised learning.The output of a classification process is a predicted type of crime based on the initial training on dataset. Meanwhile theoutput of a clustering process is several groups of classes that represent similarity among the similar instances and high differences among each group[21].

In 2013, Suzilah and Nurulhuda [22] used univariate forecasting technique to identify crime pattern in Kedah from a set of crime data from the Royal Malaysia Police (PDRM). Their study concluded and proved a hypothesis that festive seasons such as *Hari Raya Eidulfitri*, *TahunBaruCina* or Deepavali have significant association with the rise or fall of crime index. To date, this is one of the most important study on crime forecasting from the Royal Malaysian Police (PDRM) that we are able to find from the literature. However, their finding is useful as it supports Yu et. al. [15] statement that crime incident is a multi-dimensional complex phenomenon that is closely associated with temporal, spatial, societal, and ecological factors.

The clustering approach adapted by Kumar is to define the geographic boundaries of each spatial clusters [23]. With these boundaries, the changing of crime densities in a fixed size cluster is considered as the crime trend of this particular cluster. Yu [15] designs and develops the Cluster-Confidence-Rate-Boosting (CCRBoost) algorithm to efficiently select relevant local spatio-temporal patterns to construct a global crime pattern from a training set. This global crime pattern is then used for future crime prediction. The results show that the proposed CCRBoost algorithm has achieved about 80% on accuracy when tested on a data collected from a city in the United States. Besides CCRBoost [15], most of crime analytics researches were not based on Malaysian crime dataset. PredPol [24] for example, has not been tested on Malaysian crime data. In both statistics and soft computing, we need to check our assumptions before relying on a model. The “No Free Lunch” theorem [25] states that there is no one model that works best for every problem. The assumptions of a great model for one problem may not hold for another problem. The difference will be stemmed not from the data used for the predictions, but from the

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

use of different analytical techniques. Therefore, there is a need to research the effectiveness of other models and find one that works best for Malaysian Police Force.

3. Material and Methods

This section describes the datasets, information on the attributes. Later, it further discussesaboutdata pre-processing phase and machine learning methods used in this study.

3.1. Datasets Description

This study is based on datasets of apprehended criminals from four different areas in Selangor and Kuala Lumpur. The four areas chosen were Dang Wangi, Klang Utara, Klang Selatan and Kajang. These areas were chosen due to the density and number of population of the people.Table 1 shows the overall dataset and Table 2 shows the detailed dataset as below.

Table 1: Overall Dataset Description

Name of Dataset	*Apprehend*
Description	Consist of details pertaining apprehended criminals
Data set characteristics	Multivariate
Attribute characteristics	Categorical, Numerical
Total number of instances	15857
Number of attributes	10

In this study, the total number of instances is 15857. After the pre-processing steps was carried out, approximately 14322 clean data is being used.

Table 2: Detailed Dataset Descriptions

Name of Dataset	Number of Instances	Percentage of Missing Values
Apprehend 1 Klang Selatan	3423	5%
Apprehend 2 Klang Utara	1496	10%
Apprehend 3 Kajang	4577	5%
Apprehend 4 Dang Wangi	6361	4%
Apprehend 5 Klang Selatan	3423	5%

The dataset of crimes that were obtained only covering cases from Jan 2012 up until June 2014. It is noticeable thatsome inexistence information during the month of March and April 2012. Moreover, there are also missing values that mostly occurred in the *Umur* or age attribute.

3.2. Attributes Information

About 10 attributes provided by apprehended serial commercial crime database. Type of crimes are Murder(*bunuh*), Rape (*rogol*), Snatch (*Ragut*), Motorbike Theft (*Curimotosikal*), robbery with weapon (*Rompakanbersenjata*), robbery without weapon (*Rompakantanpasenjata*), night house broken (*pecahrumahmalam*) and day house broken (*pecahrumahsiang*).Table 3 shows the detailed description on each attributes. Understanding of each attribute will be helpful in order to get the maximized accuracy during the machine learning phase later on.

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

Table 3: Detailed Dataset Descriptions of Serial Commercial Crime

No	Attributes	Description	Value Descripti
1	Kontijen	State	Categorical
2	Daerah	District	Categorical
3	Balai	Police Station	Categorical
4	Tarikh KS	Date of reported crime	Continuous
5	Kesalahan	The type of crime	Categorical
6	Nama OKT	The name of criminal	Nominal
7	Jantina	Gender of the criminal	Categorical
8	Warganegara	Citizenship of the criminal	Categorical
9	Bangsa	Race of the criminal	Categorical
10	Umur	Age of criminal at the time of the crime	Continuous

3.3. Experimental Process Flow

The experiment is divided into two main phases as shown in Fig. 3 and 5 respectively. The first phase is mainly the preprocessing steps and the second phase is the machine learning steps.

