Cluster based zoning for crime information.

Richa A Patel[1], Kinjal Thakar[2], and Rina Raval[3]
[1] *Silver oak college of engineering and technology*
[2]*Assistant Professor in IT department, Silveroak College of engineering and technology*
[3]*Professor in computer department, Silveroak College of engineering and technology*

***Abstract*—Crime is an old as mankind itself in around the world. Due to the huge dataset, Prediction of crime status is highly complex. Crime status prediction is associated with types of crime in particular area, which is helpful to security department such as how to distribute police in particular areas. For Crime related solutions, many algorithm based solutions have been published. Classification and clustering algorithm are two of them, who provide guidelines with recommendations to assist decision making, accurate for people at risk. Naive Bayesian (0.898), k-Nearest Neighbor (k-NN) (0.895) and Neural Networks (0.892) were selected as the basic data mining algorithms for this process. To improve the accuracy and of the system, Support Vector Machine algorithm has been used. Furthermore, the performance of mining result is improved by using chi-square feature selection technique. Created custom matrix is divided on cluster wise zoning map.**

***Index Terms*—data mining, Crime dataset, Cluster zone, Naïve Bayesian, k-NN, SVM**

I. INTRODUCTION

find it useless and waste of time since they believe that the police are not capable of solving such The proponents were able to identify some of the major problems that were visible in the Police. These problems focused mainly on the stored data of the crime to predict the crime zone. These problems trigger the community not to report crimes happening since them crimes.

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. Therefore, data mining methods have become the main tools to analyze data and to discover knowledge from them. Here, data mining refers to an integration of multiple methods such as classification, clustering, evaluation, and data visualization.

The higher number of crimes in many country have forced to government to use the modern technologies and methods for stop the crimes or to decrease the crime ratio. Today, a high number of crimes are causing a lot of problems in many different countries. 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. Firstly, Crime dataset[6] need to makes in one proper form which is data pre-processing.

In this study, mainly use the real crime dataset[6] for the comparison of the classification algorithm. Examined classifiers are Naïve Bayesian [3], k-Nearest Neighbor (k-NN)[3] and, Support vector machine[4].

By experiment, result of three algorithms for crime dataset[6] studied and compared, and more efficient algorithms in predicting the crime status are then identified. There are many tools available for data mining. For this study, the R tool is chosen which is freely available over internet.

II. RELATED WORK

In some research work, authors have comprehensively compared different data classification techniques and their prediction accuracy for crime related work. Authors have compared Decision Tree (J48)[3], Naïve Bayesian[3], k-NN[3], Neural Networks classifiers using performance measures such as Area Under Curve (AUC) using Rapid minor tool. Authors have also compared these classifiers on various accuracy measures like Sensitivity, specificity, precision, recall and Accuracy, Kappa by implementing on R-tool.

Based on existing research, generally clustering and classification methods are used for crime related research work. 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 some research worker focuses on crime analysis[1][10] by implementing clustering algorithm on crime dataset[6] using data mining tool and there they do crime analysis[1][10] by considering crime homicide and plotting it with respect to year and got into conclusion in the form of homicide. From the clustered results it is easy to identify crime trend over years and can be used to design precaution methods for future.

Lots of country as being the first rank of the most provided cyber-crime[10] attack is a huge problem. Cyber-crime[10] can disrupt stability of the country. The second situation of cyber-crime[10] is internet risk[7][8] for everyone, especially for children. Therefore based on this kind of crime dataset[6] data mining techniques are applied for better solution.

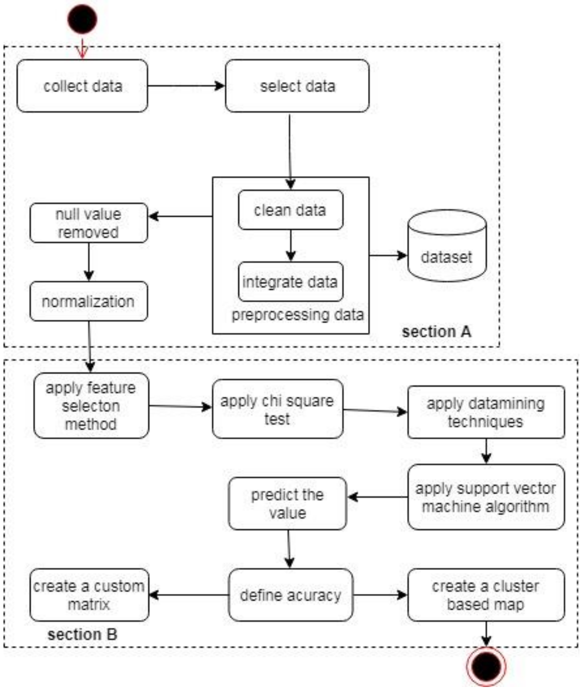
III. PROPOSED WORKFLOW

Flow of diagram:

Section A: Data collection

1. Collect Data: Data used to perform data mining analysis will be collected from police department or published by government of country in web-based format where all kinds of crime data were recorded.

