KNN CLASSIFIER AND NAÏVE BAYSE CLASSIFIER FOR CRIME PREDICTION IN SAN FRANCISCO CONTEXT

Noora Abdulrahman and Wala Abedalkhader

Department of Engineering Systems and Management, Masdar Institute of Science and Technology, Abu Dhabi, the United Arab Emirates

ABSTRACT

In this paper we propose an approach for crime prediction and classification using data mining for San Francisco. The approach is comparing two types of classifications: the K-NN classifier and the Naïve Bayes classifier. In the K-NN classifier, two different techniques were performed uniform and inverse. While in the Naïve Bayes, Gaussian, Bernoulli, and Multinomial techniques were tested. Validation and cross validation were used to test the result of each technique. The experimental results show that we can obtain a higher classification accuracy by using multinomial Naïve Bayes using cross validation.

KEYWORDS

Classification, K-NN Classifier, Naïve Bayes, Data Mining, Gaussian, Bernoulli, Multinomial, Uniform, Inverse & Python

1. INTRODUCTION

The crime rate is increasing and it is very difficult to predict crimes. However, within the last few decades, and with the developing technology, data mining provided tools that enable crime prediction. Predicting the crime will not prevent it 100% from happening, but to some extent it will provide security to crime sensitive areas.

It might be very challenging to find a pattern in a crime and then to predict the type of crime. Crime prediction goes through different steps and analysis to be determined. The first step is data collection, where related data are collected from different sources. Data has many types and shapes including categorical and numerical. Then based on the type of data the method of data mining that should be used is decided on. Some data mining methods are classification, clustering, and regression. Then comes the pattern identification phase where the trends and patterns in crime have to be identified using some algorithms. After that, the prediction, visualization, and evaluation phases are used to study the results and the behaviour.

2. LITERATURE REVIEW

Based on studies, there are several crime data mining methods, including regression, classification, and clustering. Each method is used in different scenarios based on the type of available data, training set and testing sets. This section summarizes some of the successful methods in crime prediction that other scholars developed.

Tayala proposed an approach for crime detection and criminal identification in India using data mining techniques [1]. Their approach is divided into six modules including: data extraction, data pre-processing, clustering, Google map representation, classification, and WEKA implementation. They used k-means clustering for analysing crime detection that generates two

DOI : 10.5121/ijdms.2017.9401

crime clusters based on similar crime attributes. Then KNN classification was used for criminal identification and prediction, while WEKA was used for crime verification of their results.

Nath also focused on the clustering algorithm to detect crimes' patterns and speed up the process of crime solving, specially the k-means clustering [2]. A semi-supervised learning technique was used for knowledge discovery from crime records to help increase the predictive accuracy. They also developed a weighting scheme for attributes to deal with limitations of clustering tools and techniques.

Wang et al. introduced a new crime hotspot mapping tool called Hotspot Optimization Tool (HOT), which is an application of spatial data miming to the field of hotspot mapping that can capture the differences between two classes in a spatial dataset [3]. The study was done to a north-eastern city in the United States and HOT was able to accurately map the crime hotspots.

Oatley and Ewart developed a project to assist the West Midlands Police in the UK with high volume crime, burglary from dwelling houses [4]. They created a software that utilizes mapping and visualization tools that are capable of a range of sophisticated predictions. The statistical methods employed the data-mining technology of neural network to determine the causality in this domain. The predictions on the likelihood of burglary is calculated by combining all the evidence into a Bayesian belief network that is embedded in the developed software system.

Sun et al. compared three typical classification algorithms, including C4.5 algorithm, Naive Bayesian algorithm and KNN algorithm in order to obtain high accuracy [5]. The results show that a better classification accuracy can be obtained by combining KNN classification algorithm and GBWKNN missing data filling algorithm that is based on grey relational analysis (GRA) theory.