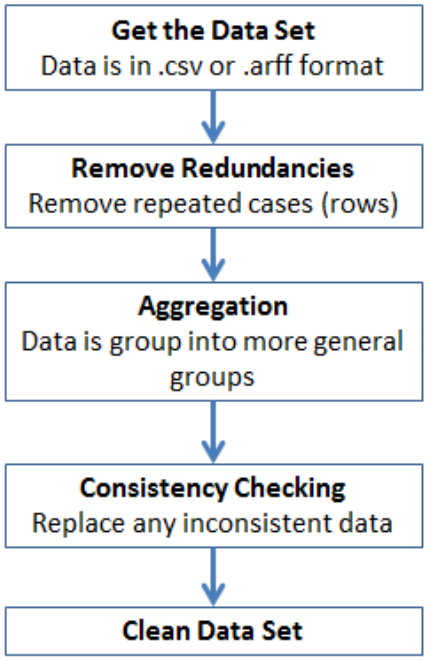

Fig. 3: Data Pre-processing Steps

Initially, the pre-processing step was conducted manually using Microsoft Excel. Fig. 3 illustrates the detail procedure. This is done for eliminating redundant data. Then, data is re-group into more general type of groups. For example, in Fig. 4, for the attribute *TarikhKes* or date of case, using the concept hierarchy, we divide the date into year and month. The same concept was also applied to the attribute *Warganegara* and *Bangsa*.

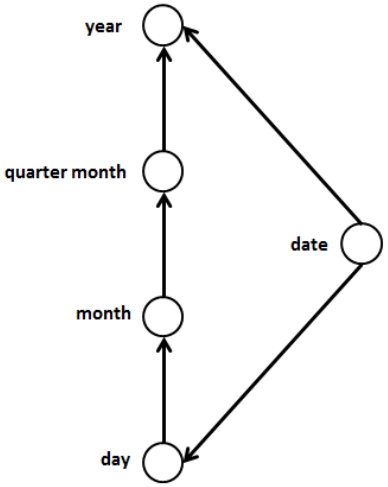

Fig. 4: Concept Hierarchy for attribute *TarikhKes*

Then, the dataset will be further generalized by categorizing the atribute*Umur*. Table 4 depicts the new group of instances used.

Table 4: Detailed Dataset Descriptions

New Group of Instances	Instances Name
< 15	Kids
15 - 20	Teens
21 – 30	Young
31 – 40	Adult
41 – 50	Old
> 50	Pensioner

After performing data transformation, the last step is to check further on its inconsistencies. Similarly, this time consuming step is performed manually. Then, the overall distribution of a cleaned data is then captured and analysed.Before undergoing to the next phase, the cleaned data will need to be filtered. WEKA software was used to filter the clean data before undergoing the machine learning phase.

Initially,about four separate datasets are combined into one main databasethe clean dataset. Then, supervised filtering techniques are applied to resam-

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

ple and further reduce the number of instances. In data sampling, we applied 40% until 90% random samplingfrom the whole set of cleaned data.

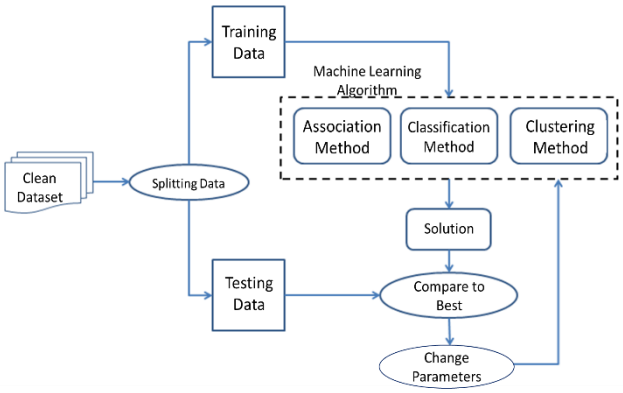

Fig. 5: Training and testing data are divided using unsupervisedand supervised machine learning methods and splitting percentage in the classification phase.

Next, missing values are replaced using unsupervised filtering techniques namelymean and mode filtering. Once the filtered dataset isfully transformed, machine learning phase is activated. Fig. 5 shows the work flow used in the second phase.The process starts with the splitting of the whole transformed data into training and testing dataset. Considering a strong tool for analysis, association method aims to investigate the crime pattern in hoping to find regularities [5].The data were then trained on both supervised and unsupervised learning algorithm such as decision table, lazy classifier, decision trees and K-means clustering.

3.4. K-Means Clustering

Clustering techniques had been used by several researchers for analysis and modelling [19-21]. In this research, clustering is used to group the serial commercial criminal and population attributes (*d* points) into group, κthat could represent object or a specific type of crime.

Start

Step 1: Consider κ points to be clustered. This will be the initial centroids.

Step 2: Each objectis assigned to the group, κ that is most similar to the centroid.

Step 3: The positions of κ centroids are recalculated after all objects, *d* have been assigned.

Step 4: Reiterate steps 2 and 3, τ until no other distinguished centroid can be found.

End

Advantages of K-means clustering are fast, robust and easy to understand. Its complexity could be calculated as follows:

$$O(\tau\kappa\eta d) \quad (1)$$

whereη is number of objects, κ is number of clusters, *d* is number of dimension of each object, and τ is the number of iterations. Normally, $\kappa, \tau, d << \eta$.