2. Data Selection: crime data is selected from the US FBI Uniform Crime Report. These dataset were collected and published by United States. The dataset consisted of 2215 total instances and 147 attributes for communities, 125 predictive, 4 non-predictive and 18 potential goal attributes. The data in each instance belong to different states in the USA. The states are represented in the form of number, every number representing its respective American state.



3. Preprocessing Data: Data preprocessing will include some basic operation for data cleaning and integrating. In this process, redundant data and data complexity will be reduced. For the first step, the goal for data cleaning is to decrease noise and handle missing values. There are a number of methods for treating records that contain missing values such as omitting the incorrect fields(s), omitting the entire record that contains the incorrect field(s), automatically entering or correcting the data with default values, deriving a model to enter or to correct the data, replacing all values with a global constant and using the imputation method to predict missing values.

In this study, some communities were removed based on occurrences of significant missing or known incorrect crime statistics. Certain attributes contain a significant number of missing values, more than 80%, for which the data was unavailable or not recorded for particular communities. These attributes with high amount of missing values were removed such as the *pctPolicWhite*, and the *pctPolicBlack*. All kinds of crime attributes (potential goal attributes) with missing values were removed because only a total number of violent crimes per 100K population attribute, which is *violentPerPop* will be considered as class. All 221 instances with *violentPerPop* missing values were removed and 1994 remained because *violentPerPop* has been chosen as the goal attribute.

For the second step, we performed data normalization, discretization, and data type transformation. For all attributes except state, min-max normalization to [0, 1] is used to avoid the large value issue as it has the advantage of protecting exactly all relationships in the data and to prevent any bias injection. Next, we also discretized the selected class, which is total number of violent crimes per 100K population (v*iolentPerPop*), into a binomial class crime status (*CrimeStatus*). In order to perform prediction, the goal class should be nominal in nature.

After discretization, we also performed data type transformation because initial class has two values, “critical” and “non-critical”. If the value in violentPerPop is less than forty percent, than the value of CrimeStatus set to “non-critical”, otherwise “critical”. The states are converted from nominal to numeric; whereby every number represents the respective American state.

Section B: Feature selection method:

Feature selection is used to remove the irrelevant or redundant attributes. Feature selection has several objectives such as enhancing model performance by avoiding over fitting in the case of supervised classification. In this study, as mentioned above, crime status (*Crime Status*) was chosen among eighteen potential goal attributes as the desirable dependent variable.

In this study, Using the Chi-square test to detect the correlated attributes as they will ruin the

Classification result because of high dependency among some attributes caused redundancy and subsequently inaccurate results. Chi-square is one of the most effective feature selection methods of classification. Chi-square feature selection has been adopted and 94 attributes are selected.

1. Data mining techniques: Dataset[6] will be prepared using this process. All data of the criminals who will under United States old or published by US government and unnecessary data such as personal information of criminal will cut-off. Remaining data were categorized in to characteristics, dangerous, type of crime and critical-non critical etc. that contain their own attributes.
2. Identify the Result: To identify the risky areas[7][8] of country based on attributes, classification data mining technique including Naïve Bayesian, k-Nearest Neighbor and Neural Networks were applied in data mining process of previous work. But, Support Vector Machine[4] will be used as an algorithm for better accuracy.

The result from this section will be used for monitoring police department to identify risk[7][8] that which area is much critical area.

Evaluation

To evaluate the proposed approach, 2*2 of table is planned to use as depicted in below table:

		Actual Result	
		critical	Non-critical
Test Results	Positive	495 (true positive)	5 (false positive)
	Negative	26 (false negative)	1375 (true negative)

By this table, we can also calculate some measures as sensitivity, specificity and

Accuracy, precision and recall value as below:

Sensitivity = 0.950
Specificity = 0.996
Precision = 0.990
Recall = 0.950
Accuracy = 0.984

Sensitivity value is used to measure the proportion of positives that are correctly identified. Specificity value is used to measures the proportion of negatives that are correctly identified. Accuracy value is used the precision of the system.

