ICIC Express Letters
Part B: Applications
Volume **10**, Number **8**, August **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **657–663**

A STUDY ON APPLICATION OF DATA MINING TECHNIQUES IN CRIME FORECASTING

Najla Al-Taleb[1] and Saqib Saeed[2]

[1]Department of Computer Science
[2]Department of Computer Information Systems
College of Computer Science and Information Technology
Imam Abdulrahman Bin Faisal University
P.O. Box 1982, Dammam 31441, Saudi Arabia
{ 2190500053; sbsaed }@iau.edu.sa

Received December 2018; accepted February 2019

Abstract. *Technology can support law enforcement agencies in crime prevention. Data mining techniques have huge potential in determining the crime patterns and this paper highlights this by conducting a study on crime dataset of Seattle Police Department. We used Support Vector Machine (SVM) and Naïve Bayes classifier algorithms for crime forecasting. The result of using 10-fold cross validation and percentage split methods showed that Naïve Bayes classifier algorithm performs slightly better than SVM classifier. This study provides insights that how data mining application can improve the efficiency and effectiveness of police department and such technologies should be used by law enforcement agencies to realize a peaceful society.*
Keywords: Data mining, Crime forecasting, Classification, SVM, Naïve Bayes

1. **Introduction.** Human behavior is not predictable; however, spatial and temporal patterns can help understand the patterns of crime. Technology can play a significant role in analyzing crime patterns to understand the circumstances that lead to a crime occurrence. There are different data mining techniques that can be used to analyze crime data such as classification, regression, and clustering. Data mining algorithms can predict crime type that is likely to occur at a specific time and location, and consequently, it will help prevent the crime [1,2].

There are many studies in the literature highlighting usage of technological applications in law enforcement [3,4]. Sathyadevan et al. used Naïve Bayes and Apriori algorithms to identify a crime pattern for a given place [5]. Similarly, Jain et al. proposed a novel method to investigate and analyze digital and physical crimes [6]. Furthermore, Ahmed et al. used Rapid Miner tool to implement Naïve Bayes classification algorithm on crime dataset to demonstrate the accuracy and efficiency of it [7]. Gupta et al. highlighted that there are peak and off peak times of crime occurrence across the year [8]. Mary concluded that multi class classifiers performance is better [9]. Shea highlighted crimes density growth in a college campus and the areas near a college campus using heat maps [10]. Kiani et al. introduced a new framework to analyze crimes by a combination of clustering and classification. The results showed that the accuracy of classification increased after parameters optimization and the classification error decreased [11]. Agarwal et al. applied K-means clustering algorithm on crime dataset of England and Wales from 1990 to 2011, and the results showed that the murder crimes were decreasing [12]. Furthermore, Tayal et al. proposed a model to detect a crime and identify the criminals in Indian Cities. The results of the evaluation of the performance of k-means using the crime features showed that the first cluster highlighted an accuracy of 93.62% and for the second cluster an

DOI: 10.24507/icicelb.10.08.657

accuracy of 93.99% was achieved [13]. Mcclendon and Meghanathan showed that linear regression was more accurate than the other techniques in comparing and analyzing crime patterns [14]. Despite these studies, there is no study which has used Support Vector Machine (SVM), so in this study we use SVM. Secondly, we use Naïve Bayes classifier, which is used heavily by different researchers to understand their performance. Our primary research question is how data mining techniques can help forecast crime patterns in Seattle city. The study will predict the crime categories that will likely occur in specific time and location in Seattle city.

The rest of this paper is structured as follows. In Section 2, methodology of this research is presented followed by findings in Section 3. Section 4 presents a discussion on the results followed by a conclusion in Section 5.

2. **Methodology.** The crime dataset that has been used in this study is a secondary data set that is available at Data.gov website, which contains the records of reported offense to the Seattle Police Department. The dataset contained 494,068 instances, and each instance represents an offense reported either by a police officer or a person from the society [15]. The dataset contained eleven attributes, namely Report Number, Occurred Date, Occurred Time, Reported Date, Reported Time, Crime Subcategory, Primary Offence, Precinct, Sector, Beat and Neighborhood. While cleaning the data, we have to omit the data for the last 2 weeks, because the quality control process lags behind for two weeks, as mentioned by the website [15]. Since last entry at the time of download was on September 28, 2018, the records from 15th-28th September 2018 are omitted. Furthermore, there were 262 instances where some values were missing so these records were omitted. Moreover, the dataset contained a lot of inconsistent data, such as the date, where sometimes it is given as text type, which required conversion. The number of instances in dataset before data cleaning was 494,068, and after the data cleaning process, it became 492,081 instances.