3.4. K-Star Classification

K-Star is a simple, instance based classifier, similar to K-Nearest Neighbor (K-NN). New serial commercial crime and population data instances, *x*, are assigned to the class that occurs most frequently amongst the k-nearest data type of crime points, $\boldsymbol{y_j}$, where $j = 1, 2 \ldots k$, k=Entropic distance is then used to retrieve the most similar instances from the data set. The K* function can be calculated as:

$$\boldsymbol{K*(y_j, x) = -lnP*(y_j, x)} \quad (2)$$

where $\boldsymbol{P*}$ is the probability of all transformational paths from instance *x* to *y*. In order to do a category prediction, the probability of an instance a being in category C is calculated by summing the probabilities from a to each instance from serial commercial crime and population that is a member of C

$$\boldsymbol{P*(c|a) = \Sigma_{b\epsilon c} P \times (b|a)} \quad (3)$$

The probabilities for each category are calculated. The relative probabilities obtained give an estimate

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

of the category distribution at the point of the instance space represented by *a*. The category with the highest probability will be chosen as the classification of the new instance.

4. Experimental Result

We divide this section into three parts. It begins with an Analysis on Serial Commercial Crime Versus Population Analysis, and followed by Serial Commercial Crime using K-means Clustering and Supervised Learning subsequently.

4.1 Serial Commercial Crime versus Population Analysis

Table 5 and Fig. 6 shows the initial result of the clean dataset belonging to the four databases.

Table 5: Overall Distribution of Crime among Districts

	Total	2012	2013	2014	
District Kajang		865	2113	1310	4288
Bangi	73	6	22	45	
Batu 14 Ulu Langat	99	24	52	23	
Batu 18 Ulu Langat	22	4	15	3	
Batu 9 Cheras	679	170	334	175	
Bdr Baru Bangi	175	10	108	57	
Beranang	86	6	38	42	
Kajang	2885	596	1385	904	
Semenyih	269	49	159	61	
District Klang Utara		126	679	482	1287
Kapar	198	69	75	54	
Meru	128	13	68	47	
Sg Kapar	470	29	281	160	
Bandar Baru Klang	282	9	149	124	
Bandar Sultan Suleiman	60	1	27	32	
Bukit Raja	149	5	79	65	
District Klang Selatan		102	1688	856	2646
Klang	2095	37	1364	694	
Pandamaran	329	20	229	80	
Pelabuhan Kelang	222	45	95	82	
District Dang Wangi		1335	3317	1449	6101
Chow Kit	156	0	0	156	
Jln Bandar (H.S.LEE)	3153	626	1880	647	
Jln Dang Wangi	2792	709	1437	646	

The dataset is distributed normally from 2012 up to 2014. The highest number of crime could be seen in Area Dang Wangi district, while the lowest numbers of crime are shown by Area Klang Utara. Please take note that the cases in 2014 only covers till the month of June.Table 6 tabulates detail total number of crime cases pertaining to its types of crime and districts. Type of crimes are Murder(*bunuh*), Rape (r*ogol*), Snatch (*Ragut*), Motorbike Theft (*Curimotosikal*), armed robbery (*Rompakanbersenjata*), unarmed robbery (*Rompakantanpasenjata*), night house broken (*pecahrumahmalam*) and day house broken (*pecahrumahsiang*). RPM divides district according to Kajang, klang Utara, Klang Selatan, and Dang Wangi. Furthermore, the number of apprehended criminals seem to higher in Kajang compared to Klang Utara danKlang Selatan. Apart from that, it is notable that motorbike theft and robbery without weapon have shown a significant increment in the Kajang and followed by dang Wangi and Klang Selatan.

On the other hand, Table 7 is the total of population reported from Department of Statistics (DOS), Malaysia and the percentage of crimes according to its demographics and districts. Here, DOS reported only three districts Kajang, Klang Utara and Klang Selatan. Unlike PDRM statistics, we can observe that Klang Selatan has higher population compared to Kajang and Klang Utara. Table 8 summarizes the percentage of serial commercial crime among total number of population. It is noticeable that Kajang has also shown the highest rate of serial commercial crime approximately 1.25% compared to other districts.Stealing, Unarmed roberry, and-Night house broken are showing significant occurrence rate ascendingly with 0.388%, 0.340% and 0.222% in between year 2012 and 2014. Besides that, armed roberry type of crime is the only district recorded on apprehended criminals with the percentage of crime rate over population merely 0.222%.

