Building a Learning Machine Classifier with Inadequate Data for Crime Prediction

Trung T. Nguyen, Amartya Hatua, and Andrew H. Sung School of Computing, the University of Southern Mississippi, Hattiesburg, MS, U.S.A. Email: {trung.nguyen, amartya.hatua, andrew.sung}@usm.edu

Abstract—In this paper, we describe a crime predicting method which forecasts the types of crimes that will occur based on location and time. In the proposed method the crime forecasting is done for the jurisdiction of Portland Police Bureau (PPB). The method comprises the following steps: data acquisition and pre-processing, linking data with demographic data from various public sources, and prediction using machine learning algorithms. In the first step, data pre-processing is done mainly by cleaning the dataset, formatting, inferring and categorizing. The dataset is then supplemented with additional publicly available census data, which mainly provides the demographic information of the area, educational background, economical and ethnic background of the people involved; thereby some of the very important features are imported to the dataset provided by PPB in statistically meaningful ways, which contribute to achieving better performance. Under sampling techniques are used to deal with the imbalanced dataset problem. Finally, the entire data is used to forecast the crime type in a particular location over a period of time using different machine learning algorithms including Support Vector Machine (SVM), Random Forest, Gradient Boosting Machines, and Neural Networks. Finally, the results are compared.

Index Terms—data mining, learning machine classifier models, missing features, random forest, gradient boosting, SVM, neural networks

I. INTRODUCTION

Crime is a common problem in nearly all societies. Several important factors like quality of life and the economic growth of a society are affected by crime. There are many reasons that cause different types of crimes. In the past, criminal behavior is believed to be the result of a possessed mind and/or body and the only way to exorcise the evil was usually by some torturous means [1]. A person's criminal behavior can be analyzed from different perspectives like his/her socio-economic background, education, psychology, etc. Researchers have done exhaustive research on these factors. Data mining and analytics have contributed to the development of many applications in medical, financial, business, science, technology and various other fields. Likewise, to obtain a better understanding of crime, machine learning can be used for crime data analytics.

Analysis and forecasting the nature of crime has been done based mainly on the criminal's economic status, psychology, race, social background, and the demographics of a particular location. In the article by Gottfredson [2], the author discussed how to make a prediction whether a person will be criminal. On the other hand, he summarized and reviewed many of the previous works in order to identify general problems, limitations, potential methods and general nature of prediction problems of crime. The scope of that paper was limited to individual prediction, it did not address global prediction problems like predicting the number of offenses or offenders to be expected at a given time and place. In [3] Hongzhi et al. used improved fuzzy BP neural network to crime prediction. There is no mention of place, time and type of the crime. In [4,5] Mohler used crime hotspot to forecast a particular type of crime (gun crime) in Chicago; they did not address the others issues like other types of crime and occurrence time of those crimes. In [6] Tahani et al. focused on all the three major aspects of crime forecasting: place of crime, time of crime and type of crime. They performed the experiment using some machine learning algorithms for the state of Colorado and California of the United States. In their research, they used only the dataset with its information based on the National Incident Based Reporting System (NIBRS) [7], where information related to crime type, crime time and crime place was present but they did not consider any information about demographic, economic and ethnic details of criminals.

In order to forecast crimes successfully, we need to forecast the three main parameters of a particular crime: its type, location and time. Also, the methodology of crime prediction should consider the pattern of previously happened crimes and the other external factors like demographic, economic and ethnic details of criminals. In the present article, we have taken care of all the above mentioned factors. Our main objective is to forecast a crime along with its type, location and time.

In the following we describe data pre-processing, prediction methods, results and conclusion.

II. PROPOSED METHODOLOGY

Our proposed methodology can be broadly divided into four phases: Data acquisition, Data preprocessing, Application of Classification algorithm, Finding result

Manuscript received August 1, 2016; revised May 1, 2017.

and drawing the conclusion. Diagrammatic representation of proposed methodology is given in Fig. 1.



Figure 1. Proposed methodology

The data is acquired from Portland Police Bureau (PPB) and the public government source American FactFinder. Detailed descriptions about each of the phases are provided below.