In data reduction process, we reduced the attributes that did not affect the results of the analysis but improved efficiency [2]. Dimensionality reduction has been used in this study and we only used Crime Subcategory, Occurred Date, Occurred Time, and Neighborhood attributes. Furthermore, we only selected the crime data of only last 1 year, so data older than 15th September 2017 was omitted, resulting in 52,545 instances. In the transformation step, two new numerical attributes, Year and Month, have been constructed by decomposing Occurred Date attribute, because the analysis and prediction need month information. In the discretization step, Occurred Time attribute, has been discretized into six hours intervals to improve the accuracy of the model. Moreover, the Crime Subcategory attribute values have been categorized into six distinct values to improve the accuracy of the model. Table 1 presents data set description after preprocessing.

The correlation coefficient has been calculated to understand the relationship between the attributes of the dataset. Table 2 shows the correlation coefficients between the class attribute and each attribute in the dataset.

Table 1. Dataset description after preprocessing

Attribute	**Data Description**	**Type**
Occurred Time	Offense occurrence time	Nominal
Crime Subcategory	Crime Subcategory	Nominal
Neighborhood	Where the crime occurred	Nominal
Month	Offense occurrence month	Numeric
Year	Offense occurrence year	Numeric

TABLE 2. Correlation coefficients between class attribute and each attribute

Attributes Pair	Correlation Coefficients
(Occurred Time, Crime Subcategory)	0.04669
(Neighborhood, Crime Subcategory)	0.01968
(Year, Crime Subcategory)	0.00585
(Month, Crime Subcategory)	0.00266

TABLE 3. Statistical analysis of numeric attributes in the dataset after preprocessing

Attribute	Mean	Standard Deviation	Maximum	Minimum
Month	6.62	3.385	12	1
Year	2017.681	0.466	2018	2017

Table 3 provides a statistical analysis of numeric attributes in the dataset after preprocessing. The statistical analysis includes the mean, minimum, maximum, and standard deviation.

We used Weka toolkit to implement SVM and Naïve Bayes classification algorithms [16]. We selected 10-fold cross validation and percentage splits methods to be used for evaluation and comparison of accuracy and recall (or sensitivity) of the classification models.

3. **Findings.** In this section, the results of analysis of Seattle city crime dataset are presented. Figure 1 presents a statistical analysis of the crime types that occurred in the last twelve months in Seattle city. Each bar represents the number of crimes that occurred for each crime category such as theft, car offense, sex offense, drug and alcohol, assault and other crime types. Whereas, in Figure 2, a statistical analysis of the crime occurrence time in the last twelve months in Seattle city is presented. In Figure 3, a statistical analysis of the crime occurrence month in Seattle city is presented. The figure shows the number of crime incidents that occurred in each month for the last twelve months.