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

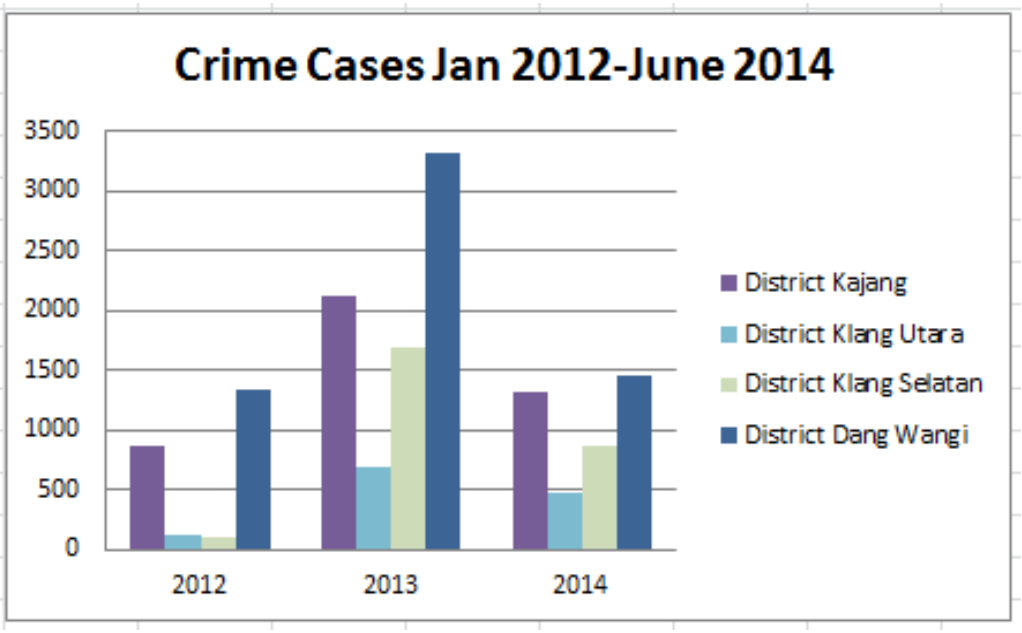

Fig. 6: Distribution of Crime Cases

Table 6: Serial Commercial Crime Statistic of the Chosen Districts

	Total	Murder	Rape	Snatch	Steal	Armed Robbery	Unarmed Robbery	Broke house night	Broke house noon	
District Kajang										**4288**
Bangi	73	1	3	0	23	0	7	14	20	
Batu 14 Ulu Langat	99	1	4	0	53	0	16	20	3	
Batu 18 Ulu Langat	22	0	4	0	13	0	1	3	0	
Batu 9 Cheras	679	16	15	13	208	11	159	127	45	
Bdr Baru Bangi	175	1	3	1	69	1	19	44	32	
Beranang	86	1	2	0	37	0	22	18	3	
Kajang	2885	69	95	50	844	51	897	469	184	
Semenyih	269	0	8	3	83	7	46	67	23	
		89	**134**	**67**	**1330**	**70**	**1167**	**762**	**310**	
District Klang Utara										**1287**
Kapar	198	5	3	2	88	0	41	10	4	
Meru	128	2	8	1	63	0	24	14	3	
Sg Kapar	470	9	10	3	146	0	176	45	8	
Bandar Baru Klang	282	14	3	4	86	0	62	78	14	
Bandar Sultan Suleima	60	0	0	0	21	0	19	7	3	
Bukit Raja	149	9	1	2	39	0	32	10	5	
		39	**25**	**12**	**443**	**0**	**354**	**164**	**37**	
District Klang Selatan										**2646**
Klang	2095	40	52	34	679	0	542	427	121	
Pandamaran	329	7	12	2	108	0	47	92	29	
Pelabuhan Kelang	222	0	2	6	98	0	33	50	10	
		47	**66**	**42**	**885**	**0**	**622**	**569**	**160**	
District Dang Wangi										**6101**
Chow Kit	156	7	0	4	77	0	46	9	1	
Jln Bandar (H.S.LEE)	3153	36	59	453	934	8	1281	135	31	
Jln Dang Wangi	2792	36	34	123	1127	16	1038	131	37	
		79	**93**	**580**	**2138**	**24**	**2365**	**275**	**69**	

Table 7: Detail demography versus total of population obtained source from Department of Statistics, Malaysia.

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

	Kajang		Klang Utara		Klang Selatan	
Race: Malay	206,805	60.4%	136805	50.0%	232,254	40.8%
Race: Chinese	65,992	19.3%	67363	24.6%	152,582	26.8%
Race: Indian	33,536	9.8%	40274	14.7%	121,533	21.4%
Race: Other Bumiputra	3,019	0.9%	1123	0.4%	11,146	2.0%
Race: Others	2,433	0.7%	1066	0.4%	2,994	0.5%
Citizen: Malaysian citizen	311,785	91.0%	246631	90.2%	520,509	91.5%
Citizen: Non-Malaysian citizen	30,872	9.0%	26808	9.8%	48,198	8.5%
Age: 10-14	25546	7.5%	24891	9.1%	47940	8.4%
Age: 15-24	79283	23.1%	51249	18.7%	103376	18.2%
Age: 25-34	75724	22.1%	58190	21.3%	121455	21.4%
Age: 35-44	46378	13.5%	40107	14.7%	86022	15.1%
Age: 45-54	30407	8.9%	26836	9.8%	57308	10.1%
Age: >54	85319	24.9%	72166	26.4%	152606	26.8%
Sex: Male	176206	51.4%	147057	53.8%	300217	52.8%
Sex: Female	166451	48.6%	126382	46.2%	268490	47.2%
Household	80785		64350		137644	
Living quarters	92000		76561		157478	
Total population	342,657		273439		568,707	