III DATA PREPROCESSING

One of the key contribution of this article is data preprocessing. In previous researches, either only the crime occurrence data obtained from police were considered or data related to the criminals were considered, while in this research both aspects are considered at the same time. In addition, several other data preprocessing techniques are used as described next.

A. Description about Dataset

In our experiments, we have used data from two different sources: first, the dataset provided by the PPB is for the period of March 1, 2012, through September 30, 2016 [8], and the dataset from the American FactFinder website [9]. The data in PPB dataset is listed in calls-forservice (CFS) records giving the following information: Category of crime, Call group, Final case type, Case description, Occurrence date, X and Y coordinate of the location of the crime and Census tract. The data in American FactFinder is the census data of Portland area. From this data, we obtained information about economic, education, employment and racial background of people in this area.

B. Data Cleaning

When we examine the data from PPB, there were some missing value in census tract information. Those data points are ignored as we have enough to perform our experiments.

C. Data Reduction

In the dataset describing crime type, four different types of parameters are used such as Category, Call groups, Final case type and Case descriptions. Out of these four types, we are going to forecast only the final case type. Other parameters are not important in this case. So, we have not included those parameters in our experiments. Thereafter, we have performed dimensionality reduction and get a dataset with a reduced number of dimension: our new dataset has a total of five parameters (Final case type, Case description, Occurrence date, X and Y coordinate of the location of the crime and Census tract).

D. Data Transformation

In the dataset provided by PPB, one of the fields is census tract. The census tract information was not given in the standard format; we converted that into elevendigit census tract value, because in the later part of our experiments it is necessary to map PPB data with data from the American FactFinder dataset. As all the PPB data are collected from Multnomah county of Oregon state so the first five digits are 41051. Then census tract value in the PPB dataset will be leading zero padded, to form six digit numbers of Multnomah county. Now to get the standard census tract value with eleven digits number, we have appended 41051 before the six digits number.

The total area of each of the census tract is different than others. So, we divided each of the census tract into small clusters with a fixed area of 0.71 square mile. The number of clusters in a particular census tract thus depends on the total area of the census tract. We derived a new parameter named Area Code, to identify the location of a crime using census tract of that place and the cluster number of that place. In our original dataset, X and Y coordinate of the crime location is provided, from this information we have created clusters of an area. If a crime is committed in cluster number "MM" of census tract "41051000XXX", then Area Code of that crime will be "41051000XXXMM". This is a unique id for location and in the rest of this article we used this parameter to identify the location of a crime.

E. Data Discretization

In the dataset provided by PPB, one of the fields is date. However, our objective is to predict the crime within a span of seven days. Therefore, instead of a particular date we converted that to its corresponding day of the week and week in the year. Then we can divide all the crimes into their occurrence day out of the seven days of the week, and occurrence week out of the 53 weeks of the year. That makes it easy to handle the data and it also helps to achieve our target more easily.

F. Data Integration

Data integration is one of the most important steps of this work on crime prediction. Economic, Demographic, Educational and Ethnic information about the people of Multnomah County of the state of Oregon are collected from the census data provided by American FactFinder and then integrated with the data provided by PPB. A total of 21 features are added, and description of the features are provided in Table I below.

TABLE I. FEATURES ADDED BY CENSUS DATA

Demographic ID	Description	
HC01_EST_VC01	Total; Estimated ; Population for whom poverty	
	status is determined	
HC02 EST VC01	Below poverty level; Estimate; Population for	
HC02_ES1_VC01	whom poverty status is determined	
UC02 EST VC01	Percent below poverty level; Estimate;	
HC05_ES1_VC01	Population for whom poverty status is	

