ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 6, Issue 4, April 2017

A NOVEL CLASSIFICATION MODEL FOR ANALYSIS OF A CRIME USING NAÏVE BYES AND KNN IN DATA MINING

SHIVRAJ SINGH DEOPA[1], ABHISHEK KUMAR [2], KUNEEK GUPTA[3]

Dr. SHASHI KANT SINGH[4]

Galgotias college of engineering & technology greater Noida

ABSTRACT

In this paper we projected a model for crime prediction using Naïve byes in data mining. The proposed model will be able to extract crime data by using KNN classification and clustering to categorize crime records in three categories such as property offence, all offence and violence offence. In this model we are developing three functions namely, first as Mining Specific Client's Crime Record, second as Find the users' Component attribute to evaluate all state in dataset and third as find the multi-users Software specify for the status which is resolved by the end user point. To obtain the best result and to achieve the most stable crime result we are modifying the decision threshold and we are calculating estimated error, absolute error and percentage error which shows that our proposed model is quite accurate.

Keywords: classification, data mining,dataset, Naïve bayes etc.

I. INTRODUCTION

Due to increasing the amount of data, a need to develop technologies to analyze data in different fields, such as business, medicine and education, has emerged [1]. Therefore, data mining methods have become the main tools to analyze data and to discover knowledge from them [2]. Here, data mining refers to an integration of multiple methods such as classification, clustering, evaluation, and data visualization [3]. One of these data which needs data mining techniques to discover and predict underlying patterns are crime data [4]. A high number of crimes in different countries have forced governments to use modern technologies and methods to control and to prevent crimes. Data mining techniques are able to identify patterns rapidly for detecting future criminal actions [5]. This is because manual interpretations of crime data are limited due to the size of data as well as the complexity among different crime attributes. Data mining methods accelerate crime analytics, provide better analysis and produce realtime solutions to save considerable resources and time [6]. Today, a high number of crimes are causing a lot of problems in many different countries. In fact, scientists are spending time studying crime and criminal behaviors in order to understand the characteristics of crime and to discover crime patterns. It is known that criminals follow repetitive behaviour patterns, so analyzing their behaviors can help to capture relations among events from past crimes [7]. In this research, crime research studies are integrated by data mining techniques to identify the patterns and to achieve more accurate results. To analyze crimes, there are several characteristics such as different races in a society, income groups, age groups, family structure (single, divorced, married), level of education, the locality where people live, number of police officers allocated to a locality, number of employed and unemployed people among others [8]. Dealing with crime data is very challenging as the size of crime data grows very fast, so it can cause storage and analysis problems. In particular, issues arise as to how to choose accurate techniques for analyzing data due to the inconsistency and inadequacy of these kinds of data. These issues motivate scientists to conduct research on these kinds of data to enhance crime data analysis. The objective of this evaluation is twofold. First, it determines whether the feature selection technique is

useful to infer better classification accuracy and performance. Second, it compares the different classifiers in terms of AUC for choosing more accurate algorithms to classify crime status in the United States of America for obtaining a deeper insight into crime. In this study, a real crime dataset is used for data mining from UCI Machine Learning Repository. Five different Classification Algorithms are used to classify dataset based on a binominal class, the crime status. Examined classifiers are Naïve Bayesian, Decision Tree (J48), Support Vector Machine (SVM), Neural Networks (MultilayerPerceptron) and k-Nearest Neighbor.

II. RELATED WORK

Much of the current work is focused in two major directions: (i) predicting surges and hotspots of crime, and (ii) understanding patterns of criminal behavior that could help in solving criminal investigations. Important contributions towards the former include [1] by Bogomolov et al, who try to predict whether any particular area in London will be a crime hotspot or not, using anonymized behavioural data from mobile networks as well as demographic data. In [2], Chung-Hsien Yu et al use classification techniques to classify neighbourhoods in a city as hotspots of residential burglary, using a variety of classification algorithms such as Support Vector Machines, Naive Bayes, and Neural Networks. (More work on the usefulness of Support Vector Machines for hotspot detection can be found in [3]). Toole et al demonstrated in [4], by analyzing crime records for the city of Philadelphia, that significant spatio-temporal correlations exist in crime data, and they were able to identify clusters of neighbourhoods whose crime rates were affected simultaneously by external forces. They also noted significant correlations in crime across weekly time scales. Towards the second objective of understanding patterns of criminal behavior, significant contributions have been made by Tong Wang et al in [5], in finding patterns in criminal activity and identifying individuals or groups of individuals who might have committed particular crimes. Their approach was to identify a common modus operandi across crimes, which could then be linked to groups or individuals who might commit the crime. For this, the authors proposed a new machine learning method called Series Finder, which was trained to recognize patterns in housebreak incidents in Cambridge, Massachusetts.