Grubesic studied the use of cluster analysis and GIS for detecting hot spots [6]. The study outlines several problems of optimization based on cluster analysis for crime hot spot detection. He suggests the adaption of an existing statistical and geometric technique, including the statistical test of cluster significance and geometric properties of cluster partitions.

Tahani et al. focuses on finding spatial and temporal criminal hotspots using statistical analysis in Denver, CO and Los Angeles, CA [7]. Then, Apriori algorithm is conducted to produce frequent patterns for criminal hotspots. In addition, they used Decision Tree classifier and Naïve Bayesian classifier in order to predict potential crime types.

Leong and Sung summarizes the spatio-temporal pattern analysis approaches for crime analysis [8]. They discuss the knowledge that could be obtained from these patterns and what approaches of various data mining techniques could be used.

Chang et al. developed a new spatio-temporal data analysis approach to discover abnormal spatio-temporal clustering patterns [9]. They proposed a quantitative evaluation framework and used it to compare against a widely used space–time scan statistic-based method. Their approach is based on a robust clustering engine to detect abnormal areas with irregular shapes more accurately than the space– time scan statistic-based method.

Sathyadevan and Gangadharan created a system that can predict regions that have high probability for crime occurrence using naïve Bayes algorithm that gave them 90% accuracy [10]. They tested the accuracy of classification and prediction based on different test sets. Their system takes factors/attributes of a place and Apriori algorithm gives the frequent patterns of that place. The pattern is used for building a model for the decision tree.

Sharma presents a paper that employs decision tree-based classification approach to detect criminal activities [11]. In this study he found that an advanced decision tree classifier and feature selection method can provide good classification result for suspicious e-mail detection. The results from his experiment show that Advanced ID3 algorithm has better classification accuracy compared with the traditional ID3 Algorithm.

Malathi and Baboo focus on developing a crime analysis tool using different data mining techniques, including MV algorithm and Apriori algorithm [12]. The tool follows four main steps: data cleaning, clustering, classification, and outlier detection. The outlier detection step is mainly used to identify crimes that might happen in the future. The results show that the tool is effective in terms of analysis speed, identifying common crime patterns and future prediction.

Yumuna and Bhuvaneswari proposed a study to analyze and predict crimes using clustering techniques including K-means and DBScan algorithm [13]. They concluded that using clustering helps to overcome the problem of large amounts of data and stem from various formats.

Wang et al. propose a pattern detection algorithm called Series Finder [14]. This algorithm grows a pattern of discovered crimes from within a database by starting from a seed of few crimes. It is able to pinpoint patterns more accurately than other similar methods.

Iqbal et al. conduct a study that compares two different classification algorithms, Naïve Bayesian and Decision Tree to predict the crime category for different states in the USA. The results show that Decision Tree algorithm performed better than Naïve Bayesian algorithm and achieved 83.9519% Accuracy in predicting the crime category for different states of the USA [15].

Oatley et al. developed a decision support system based on data mining algorithms to match and predict crimes [16]. The crime matching techniques that they used are case-based reasoning, logic programming and ontologies, and naïve Bayes augmented with spatio-temporal features. Crime prediction techniques are survival analysis and Bayesian networks. However, the results of the Bayesian network have never been validated because it consists of many arbitrary decisions and was intended as a prototype.

Yu et al. discuss crime forecasting using different classification techniques, SVM, J48, Neural, and INN [17]. They calculated the accuracy and F1 for each method. The results show that as the data SVM is better for big data sets, while Naïve Bayes is better for smaller data sets.

Chandra and Gupta proposed a novel distance-based semi-supervised clustering algorithm functional link neural network (FLNN) [18]. The motivation of their work is to overcome the disadvantages of the pair-wise constrains, which is the base for most of the semi-supervised clustering techniques. Mainly, it fails to address the problem of dealing with attributes having different weights. In most of the real-life applications, all attributes do not have equal importance and hence the same weights cannot be assigned for each attribute.