Table 8: Percentage of total serial commercial crime and total number of populations

	Total Population	**Total crime case**	**Murder**	**Rape**	**Snatch**	**Steal**	**Armed Robbery**	**Unarmed Roberry**	**Night Broken House**	**Day Broken House**
Kajang	342,657	4288	89	134	67	1330	70	1167	762	310
Percentage		**1.25%**	**0.025%**	**0.039%**	**0.02%**	**0.388%**	**0.204%**	**0.34%**	**0.222%**	**0.094%**
Klang Utara	273,439	1287	39	25	12	443	0	354	164	37
Percentage		0.47%	0.0143%	0.00037%	0.004%	0.162%	0	0.129%		
Klang Selatan	568,707	2646	47	66	42	885	0	622	569	160
Percentage		0.47%	0.0082%	0.012%	0.01%	0.16%	0	0.11%	0.1%	0.03%

4.2 Serial Commercial Crime using K-means Clustering

Here are some of the preliminary results obtained after the supervised and unsupervised learning process.For the unsupervised learning, K-means clustering technique is chosen due to the fact that crime usually varies in nature and changes over time. We conductseveral experiments starting from 2 to 11 clusters as shown in table9. Table 10shows the output of K-means algorithm based on 4, 8 and 11 clusters respectively. Each cluster represents a pattern and the number of cases groups belongs to respective cluster. The dataset is divided into 66% training and the rest is for testing.

Table 10 depicts the experiment output based on 6527 instances during testing, while table 9, shows some samples of output based on 4, 8 and 11 clusters. The numbers of clusters are chosen randomly. From table 10(a) is there are four clusters that represent four types of crime with its possible attributes that link to it. In Table 10(b), motorcycle thefts were mostly Malay male and could be either at the age group of 15 to 30 years old. On the other hand, Table 10(c) shows that snatch theft category is usually committed by a noncitizen age around 31 to 40 years old.Additional, from this clustering results, we may also conclude that type of serial commercial crime are seasonal occurrence. For example, gang robbery type of crime is usually appearing or occuring during January, May and October. On top of that, Stealing motorbike is common activity during May and June (mid-year school holiday season) that had been committed among the adults or teen criminals. Therefore, the police as well community should able to plan better and adaptive crime prevention activities according to those season.

Table 9: Results after applying K-Means Clustering Algorithm onto Commercial Crime Datasets

Number of Cluster	**Iteration s**	**Within cluster sum of squared errors**
2	3	28424.0
3	4	22886.0
4	4	22532.0
5	3	22356.0
6	3	22176.0
7	3	20752.0

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

8	3	20362.0
9	3	20251.0
10	4	17902.0
11	3	17441.0

Table 10: Result of K-Means Clustering for (a) 4, (b) 8 and (c) 11 number of cluster respectively

(a)

	Cluster 1 (1621)	Cluster 2 (1591)	Cluster 3 (3007)	Cluster 4 (308)
Kontijen	Kuala Lumpur	Kuala Lumpur	Selangor	Kuala Lumpur
Daerah	Dang Wangi	Dang Wangi	Kajang	Dang Wangi
Balai	Jln Bandar	Jln Dang Wangi	Kajang	Jln Bandar
Bulan	Januari	Mei	Januari	Oktober
Jantina	Lelaki	Lelaki	Lelaki	Lelaki
Kump Bangsa	W.Asing	Melayu	India	Melayu
Kump Umur	Young	Adult	Young	Young
Kesalahan	Samun Berkawan	Curi Motosikal	Samun	Samun Berkawan

(b)

	Cluster 1 (1160)	Cluster 2 (852)	Cluster 3 (2478)	Cluster 4 (250)
Kontijen	Kuala Lumpur	Kuala Lumpur	Selangor	Kuala Lumpur
Daerah	Dang Wangi	Dang Wangi	Kajang	Dang Wangi
Balai	Jln Bandar	Jln Dang Wangi	Kajang	Jln Bandar
Bulan	Januari	Mei	Januari	Oktober
Jantina	Lelaki	Lelaki	Lelaki	Lelaki
Kump Bangsa	W.Asing	Melayu	India	Melayu
Kump Umur	Young	Adult	Adult	Young
Kesalahan	Samun Berkawan	Samun Berkawan	Samun	Samun Berkawan
	Cluster 5 (231)	**Cluster 6 (175)**	**Cluster 7 (1008)**	**Cluster 8 (328)**
Kontijen	Kuala Lumpur	Kuala Lumpur	Selangor	Kuala Lumpur
Daerah	Dang Wangi	Dang Wangi	Kajang	Dang Wangi
Balai	Jln Bandar	Jln Bandar	Kajang	Jln Dang Wangi
Bulan	November	April	Jun	Mei
Jantina	Lelaki	Lelaki	Lelaki	Lelaki
Kump Bangsa	W.Asing	W.Asing	Melayu	Melayu
Kump Umur	Adult	Young	Teens	Young
Kesalahan	Samun	Samun	Curi Motosikal	Curi Motosikal

(c)