III. METHOD

There are numerous types of data mining approaches; some of the foremost data mining methods are known as naïve byes. In this work, we will use classification KNN methods in instruction to analyze crime pattern aim to decrease and prevention the crimes as much as possible.[7]

We propose our individual actions that may be useful for several domain areas.

Algorithm 1:Data mining. Mining Specific Client's Crime Record.

Input: Database D of transactions, Specific Product attributes.

Output: Sites (info) used by individual user

Method

1. Accept input_Atr (Specific Attribute type)

2. for (int i=0; i<= D.size; i++)

3. if (input_Atr = = D_Atr)

4. Extract information as info from database

5. Return info

6. end

Depending on the users' interest rate reducing or increasing through the time, they can change their location structure to attract more users from the aspect of precise prediction benefits. On the other hand, the statistician can view from the analysis perspective. We proposed the procedure especially for maintainers and developers as a result of Window Indiage over a precise period of time.[6]

Algorithm 2: Find the users' Component attribute to evaluate all state in dataset.

Input: Database D of transactions, Specific component.

Output: total transaction of Component attribute.

Steps

1. count = 0; temp[ ] =null; // initialization

2.while(component_state==All
component_state==Property
component_state==Violence)

3. if (temp[ ] = =null)

4. temp [ ] = D_Atr

5. count ++

6. else

7. while (temp [ ])

8. if (temp [ ]== D_Atr)

9. do nothing

10. else temp[ ] = = D_Atr

11. count ++

12. return count // total number of component attribute for class dependency

Algorithm 3: Find the multi-users Software specify for the status which is resolved by the end user point.

Input: Database D of transactions, status to count the total number of Indiage during a specific period

Output: total number of users which specify the class of the status.

Method

1. Accept status parameter

2. count = 0; temp [ ] = null;

3. while (D_Status== All &&

D_Status== Property || D_Status== Violence)

Data Set

In this research, we will reflect crime database as a training dataset used for crime record Prediction. The revealed database covers real data ideals from crime and criminal attributes. We will also consider 70 percent as training value of the proposed model and 30 percent for testing. The size of data set complete details of all 95 attributes and 3 touple can be developed from the UCI machine learning repository website.

The persistence of this training is to discover the applicability of data mining technique in the efforts of crime analyze and prevention. The data was collected from some Crime departments in India.

Data Evaluation

After applying the data mining methods comes the data set classifying the initiate results, in form of stimulating data representing knowledge depending on interestingness measures. These measures are necessary for the efficient discovery of data of value to the given user. Such measures can be used after the data mining step in order to rank the open data according to their interestingness, filtering out the uninteresting ones. More importantly, such measures can be used to guide and constrain the discovery process, improving the examine efficacy by thinning away subsets of the form space.

Naïve Byes Classification

Classification is a famous managed learning technique in data mining. It is used to spiteful meaningful information from large datasets and can be efficiently used for predicting unidentified classes. In this investigation, classification is applied to a crime dataset to predict 'Crime Category' for different states of the India. The crime dataset used in this research is real in environment; it was collected from socio-economic data from different areas in India. This paper compares the two different classification algorithms namely, Naïve Bayesian (NB) and Enhance Naïve Byes(ENB) for predicting 'Crime Category' for different states in INDIA.

KNN

In binary (two class) classification difficulties, it is useful to select k to be an odd number as this avoids secured votes. The K-Nearest Neighbor procedure is between the modest of all machine learning procedures: an object is classified by a mainstream vote of its neighbors, with the purpose presence assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, classically small). Typically Euclidean distance is used as the distance metric; though this is only appropriate to continuous variables. In cases such as text classification, alternative metric such as the intersection metric or Hamming distance, for example, can be used.