Cluster based zoning map:

Using clusters for classification makes the model simpler and increases the accuracy at the same time and we can predict the result as a supervised cluster. Mean-shift is a clustering approach where each object is moved to the densest area in its vicinity, based on kernel density estimation. Eventually, objects converge to local maxima of density. Similar to k-means clustering, these "density attractors" can serve as representatives for the data set, but mean-shift can detect arbitrary-shaped clusters similar to DBSCAN. Besides that, the applicability of the mean-shift algorithm to multidimensional data is hindered by the unsmooth behavior of the kernel density estimate, which results in over-fragmentation of cluster tails.

Zone wise created cluster map in R tool is given below:

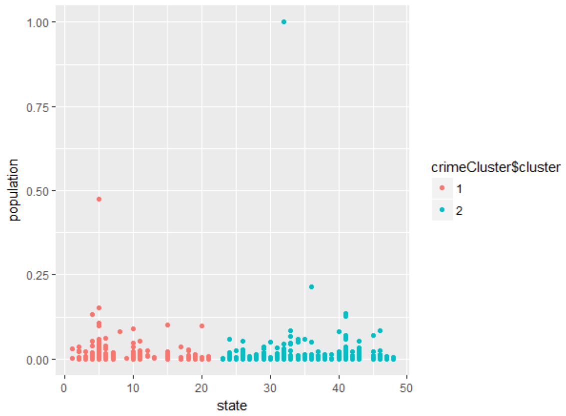


Figure 1: zone wise cluster map

IV. ALGORITHM FOR PROPOSED SYSTEM

“Support Vector Machine” (SVM)[4] is a supervised machine learning algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

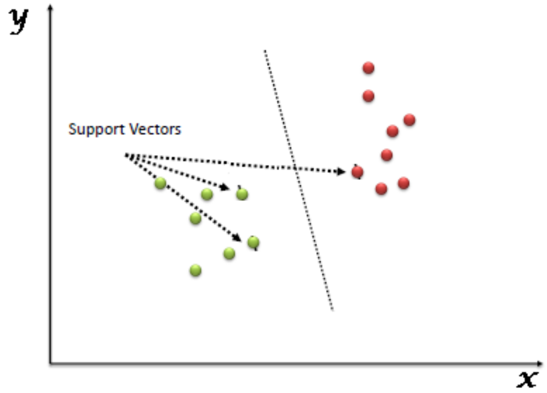


Figure 2 support vector machine

This section lists some suggestions for how to best prepare your training data when learning an SVM[4] model.

- **Numerical Inputs**: SVM[4] assumes that your inputs are numeric. If you have categorical inputs you may need to convert them to binary dummy variables (one variable for each category).
- **Binary Classification**: Basic SVM[4] as described in this post is intended for binary (two-class) classification problems. Although, extensions have been developed for regression and multi-class classification.

$$\vec{w}x_i - b \geq 1 \text{ if } \theta_i = 1$$
$$\vec{w}x_i - b \leq 1 \text{ if } \theta_i = -1$$

Where normal vector to the hyperplane, θi is denotes classes & xi denotes features. The Distance between two hyperplanes is, to maximize this distance denominator value should be minimized.
For proper classification, we can build a combined equation:

$$\|\vec{w}\|_{min} \text{ for } \theta_i \left(\vec{w}x_i - b\right) \geq 1 \;\; \forall i = 1, 2, \ldots., n$$

Non-Linearly Separable: To build classifier for non-linear data, we try to minimize.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\vec{w}\cdot\vec{x}_i - b)\right)\right] + \lambda\|\vec{w}\|^2,$$

Here, max() method will be zero(0), if x_i is on the correct side of the margin. For data that is on opposite side of the margin, the function’s value is proportional to the distance from the margin. where, determines tradeoff b/w increasing the margin size and that is on correct side of the margin.

V. CONCLUSION

The aim of this study is to classify the given specified experimental dataset[6] into two categories which are critical and non-critical. In this regard, we used three 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[4] algorithm presents the best accuracy, precision and recall, specifically by using Chi-square feature selection technique. We have shown via exploratory comparisons in terms of custom matrix that Naïve Bayesian and k-Nearest Neighbor predict lower than the Support Vector Machine[4] due to the nature of this dataset[6]. Through the implementation of Chi-square feature selection technique in R Studio, it is demonstrated feature selection is an important phase to enhance the mining quality. We can use support vector machine[4] algorithm to get more accuracy instead of simple naïve Bayesian[3], k-NN[3] and J48 algorithm. We can divide the result in cluster zone to easy prediction of critical and noncritical areas.