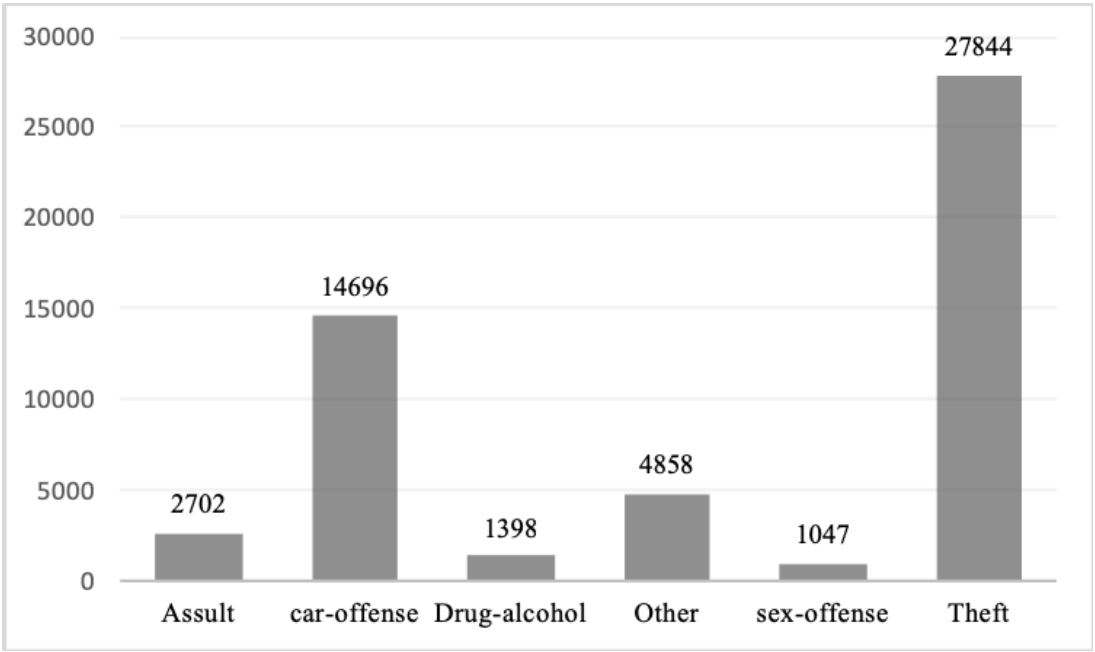

FIGURE 1. Number of crimes that occurred in Seattle city per category

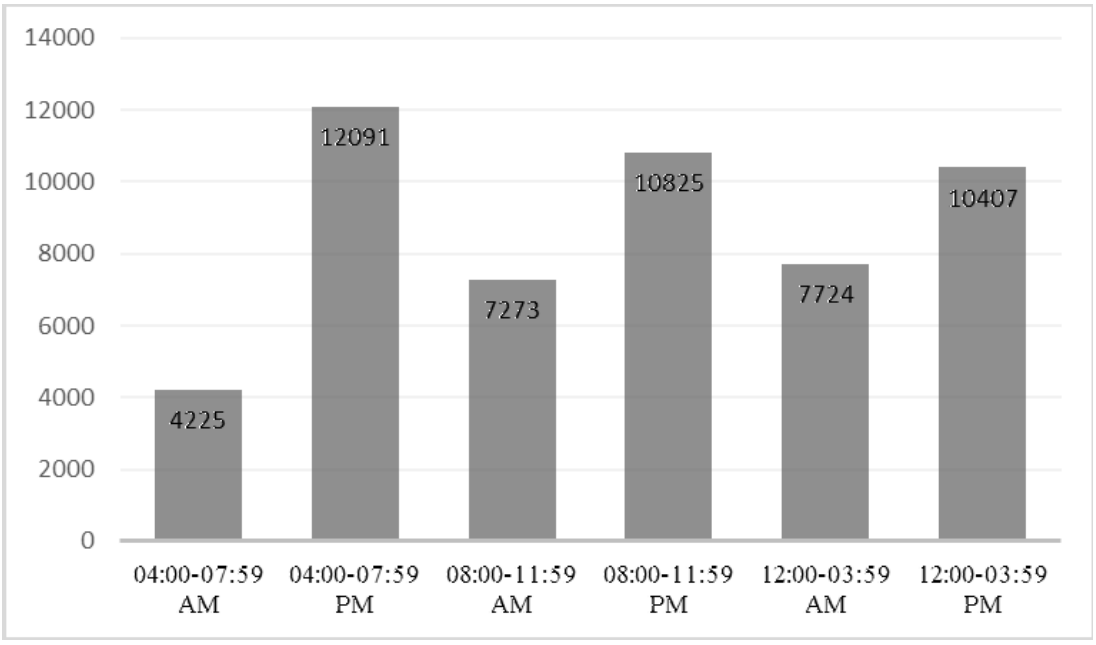

FIGURE 2. Number of crimes that occurred in Seattle city per time period

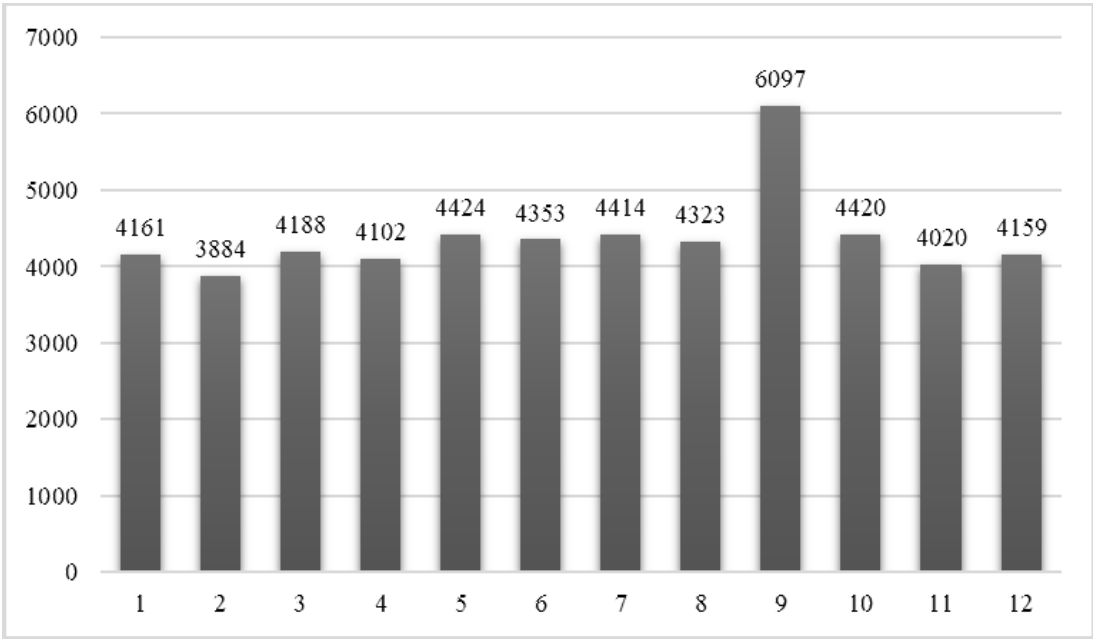

FIGURE 3. Number of crimes that occurred in Seattle city per month

Figure 4 presents a relationship between type of crime and time zone that occurred in different neighborhood.

This paper used 10-fold cross validation method to evaluate the accuracy of SVM and Naïve Bayes classification algorithms. The result showed that SVM correctly classified the crime category with an accuracy of 52.99%, while Naïve Bayes correctly classified the crime category with an accuracy of 54.06%. The recall value represents the weighted average value of the SVM and Naïve Bayes classification algorithms recall. The results of the recall, which represent the percentage of how many tuples have been correctly classified to the crime category that belongs to by SVM and Naïve Bayes classification algorithms, are presented in Table 4.

