A review on various classification algorithms to predict the crime status

Richa A Patel[1], Kinjal Thakar[2], and. Rina Raval[3]
[1] *silveroak college of enginnering and technology*
[2] *assistant professor in IT department, silveroak college of enginnering and technology*
[3] *professor in computer department,silveroak college of enginnering and technology*

Abstract- **in recent past, crime rate is becoming high in most highly populated country. Security is an aspect which is given higher priority by all governments in the world hence the significant task of to predict the crimes over data history. Now a days increasing user of the internet in the world is going rapidly hence cybercrime rate is also increasing. In data mining lots of algorithms are available to solve the mining problems. The crime profiling and zoning can be modelled with utilization of data mining. The predicted crime status can help to police department for investigation of criminal. Classification and clustering may prove the better methods for predict the crime status in data mining. Naïve Bayesian, k-Nearest Neighbour[3], Neural Networks may prove much better classifiers in comparisons of decision tree and support vector machine in data mining.**

Index Terms- **Data mining in crime data, Classification, clustering, predict crime status, K means algorithm [1], K-NN [3], decision tree[3], J48, SVM[4].**

I. INTRODUCTION

“Crime is as old as mankind itself.” Schafer stated that ever since the biblical misconduct happened during the period of Adam and Eve, although cultures of humanity have developed and rules have been formed since then, violence has continued. Crime has been present from the very start of humanity and has never stopped. Furthermore, crime has become a “common societal phenomenon” that it is deliberated now as part of an organization’s functional element. System captures vital information of a crime when it is being reported. Documents provide by the complaint can be uploaded into the system. This system was patterned from existing systems and the actual practice of the precinct. All necessary documents such as incident report, endorsement letters and affidavits can be printed using the system. Tracking of the status of the complaints can be performed with the system.

The main goal intended for the study of this project is to be able to explore crime management and information system development concepts in applying to community-based crime prevention. The main scope of the research study is the accuracy and reliability of crimes in India. It mainly focuses on the management of crimes from the initial reporting of the crime until the investigation process.

Classification is used method in crime dataset [6]. Main goal of employing classification algorithm would be always to create perfect prediction every and every time with high accuracy. By employing classification technique, we are able to predict potential target from previous data set.

Classification works on two Different types of data sets: (1) Training data set, and (2) Testing data set [6]. A model is assembled using training data set, and performs with the prediction by simply employing the model on testing data set. Class label of training data set are known for us.

SVM[4] and KNN [1] are data Mining algorithms utilized on selection of application like opinion classification, spam detection, etc. SVM[4]is a discriminative classifier formally defined by a separating hyper plane. Given labelled training data, the algorithm outputs an optional hyper plane which categorized new examples. SVM[4] is a supervised learning.

In two dimensional space this hyper plane is a line dividing a plan in two parts where in each class lay in either side. In machine learning, support vector machines are supervised learning algorithms that analyse data used for classification and regression analysis. SVM[4] are among the best “off-the-shelf” supervised learning algorithm.

II. RELATED WORK

Data mining in crime dataset[6] Has been created in a variety of application. Predict the crime status is just one of these. Safety [7] and risk has been analysed by many researchers. There's broad field of research for example its risk[7][8], criminal definition, applied algorithms and its prediction. Crime prediction guideline was chosen to give decision-making and recommendation process. SVM[4] ,KNN[3] and decision tree[3] were chosen as the algorithm for this particular procedure.

In the world's every country crime has been created repeatedly. The main goal of this research to protect the people from criminal. And these kinds of research also help to police to identify the criminal. They may also increase the safety [7] of the people and decrease the risk [7][8] of the population of any country.

here is also various kind of crime happened such as cybercrime [12]. Increasing user of the internet in the world is going rapidly, by the same condition in every country. This increasing is triggering the vulnerability of securing information technology advantage and user privacy. Cybercrime [12] is all about the crimes in which communication channel and communication device has been used directly or indirectly as a medium whether it is a Laptop, Desktop, PDA, Mobile phones, watches etc.

Children use Internet in a daily activity both at home and school. Despite the benefits that Internet might has for children, there are risks [7][8] surrounded children that we must recognize and be careful of it. Internet use has serious risks including child sexual harassment and child pornography. Unfortunately, cybercriminals [10] take advantages of technological advancement and exposed young people as they are the most valuable at any society.

III. EXISTING WORK

If Based on existing research, it has been identified that data mining techniques aid the process of crime detection. Some examples of data mining techniques usage to analyze crime data are classification and machine learning algorithms. employ an ensemble of data mining. classification techniques for crime forecasting. Several classification methods that are included in the study are One Nearest Neighbor (1NN), Decision Tree[3] (J48), Support Vector Machine (SVM)[4], Neural Network (Neural) with 2-layer network, and Naïve Bayesian (Bayes).