Chung et al. utilized four identity fields: name, address, date-of-birth, and social security number and compared each corresponding field for a pair of criminal identity records to analyze and predict Tucson Police Department (TPD) data [19] using Social Network Analysis (SNA). Then, an overall verification between the true and predicted data was computed by calculating the Euclidean Distance of disagreement measures over all attribute fields.

Chakravorty et al. had to deal with structured variables that are clearly pre-identified and usually collected by a single entity, as well as unstructured variables that are collected from different resources with different structure data sets [20]. NLP technique (Natural language processing, NLP is a set of techniques for using computers to detect in human language, the kinds of things that humans detect automatically), has been applied to the data collected to identify the places where murder has occurred in these two years. While for structured data, a k-means cluster analysis has been applied to the crime data.

3. PROBLEM DEFINITION

Data mining is used in this research to predict the category of crime that occurred from a dataset of nearly 12 years of crime report from across all San Francisco. This research is focused on the Kaggle data set from the “San Francisco Crime Classification” competition.

The model is designed using 8 features: dates, category, description, day-of-week, police department district, resolution, address, and X & Y coordinates, each set contains in excess of 800,000 reports. Given the time and location we should be able to predict the category of crime.

We will compare and contrast two different methods of data mining classification, K-NN classifier and Naïve Bayes classifier. For each method, codes will be developed and applied to the real crime data from the year 2003 to 2015. For each of the methods, different techniques were applied to compare which will give better results and will be more accurate.

4. DATA

The used dataset contains incidents derived from the SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... Belong to test set, week 2,4,6,8 belong to the training set.

A total of (878050) crimes are provided in the training data and the test data for which the categories will be predicted are of the same number. The data fields are:

- Dates - timestamp of the crime incident
- Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.
- Descript - detailed description of the crime incident (only in train.csv)
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District
- Resolution - how the crime incident was resolved (only in train.csv)
- Address - the approximate street address of the crime incident
- X - longitude
- Y – latitude

5. METHODOLOGY

In this research, Python was used to explore training data, make regression analysis and predict categories for test data, in order to get the best correlation between the features (Date, Pd-District, Address, Day of Week, Description, Resolution, X and Y) and the target value (Category of Crime). All nominal values were converted into binary values by converting the values of the attributes into separate new attributes and give them values of either 0 or1.

Several trials of different Regression methods were used on the training data by splitting it into two sets; training and validation, both validation and cross validation were conducted, the method with the least Log loss was applied to predict the results for the test data.

Two main Algorithms were used in this research. The first algorithm is K-nearest neighbours, KNN is a supervised learning algorithm for either classification or regression, using two different distances-weighting functions. The first function is uniform, all points in each neighbourhood are weighted equally. The second function is inverse, weight points by the inverse of their distance. In this case, closer neighbours of a query point will have a greater influence than neighbours which are further away.

The second algorithm is Naive Bayes, which is a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features, using three functions. The first function is Bernoulli, which implements the naive Bayes training and classification algorithms for data that are distributed according to multivariate Bernoulli distributions, it is useful if your feature vectors are binary. The second function is multinomial, which implements the naive Bayes algorithm for multinomial distributed data, it is typically used for discrete counts. The third function is Gaussian, where the likelihood of the features is assumed to be Gaussian, instead of discrete counts, we have continuous features. In Python, the Scikit-learn library functions were used to conduct regression and classification.

KNN has some advantages and disadvantages compared to Naïve Bayes. An advantage is that KNN's decision boundary can take any form, Naïve Bayes can only have linear, elliptic, or parabolic decision boundaries. Also Naive Bayes is not good with correlated attributes, if the distinguishing characteristic of classification is not the marginal distributions but correlation, then NB won't be a good choice. Naive Bayes can also be misled by the absence of an attribute. One of the disadvantages is that KNN does not recognize the most important attributes, the distance is the only criteria used. Additionally, it is non-parametric, and thus not as interpretable as NB, KNN cannot provide any relationships between the distribution of attributes and classes. KNN does not handle the missing data properly, Naïve Bayes just excludes the attribute of missing data. In KNN, the value of K needs to be tuned and an optimal value needs to be assigned. Another disadvantage is that KNN is slower in processing during prediction, with large amounts of data the difference in speed is significant.