K nearest neighbors is a modest process that suppliesall obtainable cases and classifies new cases based on a comparison measure (e.g., distance functions). KNN has been used in arithmetical assessment and pattern recognition previously in the opening of 1970's as a non-parametric method.

K nearest neighbor procedure is very simple. It works based on smallest distance from the query instance to the training samples to regulate the K-nearest neighbors. The data for KNN procedure contain of numerous attribute names that will be used to categorize. The data of KNN can be any measurement scale from nominal, to quantitative scale.

The KNN procedure is presented in the following form:

Input: D, the set of k training objects, and test object z= (x', y').

Process: Calculate d(x', x), the distance between z and every object, (x, y)∈D.

Select $D_z \subseteq D$, the set of k closet training objects to z.

Output: $y' = \text{argmax}_v \sum_{(xi, yi) \subseteq Dz} I(v = y_i)$

- v is a class label
- y_i is the class label for the i^{th} nearest neighbors
- I (.) is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

In this scheme, KNN set of rules is used the appropriate result by mixing the Euclidean distance between the numerous kinds of distance metric. The Euclidean distance is as shown in below:

$$d_{ij} = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \left(x_{i2} - x_{j2}\right)^2 + \cdots + \left(x_{ip} - x_{jp}\right)^2} \quad (1)$$

Where

d_{ij} = the distance between the training objects and test object

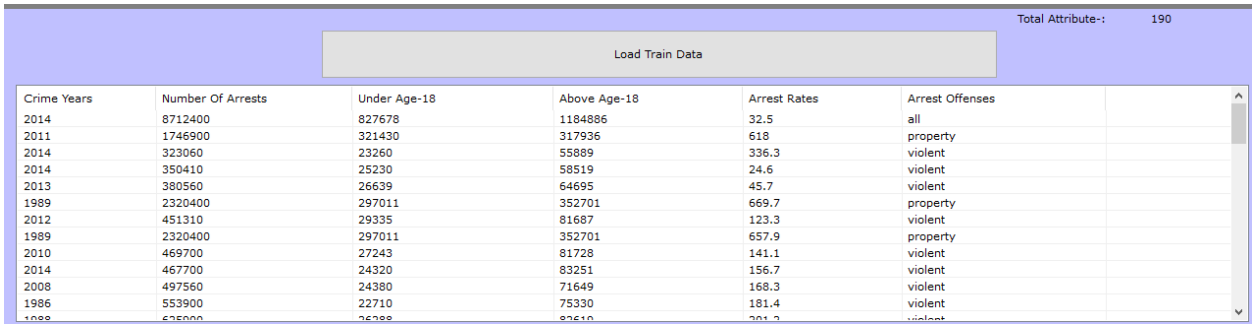
x_i = input data for test object

x_j = data for training objects stored in the database

IV. RESULT AND DISCUSSION

In every classical, the accuracy plays an important role in the approval of that model for the application. The crime Dataset classify on the basis of age factor with respect to crime year. And have to calculate fix state parameter from the dataset which perform the better efficiency of naïve bayes classification.

The results from the Graph and its explanation clearly show that % error of Hybrid Naïve bayes and KNN. Hybrid method performed better in predicting all the classes, namely All, Property and Violent. Graph illustrates the Accuracy for both the algorithms used in the experiment.



Crime Years	Number Of Arrests	Under Age-18	Above Age-18	Arrest Rates	Arrest Offenses
2014	8712400	827678	1184886	32.5	all
2011	1746900	321430	317936	618	property
2014	323060	23260	55889	336.3	violent
2014	350410	25230	58519	24.6	violent
2013	380560	26639	64695	45.7	violent
1989	2320400	297011	352701	669.7	property
2012	451310	29335	81687	123.3	violent
1989	2320400	297011	352701	657.9	property
2010	469700	27243	81728	141.1	violent
2014	467700	24320	83251	156.7	violent
2008	497560	24380	71649	168.3	violent
1986	553900	22710	75330	181.4	violent

FIG 1: Load data for training using KNN

In the fig 1, We Train the dataset of crime on the basis of age factor for overall rate and no of arrests.