Detecting patterns of serial criminal behaviors and crime activity geographically by using clustering in are also used for pattern recognitions and predictions.in his study, applies clustering by considering the geographical approach which shows regional crimes on a map and clusters crimes according to their types by using a combination of K-means Clustering Algorithm and Weighting Algorithm. Clustering and graph representations are also used to obtain similar crime and group classes of criminals, as well as to visualize the results. Clustering features such as shape, size, and distribution are able to help understand more details about relevant crimes including a clustering analysis on US State database. Association mining is one of the acceptable methods to discover the underlying novel patterns on a large volume of crime data. Other techniques, such as semantic analysis and text mining are used to extract entity extraction from FBI bulletins.

In a fuzzy association rules mining application for community crime pattern discovery is proposed. The application produced interesting and meaningful rules at regional and national levels and, to extract novel rules, a relative support metric is defined. It employs temporal association rule mining when the amount of data is growing.it present a new distance measure to evaluate individual criminals using the profiles to cluster them and enable recognition of criminals' classes. This research also present a particular distance measure for a combination of the profile differences with crime frequency and change of criminal behavior over time.

IV. PROPOSED WORKFLOW

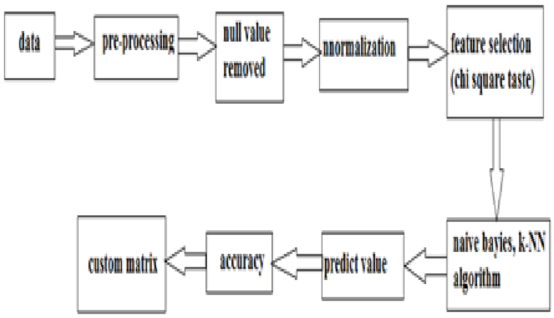

Figure 1.existing Work flow for crime status

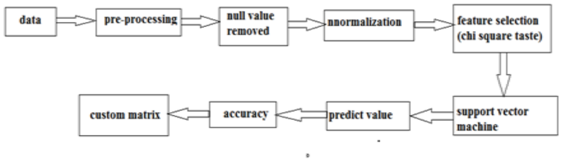

Figure 2.proposed Work flow for crime status

V. FLOW OF THE DIAGRAM

Process of Acute kidney custom matrix Will Be split in to 3 segments. The first step is really for Identify the data format of the given crime data. There'll soon be couple actions in section, that can soon be achieved as data pre-processing [7]. It's going to include Data set from released by any country. From then on, data need to be convert in one proper format and then need to be remove null value from dataset. Initial data set will incorporate criminal's information and which kind of crime, etc.

Data Pre-Processing will comprise some fundamental Performance for data cleaning and integrating. Data redundancy and data complexity is going to be lowered in such steps. To identify the risk zoning of particular areas predicated on such features, Classification data mining technique KNN[3] or SVM[4] is going to be properly used. By employing this sort of algorithm, we could possibly secure more accuracy without time complexity.

Secondly Section is going to be for applying feature selection method. Where chie square teast will provide the two categorical variables for single population to assist decision making, looking after area in risk and crime rates to increase survival and also increase risk zone for people. Support vector machine algorithm is going to be generated based on principle, which create a more accuracy in crime database.

method	Precision (%)		recall (%)	
	Set1(1:44)	Set2(2:94)	Set1(1:44)	Set2(2:94)
Naïve Bayesian	86.7	87.5	84.6	84.4
Support Vector Machine (SVM)	84.6	85.7	85.0	86.3
Neural Network (Multilayer Perceptron)	85.0	86.1	85.6	86.5
k-Nearest Neighbor (K=10)	85.9	87.3	87.5	88.0

Figure3. precision and recall.

Third Section is really for evaluating the end result of several measures such as sensitivity, precision, recall and accuracy.

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 88.00 which is good for this model as it's above 87.5.

Recall = TP/TP+FN.

VI.COMPARISION

the precision and recall values for all five classifiers are not significantly differentiated from each other and also there is no dominating relation between ROC curves in the entire range. In this situation, AUC provides a good summary for comparing the classifiers. It also compared accuracy and the Area Under Curve (AUC) with different classifiers in various dataset. They conclude that the best tool for classifier comparison is AUC which helps users to better understand the performance of the classifiers.