	Cluster 1 (1031)	Cluster 2 (761)	Cluster 3 (1581)	Cluster 4 (243)
Kontijen	Kuala Lumpur	Kuala Lumpur	Selangor	Kuala Lumpur
Daerah	Dang Wangi	Dang Wangi	Kajang	Dang Wangi
Balai	Jln Bandar	Jln Dang Wangi	Kajang	Jln Bandar
Bulan	Januari	Mei	Januari	Oktober
Jantina	Lelaki	Lelaki	Lelaki	Lelaki
Kump Bangsa	W.Asing	Melayu	India	Melayu
Kump Umur	Young	Adult	Adult	Young
Kesalahan	Samun Berkawan	Samun Berkawan	Samun	Samun Berkawan
	Cluster 5 (231)	**Cluster 6 (175)**	**Cluster 7 (849)**	**Cluster 8 (329)**
Kontijen	Kuala Lumpur	Kuala Lumpur	Selangor	Kuala Lumpur
Daerah	Dang Wangi	Dang Wangi	Kajang	Dang Wangi
Balai	Jln Bandar	Jln Bandar	Kajang	Jln Dang Wangi
Bulan	November	April	September	Mei
Jantina	Lelaki	Lelaki	Lelaki	Lelaki
Kump Bangsa	W.Asing	W.Asing	Melayu	Melayu
Kump Umur	Adult	Young	Young	Young
Kesalahan	Samun	Samun	Curi Motosikal	Curi Motosikal
	Cluster 9 (105)	**Cluster 10 (839)**	**Cluster 11 (383)**	
Kontijen	Kuala Lumpur	Selangor	Selangor	
Daerah	Dang Wangi	Klang Selatan	Klang Selatan	
Balai	Jln Bandar	Klang	Klang	
Bulan	Julai	Mac	November	
Jantina	Lelaki	Lelaki	Lelaki	
Kump Bangsa	W.Asing	W.Asing	Melayu	
Kump Umur	Adult	Adult	Teens	
Kesalahan	Ragut	Pecah Rumah Mal	Curi Motosikal	

4.3 Serial Commercial Crime using Supervised Clustering

Apart from clustering approach, we also conducted classification train and test using K-Star, Random Forest, and Random Tree, J48,RepTree, Bayesian Net, IBK, and Decision Table by using contingent district, police office, month, gender, races, age as the attribute or features to predict or classify the types of offence as the target value. Here, we test on split-percentage from 95%-5% until 10%-90%. We observe that split percentage of 80:20 is sufficient to achieve optimal learning model.

Figure 7 depicts the initial accuracy of KStar Algorithm, one of the classification methods that have been applied. Details on the classification results could be referredto Appendix A.1.At this preliminarystage of research, the highest accuracy obtained currently is using K-Star lazy classifier. This classifier has had very good results handling missing values, smoothness problems and coping with mixed values. Both Random forest and Random tree achieve acceptable scores which are slightly below K-Star. Additionally, this analysis shows that those >70% of accuracy are prominent and distinctive crime pattern that can be associated and correlated with demographic or population factors. Therefore, further and strategic plans on combating crime are able to drive better and generic crime prevention activities such as smart crime free community partnership or to uphold better zero-crime policy in relation to demographic factors or attributes. The rest 30% cases in which less correlated may be independent in which the law authorities and officers able to handle them according to personalized combating plan.

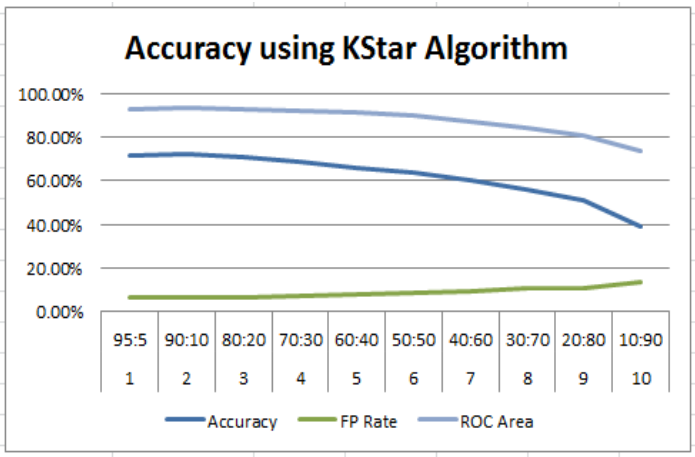

Fig. 7: Graph of K-Star Accuracy Algorithm

5. Conclusion

This study had paid an extra focus on pre-processing phase due to the importance of a clean and accurate data in pattern mining. Therigorous experimental results are presented. We can conclude that K*Star classification method achieve higher scores compared to J48, Decision table, and, Random forest and Random Tree. This is an iterative process to find the best solution. As conclusion, having accurate dataset will help in obtaining an output that is useful and meaningful. The results of this study will be further improved by adding the demographic and statistic information to the current crime dataset.

TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

Acknowledgement

The author acknowledges the support of NKRA Safe City Program, Royal Malaysia Police and Federal Town & Country Planning Department Peninsular Malaysia to enable this concept to be further research for this project. This work was supported by the Grant Code PP-FTSM-2019 and AP/2017/005/2.