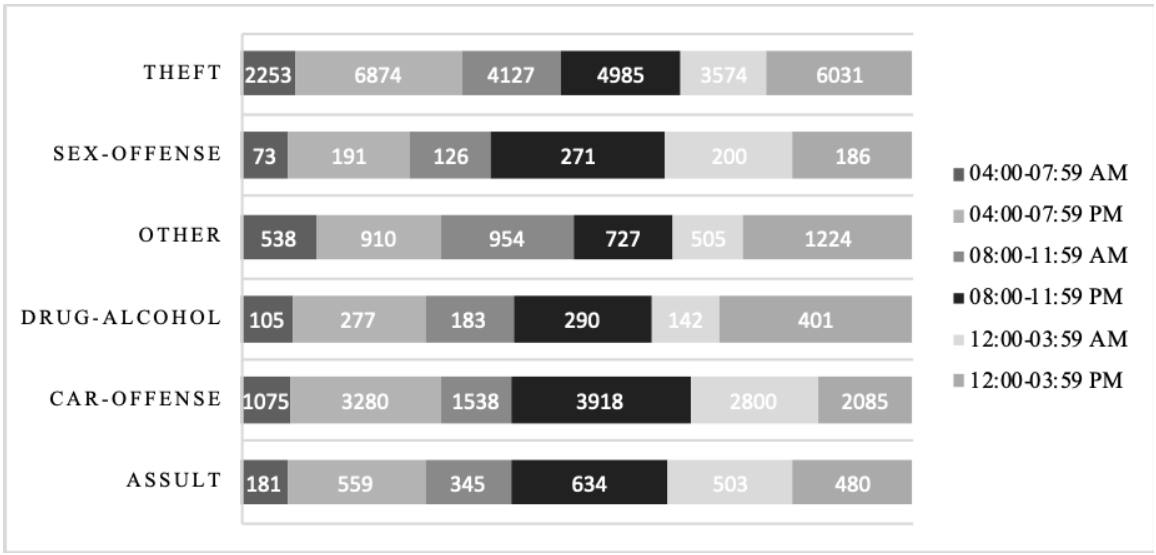

FIGURE 4. Number of crimes in different time zones and crime categories

TABLE 4. Classification evaluation using 10-fold cross validation

Algorithm	Correctly Classified	Incorrectly Classified	Recall
SVM	52.99%	47.01%	0.53
Naïve Bayes	54.06%	45.94%	0.54

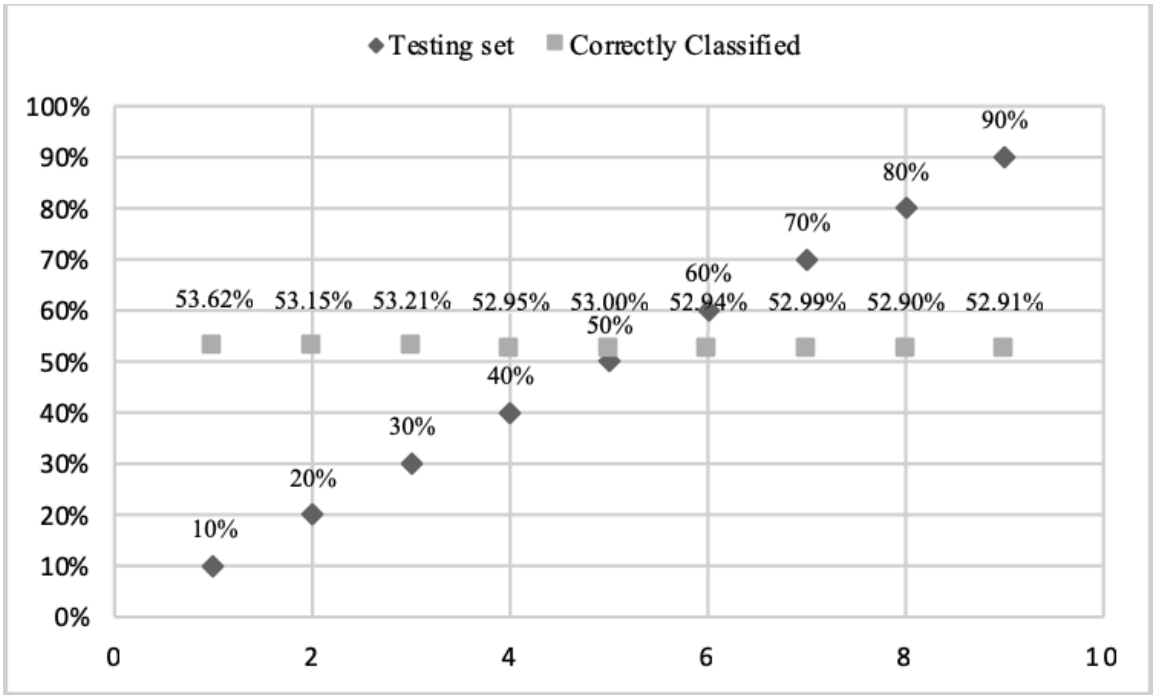

FIGURE 5. Percentage split for SVM

Percentage splits method has also been used in this paper to evaluate the accuracy of SVM and Naïve Bayes classification algorithms. Figure 5 represents the effect of increasing the test set percentage on the accuracy of the classification of SVM algorithm. Figure 6 represents the effect of increasing the test set percentage on the accuracy of the classification of Naïve Bayes algorithm.