Demographic ID	Description		
	determined		
	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC14	LATINO ORIGIN		
	- One race - White		
	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC15	LATINO ORIGIN		
	One race - Black or African American		
	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC16	LATINO ORIGIN		
	- One race - American Indian and Alaska Native		
	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC17	LATINO ORIGIN		
	- One race - Asian		
	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC18	LATINO ORIGIN		
	- One race - Native Hawaiian and Other Pacific		
	Islander		
MONT FOR MONT	Total; Estimate; RACE AND HISPANIC OR		
HC01_EST_VC19	LATINO ORIGIN		
	- One race - Some other race		
HC01_EST_VC20	Total; Estimate; RACE AND HISPANIC OR		
	LATINO ORIGIN		
	- I WO OF MORE FACES		
HC01 EST VC28	ATTAINMENT		
11001_101_+010	- Less than high school graduate		
	Total; Estimate; EDUCATIONAL		
HC01_EST_VC29	ATTAINMENT		
	- High school graduate (includes equivalency)		
HC01 EST VC30	ATTAINMENT		
IIC01_LD1_VC50	- Some college, associate's degree		
	Total; Estimate; EDUCATIONAL		
HC01_EST_VC31	ATTAINMENT		
	-Bachelor's degree or higher		
UC01 EST VC26	Total; Estimate;		
HC01_E31_VC30	EMPLOYMENT STATUS - Employed		
HC01 FST VC39	Total; Estimate;		
Heor_Esr_ves/	EMPLOYMENT STATUS - Unemployed		
HC01 EST VC51	Total; Estimate; All Individuals below: - 50		
	percent of poverty level		
HC01_EST_VC52	Total; Estimate; All Individuals below: - 125		
	Total: Estimate: All Individuals below: - 150		
HC01_EST_VC53	nercent of poverty level		
	Total: Estimate: All Individuals below: - 185		
HC01_EST_VC54	percent of poverty level		
UC01 EST VC55	Total; Estimate; All Individuals below: - 200		
HC01_EST_VC55	percent of poverty level		

We observe that there exist certain relations between each of these features and the rate of crimes for a particular census tract. In the census data, the information is mostly given in the form of percentage. From this information, we generated two types of dataset. Our first dataset is created by integrating PPB data and demographic data based on census tract information provided by FactFinder. On the other hand, in the second type of dataset, we assigned values randomly for those parameters in such way that, the overall percentage is the same. In the experiment, we have 92,715 data points. Based on the data distribution from the demographic data we have, we generated missing data using random variable with pre-determined distribution. For example, the HC01_EST_VC01 attribute represents the percentage of below poverty level for each census tract. In the present context, 9.1% of people are below poverty level at census tract 100. So, in the newly generated data, 9.1% data points of census tract 100 will be assigned as 0 (which denotes below poverty level) and the rest as 1 (which denotes not below poverty level). Similarly, for all other attributes, missing data will be assigned based on the percentage mentioned in census data.

IV LEARNING MACHINE MODELS

For this crime prediction problem, we employed several machine learning algorithms to build models to accomplish the classification task and then compared the results of them. The learning machines used include Support Vector Machines or SVM [10, 11, 12], Random Forest [13], Gradient Boosting Machines [14], and multilayer neural networks (using Matlab toolbox, employing scaled conjugate gradient training and resilient backpropagation training algorithms) [15, 16, 17, 18]. These learning machines are well-known and details about them can be found in a variety of references.

V RESULT

All the models described in the previous section were trained and tested in our crime prediction tasks. The following sections present the result. The first subsection will discuss our solution of the imbalanced dataset we face. The second and third subsections describe the results for our two kinds of dataset we have after preprocessing.

A. Dealing with Imbalanced Dataset

Below is the distribution of our compilation dataset after preprocessing. In TABLE II Class 1, Class 2, Class 3, Class 4 represent the following classes: STREET CRIMES, OTHER, MOTOR VEHICLE THEFT, BURGLARY respectively.