References

[1] The Straits Times. 2018. Malaysian MP lodges police report on Altantuya murder, wants case reopened, Accessed at https://www.straitstimes.com/asia/se-asia/malaysian-mp-lodges-police-report-on-altantuya-murder-wants-case-reopened

[2] The Star. 2014. Penang cops on the hunt for fourth suspect behind grisly murders. Accessed at http://www.thestar.com.my/News/Nation/2014/12/14/Crime-Penang-Relau/

[3] The Rakyat Post. Missing French tourist found dead in Tioman. 2014. Accessed at http://www.therakyatpost.com/news/2014/09/04/missing-french-tourist-found-dead-tioman/

[4] Pemandu. 2012. Reducing Crime Overview. Accessed at http://www.pemandu.gov.my/gtp/Reducing_Crime-@-Reducing_Crime_Overview.aspxB.

[5] Agrawal, R., Imielinski, T. & Swami, A. 1993. Mining association` rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD'93). 207-216.

[6] Zaki, M. J. & Ho, C. T. 2000. Large Scale Parallel Data Mining. Springer.

[7] Han, J., Cheng, H., Xin, D. &Yan, X. 2007. Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery, Vol 15, 55-86.

[8] Bruce, C., Santos, R. B., Rodriguez, E. & Gwinn, S. 2011. Crime Pattern Definition for Tactical Analysis. Accessed at http://www.iaca.net/Publications/Whitepapers/iacawp_2011_01_crime_patterns.pdf.

[9] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. & Chau, M. 2004. Crime data mining: a general framework and some examples. Computer , Vol.37(4), 50-56.

[10] Isafiade, O.E. & Bagula, A.B. 2013. CitiSafe: Adaptive Spatial Pattern Knowledge Using Fp-Growth Algorithm for Crime Situation Recognition. Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC). 551-556.

[11] Xu, Y., Mingyang, L., Ningning, A. &Xinchao, Z. 2012. Criminal Detection Based on Social Network Analysis. Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on. 201-204.

[12] Snook, B., Zito, M., Bennell, C. & Taylor, P. J. 2005. On the Complexity and Accuracy of Geographic Profiling Strategies. Journal of Quantitative Criminology, Vol 21 (1).1-26.

[13] Ding, L., Steil, Dixon, D. B., Parrish, A. & Brown, D. A relation context oriented approach to identify strong ties in social networks, Knowledge-Based Systems, Vol 24(8). 1187-1195

[14] Amandus Jr., P. V. 2014. Background of Safe City monitoring Systems (SCMS), Federal Department of Town and Country Planning, Kuala Lumpur. Interview, 23 April 2014.

[15] Yu, C.-H., Ding, W., Chen, P., Morabito, M. Crime forecasting using spatio-temporal pattern with ensemble learning (2014) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8444 LNAI

[16] Domingos, P. 2012. A few useful things to know about machine learning. Communications of ACM 55(10):78-87.

[17] Mitchell, T. 2006.The discipline of machine learning (Technical Report CMU-ML-06-108). Carnegie Mellon University.

[18] Mohammad OsiurRahman, AiniHussain, Edgar Scavino, Hassan Basri, M.A. Hannan. 2011, Intelligent computer vision system for segregating recyclable waste papers, Expert Systems with Applications, 38(8):10398-10407.

[19] FarshidPirahansiahSitiNorul Huda Sheikh Abdullah ShahnorbanunSahran, 2013. Simultaneous Localization And Mapping Trends And Humanoid Robot Linkages Asia-Pacific Journal of Information Technology and Multimedia JurnalTeknologiMaklumatdan Multimedia Asia-Pasifik 2(2): 27 – 38 , December 2013.

[20] Siswantoro, J., Prabuwono, A.S., Abdullah, A. and Bahari, I., 2017. Hybrid Neural Network and Linear Model for Natural Produce Recognition Using Computer Vision. *Journal ICT Research and Applications*, *11*(2), pp.184-198.

[21] S. N. H. S Abdullah, Farah AqilahBohani, Zakree Ahmad Nazri, Yasmin Jeffry, Mohammed Ariff Abdullah, MdNawawiJunoh, ZainalAbidinKasim, 2018, Amenities Surrounding Commercial Serial Crime Prediction At Greater Valley And Kuala Lumpur Using K-Means Clustering/PengecamanKemudahanAwamSekitarLokasiJenayah KormesialBersiri Di LembahKlang Dan Kuala Lumpur MenggunakanKaedahGugusan K-Means. JurnalTeknologi (Sciences & Engineering) 80:4 (2018) 43–53S

[22] Suzilah Ismail, NurulhudaRamli, Short-term Crime Forecasting in Kedah, Procedia - Social and Behavioral

Sciences, Volume 91, 10 October 2013, Pages 654-660, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2013.08.466.

[23] Kumar, M.V., Chandrasekar, C.: Spatial clustering simulation on analysis of spa- tialtemporal crime hotspot for predicting crime activities. International Journal of Computer Science and Information Technologies 2(6), 2864–2867 (2011).