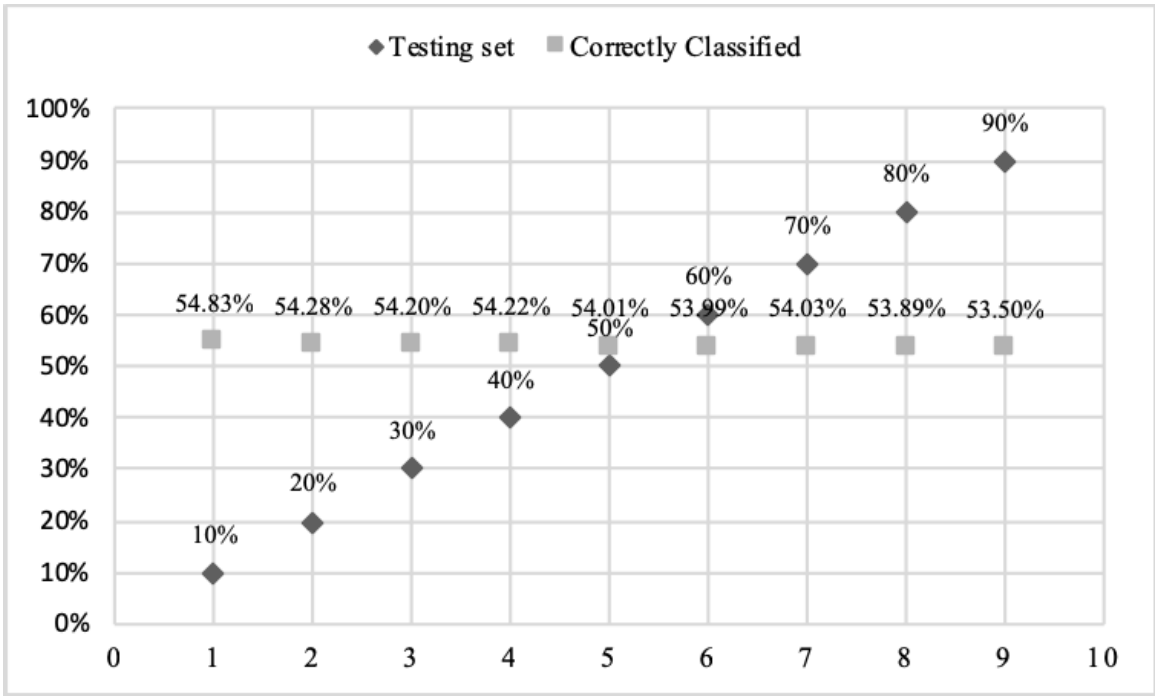

FIGURE 6. Percentage split for Naïve Bayes

4. **Discussion.** In this study, the crime record for the last twelve months in Seattle city has been analyzed and used for the crime category prediction. From the data analysis, we found that the crimes labeled under theft category were the most occurring in Seattle city while the crimes that were labeled under sex offense category were the least occurring. Also, the crimes that were categorized as car offense were second mostly occurring crimes. Regarding the time, we observed that the crimes mostly occurred in the time period from 4 PM to 7:59 PM in Seattle city while the time period from 4 AM to 7:59 AM had least crime incidents. Also, we found that September was the month that had the highest number of crime incidents while the lowest number of crime incidents were in February in the last twelve months. Moreover, from the analysis, we found that the Downtown commercial neighborhood has the highest number of crime occurrence. The crimes that were labeled under theft category were the most occurring during all period of times in Seattle city. Drug and alcohol crimes mostly occurred during 12 PM to 3:59 PM. Also, in the Bell Town neighborhood, the crimes that were labeled under car offense category were highest in number.

In terms of prediction, when we used the 10-fold cross validation method we observed that both SVM and Naïve Bayes classification algorithms showed a lower classification accuracy, but Naïve Bayes classifier performed slightly better than SVM classifier. Moreover, the recall values showed that the number of tuples that were correctly classified into their category were higher while using Naïve Bayes classification algorithm as compared to SVM classification algorithm.

Percentage splits found that SVM performed best with an accuracy of 53.62% when the training set percentage is 90% and the testing set is 10%. Also, Naïve Bayes performed best when the training set percentage is 90% and testing set is 10% with an accuracy of 54.83%. Moreover, we observed that the accuracy of SVM in percentage splits decreased each time we increased the percentage of the testing set and decreased the training set. Similarly, the accuracy of Naïve Bayes classification algorithm in percentage splits decreased each time we increased the percentage of the testing set and decreased the training set.