6. DISCUSSION AND RESULTS

Comparing the features in both the test and train files the "Description" and "Resolution" are missing in the test, therefore they were not used in the analysis, some pre analysis exploration of data was conducted. A few histograms of features values and features counts were generated for "Hour, Month, District, Day of week and Category" as shown below in order.

Looking into the hour and month counts histograms, we can see more variations occurring in the hour which is expected considering the larger number of possible values for the feature, however, to obtain more specific predictions in the regression analysis hours were taken into consideration.

Southern district has the highest number of crimes and Richmond has the least, variance is noticed among the district's counts of crimes.

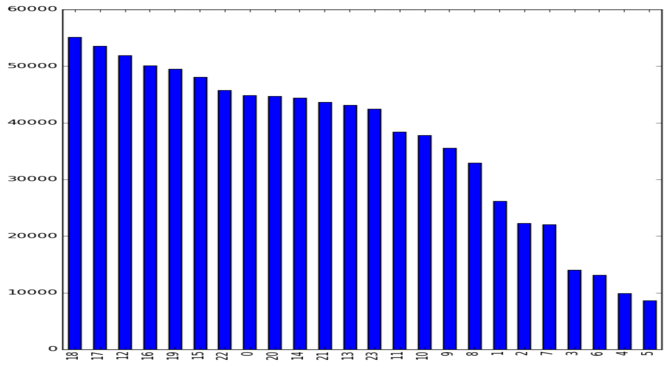


Figure 1. Hour (0-23) counts

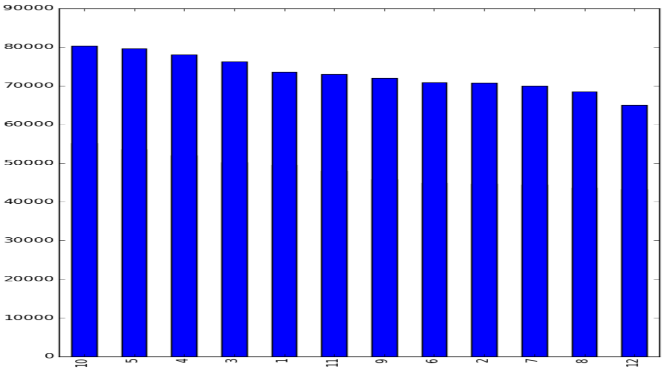


Figure 2. Month (1-12) counts

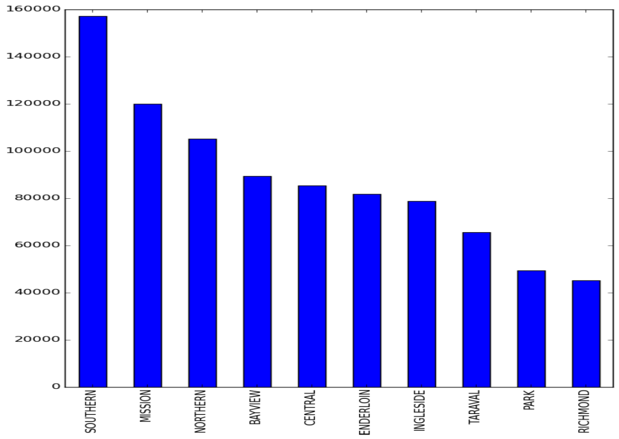


Figure 3. Pd-District counts

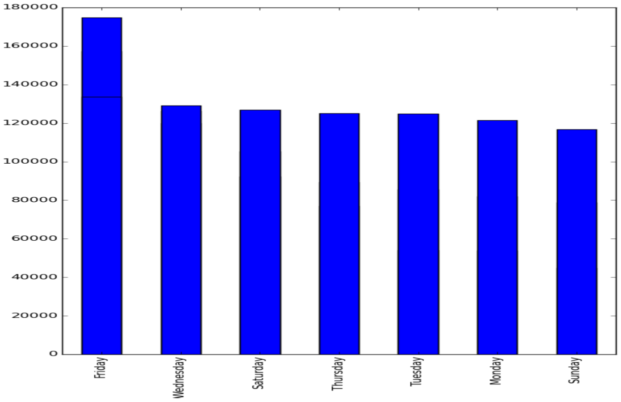


Figure 4. Days of the Week counts

Friday stands out as there is small variance between other week days. In Figure 5, we can see that there is a very large variance in categories counts Larceny/theft has the highest count and Trea has the lowest.

With regards to the Address (the values of a few hundreds) and X and Y coordinates (values of a few thousands), they had very large number of values and they both represented the location. Therefore, to obtain more specific results, Address was not used in the regression model, only X and Y coordinates were implemented.

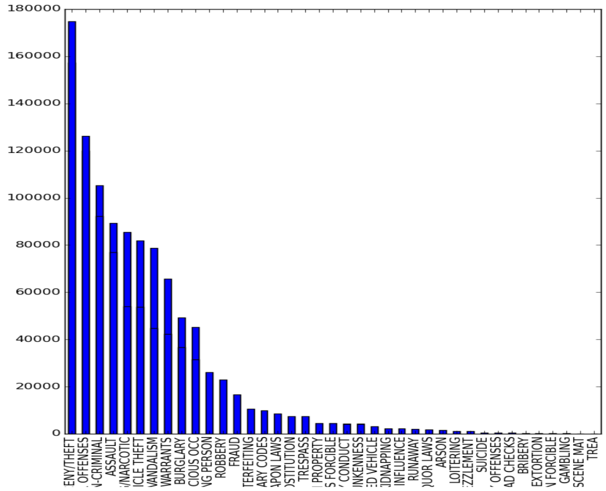