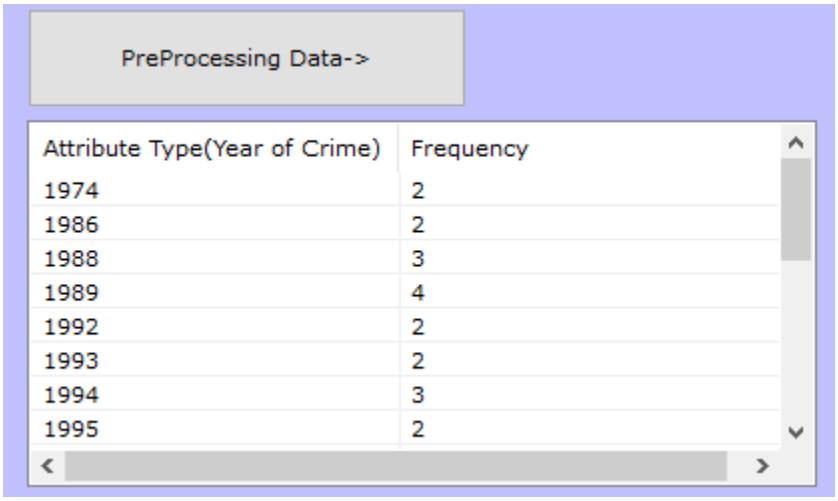
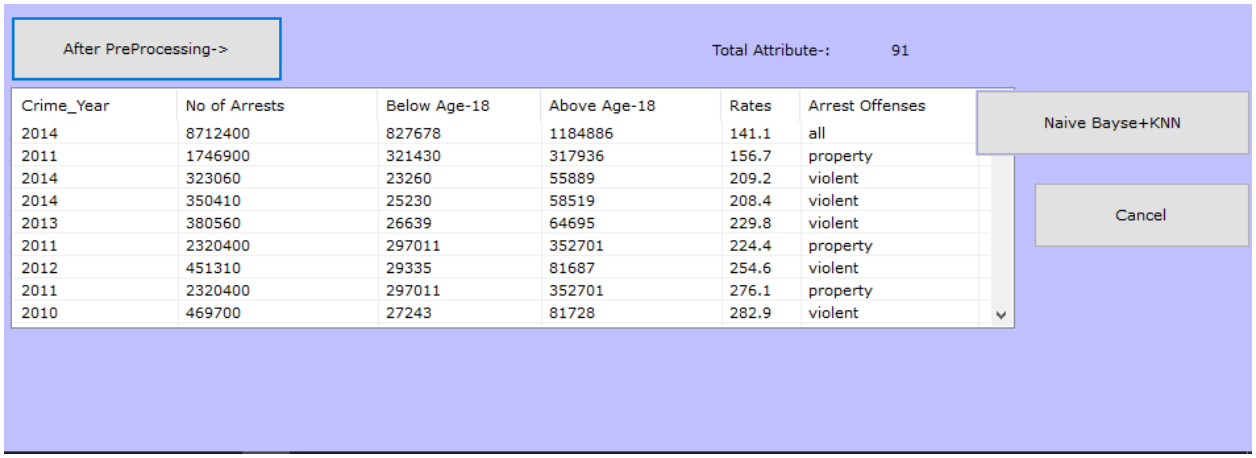


FIG 2: frequency for Crime Year for Fix State

Fig 2 shows the data preprocessing on the origin of frequency for Crime Year for Fix State.



Crime_Year	No of Arrests	Below Age-18	Above Age-18	Rates	Arrest Offenses
2014	8712400	827678	1184886	141.1	all
2011	1746900	321430	317936	156.7	property
2014	323060	23260	55889	209.2	violent
2014	350410	25230	58519	208.4	violent
2013	380560	26639	64695	229.8	violent
2011	2320400	297011	352701	224.4	property
2012	451310	29335	81687	254.6	violent
2011	2320400	297011	352701	276.1	property
2010	469700	27243	81728	282.9	violent