5. **Conclusion.** Crime is unpredictable because it is dependent on human behavior. Thus, many studies were conducted to understand the crime circumstances. In this paper, we have implemented Naïve Bayes and SVM classification algorithms on crime dataset for Seattle city to predict the crime category that is likely to occur. Moreover, we have analyzed Seattle city crime dataset to find the crime patterns and to understand the type of data and patterns in the dataset. We observed that Naïve Bayes classifier performed slightly better than SVM classifier. Finally, we found that data mining techniques can be helpful in analyzing crime data and crime forecasting, even though SVM showed a low accuracy of 53.62% as compared to 54.83% shown by Naïve Bayes. These performance values can be enhanced by using different optimization methods. For future work, we intend to take into consideration of other attributes that describe the type of crime location that can help predict the crime category rather than using the location attribute. Furthermore, we intend to find a way to improve the SVM classification to increase the accuracy of the classification and prediction of a crime category.

REFERENCES

[1] A. Agarwal, D. Chougule, A. Agrawal and D. Chimote, Application for analysis and prediction of crime data using data mining, *Int. J. Adv. Comput. Eng. Netw.*, vol.4, no.5, pp.9-12, 2016.

[2] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishing, 2012.

[3] C.-H. Yu, M. W. Ward, M. Morabito and W. Ding, Crime forecasting using data mining techniques, *IEEE the 11th Int. Conf. Data Min. Work.*, pp.779-786, 2011.

[4] C. Detotto and E. Otranto, Does crime affect economic growth?, *Kyklos*, vol.63, no.3, pp.330-345, 2010.

[5] S. Sathyadevan, M. S. Devan and S. S. Gangadharan, Crime analysis and prediction using data mining, *The 1st Int. Conf. Networks Soft Comput.*, pp.406-412, 2014.

[6] N. Jain, P. Sharma, R. Anchan, A. Bhosale, P. Anchan and D. Kalbande, Computerized forensic approach using data mining techniques, *Proc. of ACM Symp. Women Res.*, pp.55-60, 2016.

[7] W. Ahmed, Dr. S. S. Biswas and T. Nafis, Performance analysis of Naïve Bayes algorithm on crime data using rapid miner, *Int. J. Adv. Res. Comput. Sci.*, vol.8, no.5, pp.683-687, 2017.

[8] A. Gupta, A. Mohammad, A. Syed and M. N. Halgamuge, A comparative study of classification algorithms using data mining: Crime and accidents in Denver City the USA, *Int. J. Adv. Comput. Sci. Appl.*, vol.7, no.7, pp.374-381, 2016.

[9] R. R. Mary, Performance analysis of different classifiers to build a classification model and to improve the vigilance skills in crime detection using data mining techniques, *Int. J. Adv. Res. Comput. Sci.*, vol.3, no.7, pp.314-317, 2012.

[10] A. P. Shea, A visual system for mining crime mining across college campuses, *Proc. of 2017 ACM Int. Conf. Manag. Data – SIGMOD SRC'17*, pp.1-3, 2017.

[11] R. Kiani, S. Mahdavi and A. Keshavarzi, Analysis and prediction of crimes by clustering and classification, *Int. J. Adv. Res. Artif. Intell.*, vol.4, no.8, pp.11-17, 2015.

[12] J. Agarwal, R. Nagpal and R. Sehgal, Crime analysis using k-means clustering, *Int. J. Comput. Appl.*, vol.83, no.4, pp.1-4, 2013.

[13] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta and N. Tyagi, Crime detection and criminal identification in India using data mining techniques, *AI Soc.*, vol.30, no.1, pp.117-127, 2014.

[14] L. Mcclendon and N. Meghanathan, Using machine learning algorithms to analyze crime data, *Mach. Learn. Appl. An Int. J.*, vol.2, no.1, pp.1-12, 2015.

[15] *Crime Data*, data.seattle.gov, https://catalog.data.gov/dataset/crime-data-76bd0, 2018.

[16] *Weka*, https://www.cs.waikato.ac.nz/ml/weka/documentation.html.