REFERENCES

[1] D. Lacey and C. Cross, "What victims want: An examination of identity theft restoration from a victim's perspective," 28th Annual Australian and New Zealand Society of Criminology Conference, 25 November, 2015

[2] D. Lacey and P. Salmon, "It's Dark in There: Using Systems Analysis to Investigate Trust and Engagement in Dark Web Forums," Engineering, Psychology & Cognitive Ergonomics, 2015

[3] K. Turville, S. Firmin, J. Yearwood, and C. Miller, "Understanding victims of identity theft: A grounded theory approach," Proceedings of the 5th International Conference on Qualitative Research in IT & IT in Qualitative Research, Brisbane, Australia, 2010

[4] . Livingstone and L. Haddon, Comparing children's online opportunities and risks across Europe. 2009.

[5] S. Livingstone, G. Mascheroni, and K. Ólafsson, "Children ' s online risks and opportunities : Comparative findings from EU Kids Online and Net Children Go Mobile Executive summary," 2014.

[6] N. lee Quenna, "Nurture cyberkids to cut risks, UN agencies urge," INQURER.NET, 2015. [Online].Available:http://technology.inquirer.net /45110/nurture-cyberkids-to-cut-risks- un-agencies-urge. [Accessed: 01-Nov-2015].

[7] Malathi.A 1 ,Dr.S.Santhosh Baboo 2 and Anbarasi . A 31 Assistant professor ,Department of Computer Science ,Govt Arts College ,Coimbatore , India . 2 Readers , Department of Computer science , D.G. Vaishnav Collge ,Chennai , India , 2011 An intelligent Analysis of a city Crime Data Using Data Mining.

[8] Malathi , A; Santhosh Baboo , S, 2011 An Enhanced Algorithm to Predict a Future Crime using Data Mining **.**

[9] MA, J., Saul, L.K., Savage, S., and Voelker, G.M., "Learning to Detect Malicious URLs". in ACM Transactions on Intelligent Systems and Technology, New York, NY, USA, Article 30 April 2011, ACM, pp. 1245-1254. DOI=http://dl.acm.org/citation.cfm?id=1961202.

[10] S. Sheng, B. Wardman, G. Warner, L. Cranor,J. Hong, and C. Zhang. (2009) An empirical analysis of phishing blacklists. [Online]. Available: http://ceas.cc/2009/papers/ceas2009-paper-32.pdf.

[11] J. Ma, L. K. Saul, S. Safage, G. M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proc. ACM SIGKDD, Paris, France, 2009, pp. 1245-1253.

[12] R. B. Basnet, A. H. Sung, Q. Liu, "Rule-Based Phishing Attack Detection" Proc. International Conference on Security and Management (SAM'11), Las Vegas, NV, 2011, pp. 624-630.

[13] Xiaoqing GU, Hongyuan WANG, and Tongguang NI "An Efficient Approach to Detect Phishing Web" Journal of Computational Information Systems, 2013, pp.5553-5560.

[14] M. G. Alkhozae and O. A. Batarfi, "Phishing websites detected based on phishing characteristic in the webpage source code," in International Journal of Information and Communication Technology Research, vol. 1, no. 6, Oct. 2011, pp. 283–291.

[15]R. Patil, B. Dasharath Dhamdhere, K. S. Dhonde, R. G. Chinchwade and S. B. Mehetre, "A hybrid model to detect phishing-sites using clustering and Bayesian approach," *International*

Conference for Convergence for Technology-2014, Pune, 2014, pp. 1-5.

[16] C. L. Tan, K. L. Chiew and S. N. Sze, "Phishing website detection using URL-assisted brand name weighting system," *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Kuching, 2014, pp.054-059. doi: 10.1109/ISPACS.2014.7024424.

[17] J. Zhang and Y. Wang, "A real-time automatic detection of phishing URLs," *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, Changchun, 2012, pp. 1212-1216.

[18] Phish Tank. (2016, Nov.) Statistics about phishing activity and phish tank usage. [Online]. Available:http://www.phishtank.com/stats/2013/01/.

[19]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.

[20]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.