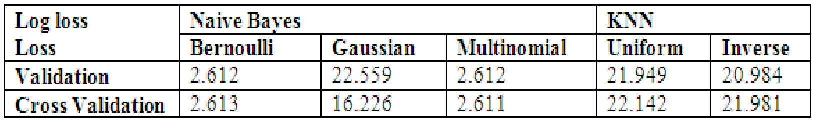


Figure 5. Categories counts

Using the five mentioned features (Day of week, District, Hour, X and Y) a regression model was built to represent the training data several times using different functions as described in the methodology.

A total of (10) trials were conducted using both Validation and Cross validation for the following: KNN- Uniform weights, KNN- distance Inverse Weights, Naïve Bayes – Bernoulli, Naïve Bayes – Multinomial, and Naïve Bayes – Gaussian. Table 1 shows the Log loss results of each run.

Table 1. Trials log loss results



Log loss	Naive Bayes			KNN	
Loss	Bernoulli	Gaussian	Multinomial	Uniform	Inverse
Validation	2.612	22.559	2.612	21.949	20.984
Cross Validation	2.613	16.226	2.611	22.142	21.981

Overall the cross validation gave better results than validation, mostly a small improvement was achieved (around 1%), which indicates that the split data was the representative of the whole set.

The least log loss values resulted fro naïve Bayes Bernoulli and Multinomial. This similarity is normal due to the binary nature of the data points in most attributes, only the X, Y were not binary and still they had very small variances which is why they did not have an impact on the final log loss very small differences were detected in favor of the multinomial.

Naïve Bayes Gaussian showed very bad results which indicates that the data does not follow a normal distribution and data was discrete not continuous.

KNN also gave very bad results, due to its inability to recognize the importance of each attribute also the value of k used might not be optimal and it was very slow during execution of regression and classification.