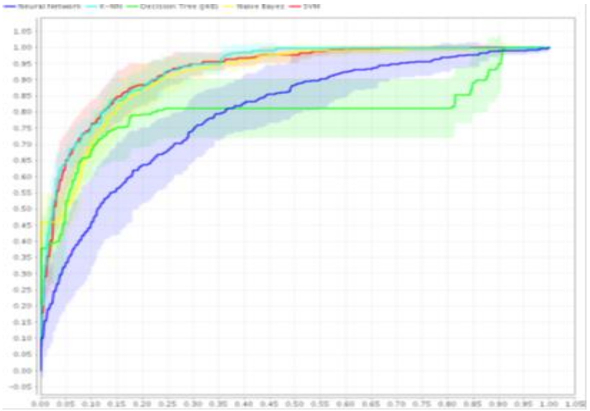

Figure 3. ROC comparison of attributes

VII. CONCLUSION

The aim of this study is to classify the given specified experimental dataset into two categories which are critical and non-critical. In this regard, we used five classification algorithms by combining two different ways of feature selection techniques, manually and Chi square, to determine more accurate classifiers. From the experimental results, support vector machine algorithm presents the best accuracy, specifically by using Chi-square feature selection technique. We have shown via exploratory comparisons in terms of AUC that Naïve Bayesian, Neural Networks, and k-Nearest Neighbor predict lower than the Support Vector Machine and Decision Tree due to the nature of this dataset. Through the implementation of Chi-square feature selection technique in RStudio, it is demonstrated feature selection is an important phase to enhance the mining quality.

REFERENCES

[1] Suresh Babu Changalasetty, Lalitha Saroja Thota, Ahmed Said Badawy and Wade Ghribi, Classification of Moving Vehicles using K-Means Clustering, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT-2015), India, Mar 2015.

[2] A. Malathi, Dr. S. Santhosh Baboo, Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters, Global Journal of Computer Science and Technology Volume 11 Issue 11 Version 1.0 July 2011.

[3] Charles X. Ling, Jin Huang, Harry Zhang, “AUC: A Better Measure than Accuracy in Comparing Learning Algorithms”. In Proceedings of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence (AI'03), pp. 329-341, 2003. [4] A.CichockiandR.

[4] Jesse Davis, Mark Goadrich, “The Relationship between Precision-Recall and ROC Curves”, In Proceedings of the 23rd International Conference on Machine Learning (ICML'06), pp. 233-240, 2006.

[5] K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana ,2012 Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model

[6] https://www.gov.uk/government/publications/offences-recorded-by-the-police-in england-and-wales-by-offence-and-police-force-area-1990-to-2011-12

[7] Y. Jiang, Y. Xiang, X. Pan, K. Li, Q. Lv, R. P. Dick, L. Shang, and M. Hannigan, “Hallway based automatic indoor floorplan construction using room fingerprints,” in Proc. of ACM Ubicomp, 2013.

[8] J. Yin, Q. Yang, and L. M. Ni, “Learning adaptive temporal radio maps for signal-strength-based location estimation,” IEEE Transactions on Mobile Computing, vol. 7, no. 7, pp. 869–883, 2008.

[9] Y. Zheng, F. Liu, and H.-P. Hsieh, “U-air: when urban air quality inference meets big data,” in Proc. of ACM SIGKDD, 2013.

[10] The Cyber Theft Ring. Available at:http://www.fbi.gov/news/stories/2010/october/cyber banking-fraud/cyber-banking-fraud-graphic accessed on Sunday 09.14 PM september 12, 2014.

[11] Indonesia Tapping Cases. Available at:http://www.tempo.co/read/news/2014/02/21/063556304/4-Kasus-Penyadapan-Besar-di-Indonesia accessed on Friday 08.00 AM 19 December 2014.

[12] 10 Prevention of Fraud Crime. Available at:http://www.fraudlabs.com/fraudlabswhitepaperpg1.htm accessed on friday 08.30 AM 19 December 2014.

[13] E. Han, “Australians conned out of $1 million in tax scams so far this year, says ACCC”, The Sydney Morning Herald, 29 July 2016.

[14] Y. Yang, M. Manoharan, and K.S. Barber, “Modelling and Analysis of Identity Threat Behaviors Through Text Mining of Identity Theft Stories,” IEEE Joint Intelligence and Security Informatics Conference (JISI), the Hague, the Netherlands, pp. 184-191, September 24-26, 2014.

[15] D. Lacey, “Guilty Before Proven Innocent: The Truth About the Identity Theft Customer-Organization Dynamic”, ID360 The Global Forum on Identity, Austin Texas, 4-5 May 20.

[16] E. Kritzinger, “Enhancing Cyber Safety Awareness among School Children in South Africa through Gaming,” in Science of Information Conference 2015, 2015, pp. 1243–1248.

[17] B. O. Neill and E. Staksrud, “Final recommendations for policy,”2014.

[18] Theguardian, “Why increasing digital Arabic content is key for global development,” 2014. [Online].Available:https://www.theguardian.com /media-network/media-networkblog/ 2014/apr/28/global-development-digital-arabic-content.

WEB REFERANCE

[1] http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

[2] https://en.wikipedia.org/wiki/Support_vector_machine

[3] http://ieeexplore.ieee.org