Figure 3: After preprocess the data using Naïve bayes and KNN

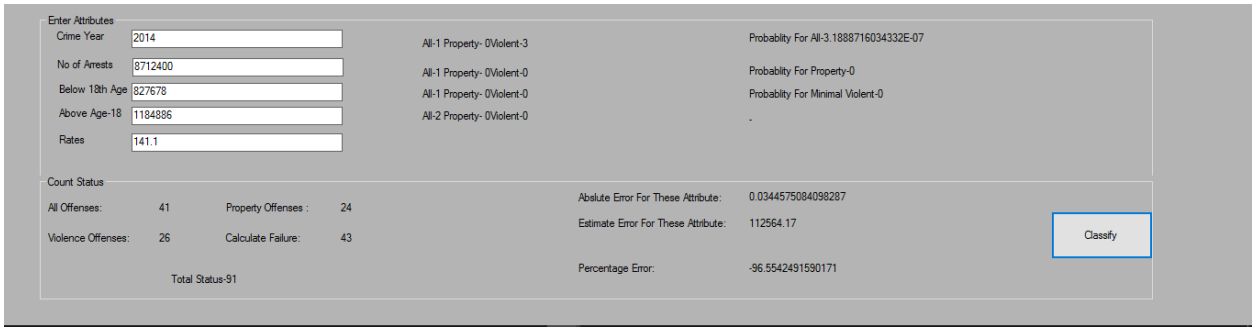


Fig 4: input random Attribute for Crime Parameter for prediction, fix state parameter for individual class and prediction value for all classes in Naïve Byes

In this paper we used hybrid method Naïve byes and KNN with respect to probability error. This is shown in fig 4.

V. CONCLUSION

The proposed model will be able to extract crime data and categorize crime records in three categories by using three functions i.e., first as Mining Specific Client's Crime Record, second as Find the users' Component attribute to evaluate all state in dataset and third as Find the multi-users Software specify for the status which is resolved by the end user point and give the optimal result. This model is useful for exacerbating the crime from the society.

In future work, we would like to examine whether it is possible to eliminate discernment and to decrease false optimistic rates without using (or knowing) the origin of a person in the prediction model and to define the effect on the accuracy of the resulting classifier. We arrange by saying that this reading validates the use fullness of discrimination-aware data mining works in practical settings.

References

[1] Bogomolov, Andrey and Lepri, Bruno and Staiano, Jacopo and Oliver, Nuria and Pianesi, Fabio and Pentland, Alex.2014. Once upon a crime: Towards crime prediction from demographics and mobile data, Proceedings of the 16th International Conference on Multimodal Interaction.

[2] Yu, Chung-Hsien and Ward, Max W and Morabito, Melissa and Ding, Wei.2011. Crime forecasting using data mining techniques, pages 779-

786, IEEE 11th International Conference on Data Mining Workshops (ICDMW)

[3] Kianmehr, Keivan and Alhajj, Reda. 2008. Effectiveness of support vector machine for crime hot-spots prediction, pages 433-458, Applied Artificial Intelligence, volume 22, number 5.

[4] Toole, Jameson L and Eagle, Nathan and Plotkin, Joshua B. 2011 (TIST), volume 2, number 4, pages 38, ACM Transactions on Intelligent Systems and Technology

[5] Wang, Tong and Rudin, Cynthia and Wagner, Daniel and Sevieri, Rich. 2013. pages 515-530, Machine Learning and Knowledge Discovery in Databases

[6.] Steven N. Durlauf and Lawrence E. Blume (Eds), 'Social Norms' in New Palgrave Dictionary of Economics, Second Edition, London: Macmillan, 2011.

[7.] Martin Innes (2003). Crime as a Signal, Crime as a Memory, Journal for Crime, Conflict and the Media, vol 1, pp 15-22.

[8.] De Knegt; H.J.; F. van Langevelde; M.B. Coughenour; A.K. Skidmore; W.F. de Boer; I.M.A. Heitkönig; N.M. Knox; R. Slotow; C. van der Waal and H.H.T. Prins (2010). Spatial autocorrelation and the scaling of species–environment relationships. Ecology 91: 2455–2465. doi:10.1890/09-1359.1

[9.] Weisstein, Eric W. "Moore Neighborhood." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/MooreNeighborhood.html

[10.]Pang-Ning Tan, Michael Steinbach, Vipin Kumar, and Addison Wesley. Introduction to Data Mining (2006) ISBN: 0-321-321136-7