Being the Best method Cross validation using Multinomial regression function from training set was used to predict categories and classify the data points in the test set, after submitting the results to Kaggle the score representing the log loss was (2.61303).

7. CONCLUSION

In this paper, we applied two data mining techniques – KNN and Naïve Bayes, to identify patterns and then to predict the classification of a crime based on time and location. In the KNN approach, we applied Uniform and Inverse versions of the technique. For the Naïve Bayes approach, we used three different types; Bernoulli, Gaussian and Multinomial.

The results of the Bernoulli and Multinomial approaches are the best among applied techniques, when evaluating the results based on the log loss function.

In our work, we have directly applied each technique on the training data set before any pre-processing. The data set was not tested for outliers or entry errors. Moreover, since the data was taken in sequential time periods, auto regression and/or moving average behavior could be observed and thus should have been treated – especially if we applied a regression model.

In addition, other techniques could be applied such as Neural Networks (ANN, INN and FLNN) and Spatio-Temporal techniques as stated in the letter review where other researchers have applied similar techniques for the prediction of crime classification based on a given time and location.

REFERENCES

[1] 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 & Society, vol. 30, pp. 117-127, 2014.

[2] S. V. Nath, "<Crime Pattern Detection Using Data Mining.pdf>."

[3] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito, "Crime hotspot mapping using the crime related factors—a spatial data mining approach," Applied Intelligence, vol. 39, pp. 772-781, 2012.

[4] G. C. Oatley and B. W. Ewart, "Crimes analysis software: 'pins in maps', clustering and Bayes net prediction," Expert Systems with Applications, vol. 25, pp. 569-588, 2003.

[5] C.-c. Sun, C.-l. Yao, X. Li, and K. Lee, "<Detecting Crime Types Using Classification Algorithms.pdf>."

[6] T. H. Grubesic, "<Detecting Hot Spots Using Cluster Analysis and GIS.pdf>."

[7] T. Almanie, R. Mirza, and E. Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots," International Journal of Data Mining & Knowledge Management Process, vol. 5, pp. 01-19, 2015.

[8] K. Leong and A. Sung, "<A review of spatio-temporal pattern analysis approaches on crime analysis.pdf>."

[9] W. Chang, D. Zeng, and H. Chen, "A stack-based prospective spatio-temporal data analysis approach," Decision Support Systems, vol. 45, pp. 697-713, 2008.

[10] S. Sathyadevan and S. Gangadharan., "<Crime Analysis and Prediction Using Data Mining.pdf>."

[11] M. Sharma, "Z-CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree," in Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on, 2014, pp. 1-6.

[12] A. Malathi and S. S. Baboo, "An enhanced algorithm to predict a future crime using data mining," 2011.

[13] S. Yamuna and N. S. Bhuvaneswari, "Datamining Techniques to Analyze and Predict Crimes," The International Journal of Engineering And Science, vol. 1, pp. 243-247, 2012.

[14] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Learning to detect patterns of crime," in Machine Learning and Knowledge Discovery in Databases, ed: Springer, 2013, pp. 515-530.

[15] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. S. Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian Journal of Science and Technology, vol. 6, pp. 4219-4225, 2013.

[16] G. Oatley, J. Zeleznikow, and B. Ewart, "Matching and predicting crimes," Applications and Innovations in Intelligent Systems XII, pp. 19-32, 2005.

[17] C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011, pp. 779-786.

[18] B. Chandra and M. Gupta, "A novel approach for distance-based semi-supervised clustering using functional link neural network," Soft Computing, vol. 17, pp. 369-379, 2013.

[19] H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, et al., "Crime data mining: an overview and case studies," in Proceedings of the 2003 annual national conference on Digital government research, 2003, pp. 1-5.

[20] S. Chakravorty, "Data mining techniques for analyzing murder related structured and unstructured data," American Journal Of Advanced Computing, vol. 2, 2015.