[24] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.

[25] Turner, G., Brantingham, D.J. and Mohler, D.G., 2014. Predictive Policing in Action in Atlanta, Georgia. The Police Chief.

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

Appendix A

Table A.1: Result of Classification using KStar Lazy Classifier Algorithm

Model	Data Allocation	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	95:5	76.43%	0.764	0.06	0.771	0.764	0.762	0.954
2	90:10	72.70%	0.727	0.064	0.739	0.727	0.723	0.949
3	80:20	71.07%	0.711	0.068	0.721	0.711	0.71	0.938
4	70:30	68.64%	0.686	0.069	0.69	0.686	0.686	0.926
5	60:40	66.08%	0.661	0.077	0.663	0.661	0.659	0.906
6	50:50	60.58%	0.606	0.091	0.607	0.606	0.602	0.881
7	40:60	57.42%	0.574	0.097	0.575	0.574	0.569	0.859
8	30:70	52.43%	0.524	0.107	0.524	0.524	0.519	0.826
9	20:80	46.58%	0.466	0.118	0.467	0.466	0.46	0.785
10	10:90	37.54%	0.375	0.13	0.372	0.375	0.371	0.722

Table A.2: Result of Classification using J48 Algorithm

Model	Data Allocation	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Number of Leaves	Size of Tree
1	90:10	58.28%	0.583	0.074	0.584	0.583	0.578	0.846	1820	2346
2	80:20	57.23%	0.572	0.079	0.571	0.572	0.568	0.843		
3	70:30	55.52%	0.555	0.084	0.555	0.555	0.55	0.831		
4	60:40	51.37%	0.514	0.089	0.511	0.514	0.51	0.811		
5	50:50	48.74%	0.487	0.099	0.484	0.487	0.483	0.789		
6	40:60	44.85%	0.449	0.109	0.44	0.449	0.44	0.76		
7	30:70	41.94%	0.419	0.111	0.416	0.419	0.414	0.736		
8	20:80	37.49%	0.375	0.128	0.364	0.375	0.364	0.687		
9	10:90	29.72%	0.297	0.119	0.302	0.297	0.296	0.636		
10	66:34	53.57%	0.536	0.088	0.533	0.536	0.53	0.822		

Table A.3: Result of Classification using Decision Table Algorithm

Model	Data Allocation	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Number of Rules
1	90:10	33.61%	0.434	0.119	0.481	0.434	0.429	0.775	1125
2	80:20	29.18%	0.446	0.116	0.477	0.446	0.443	0.765	
3	70:30	30.41%	0.411	0.134	0.443	0.411	0.402	0.761	
4	60:40	36.96%	0.413	0.131	0.436	0.413	0.404	0.751	
5	50:50	38.62%	0.386	0.137	0.399	0.386	0.373	0.739	
6	40:60	41.28%	0.37	0.142	0.381	0.37	0.354	0.716	
7	30:70	41.14%	0.304	0.169	0.373	0.304	0.287	0.621	
8	20:80	44.64%	0.292	0.176	0.325	0.292	0.268	0.619	
9	10:90	43.36%	0.336	0.174	0.284	0.336	0.273	0.645	
10	66:34	40.26%	0.403	0.134	0.421	0.403	0.392	0.757	

Table A.4: Overall Result of all the Supervised Learning Classification Algorithms

Run	Bayes Net	J48	REP-Tree	Ran-dom Forest	Ran-dom Tree	JRip	One R	PART	Kstar	IBK	LWL	Stack-ing	Bag-ging	AdaboostM1	Att Selected Classi-fier	Filtered Classi-fier
1	38.70	39.37	24.53	NEM	NEM	38.96	9.76	NEM	36.42	36.42	36.26	24.53	NEM	32.13	24.53	38.29
2	59.68	54.08	47.89	62.76	62.57	40.55	63.97	53.48	**75.34**	75.06	39.02	25.71	52.18	35.30	48.49	54.08
3	42.33	58.66	47.95	74.07	72.99	45.66	35.94	57.67	**72.70**	74.26	35.02	24.40	54.78	32.76	44.26	51.45

November-December 2019
ISSN: 0193-4120 Page No. 4786 - 4799

4	37.94	37.27	37.50	38.35	38.07	35.46	33.27	39.50	42.23	36.80	36.45	24.53	40.04	32.60	40.26	37.69
5	40.17	38.99	38.42	39.63	35.84	39.47	33.27	NEM	37.97	37.94	32.26	24.53	38.51	32.82	37.88	38.32
6	51.43	57.14	51.11	66.03	66.98	48.89	47.94	55.87	68.25	66.03	46.98	27.30	52.06	34.60	47.62	57.14
7	41.91	58.85	48.74	**74.17**	**73.31**	45.85	35.78	58.98	**73.09**	74.39	32.76	24.40	NEM	32.76	51.06	52.08
8	44.14	61.74	52.46	**73.82**	**73.75**	47.16	36.42	61.58	**73.88**	73.98	34.00	25.71	NEM	34.00	52.65	53.83