TABLE II. NUMBER OF DATA POINTS FOR DIFFERENT CLASSES

Class 1	Class 2	Class 3	Class 4	Total
20,216	70,390	1,221	924	92,751

In Table II, we can see that there is a big difference between the numbers of each crime types because of the nature of crime occurrence probability. According to Chawla [19, 20], there are 4 ways of dealing with imbalanced data:

- 1) Adjusting class prior probabilities to reflect realistic proportions
- 2) Adjusting misclassification costs to represent realistic penalties
- 3) Oversampling the minority class
- 4) Under-sampling the majority class

In this project, we apply the under-sampling the majority class technique. In our dataset, class 3 and class 4 are the minority classes and class 1 and class 2 are the majority classes because the number of samples of class 3 and class 4 are much smaller than the number of samples of class 1 and 2. Below are the steps to construct the new dataset T from the original dataset:

- We applied k-means clustering with number of cluster equals to 2000 on two majority classes (class 1 and class 2)
- From each cluster of class 1 and class 2, select k random samples and put to new dataset T. In our experiment, we chose k equals to 3 for class 1 and k equals to 5 for class 2
- Put all samples of class 3 and class 4 into new dataset T

After new dataset T is constructed, 10-fold validation is used to divide this dataset into training set and test set and fed into selected learning machine to build classification models. These models will then be validated on the original dataset to benchmark the performance.

B. Prediction Results with First Dataset

The first dataset is created by integrating processed data after preprocessing original PPB data with demographic data based on census tract that we obtained from FactFinder. After preprocessing original PPB data, we have 6 features dataset. Then using census tract as join key, we integrate 20 more demographic features from FactFinder to form our first dataset. This dataset contains 24 features with the size of more than 92,000 records. Then we applied under-sampling technique (mentioned in 5.1 above) on our first dataset to create a training dataset. The new dataset T that was constructed from above section had the with the size of 18.145 samples (6,000 samples of class 1; 10,000 samples of class 2; 1,221 samples of class 3; 924 samples of class 4). This training dataset has the almost equal number of samples of 4 different CFS classes that were defined by the Portland Police Bureau. The CFS classes are corresponding to Burglary, Street Crimes, Theft of Auto and Others. The following subsection with discuss the results of applying different machine learning algorithms on our first dataset.

1) Support Vector Machine (SVM)

After the model were trained using SVM with Gaussian kernel on the above under-sampling training dataset, we tested the model run on all the available samples of our first dataset and got the overall 57.3% correct prediction of crime types. The confusing matrix of testing SVM model and the classification accuracy for each class are described in TABLE III.

TABLE III. RESULTS USING SVM ON FIRST DATASET

	Class 1	Class 2	Class 3	Class 4
Class 1	10242	19576	45	37
Class 2	6542	42154	57	45
Class 3	2038	5423	1100	13
Class 4	1394	3237	19	829
Precision	50.66%	59.89%	90.09%	89.72%

2) Random forest

Next, we have applied Random Forest to complete the crime prediction task. We have trained random forest models using deep tree with minimum leaf size equals to 5 on our first training set, evaluated resulting models on all samples of our first dataset and got the results

mentioned in TABLE IV. At first, we have started from using 50 trees and then adding 50 trees more next loop. In our experiment, the best accuracy results occur when we set number of tree to 100, and increased the number of trees doesn't improve the results. TABLE V below displayed of the Mean Square Error (MSE) and accuracy results when training Random Forest with different number of trees. Furthermore, TABLE V showed the confusing matrix when evaluated random forest on our first dataset (number of tree = 100).

TABLE IV. ACCURACY RESULTS USING RANDOM FOREST ON FIRST DATASET

	MSE	Precision
Number of Trees=50	0.332234	66.78%
Number of Trees=100	0.327684	67.23%
Number of Trees=150	0.331587	66.84%
Number of Trees=200	0.331705	66.83%
Number of Trees=250	0.339317	66.07%

TABLE V.Result of Random Forest on First Dataset
(Number of Tree is 100)

	Class 1	Class 2	Class 3	Class 4
Class 1	11427	19375	583	455
Class 2	8704	50827	559	440
Class 3	78	153	78	3
Class 4	7	35	1	26
Accuracy	56.52%	72.2%	6.38%	2.81%

3) Gradient boosting machines

Finally, we have applied Gradient Tree Boosting to complete the prediction task. We have directly trained and evaluated this model using AdaBoost with training set and got the results mentioned in TABLE VI. Generally, it will be better to employ more decision trees for higher prediction accuracy (lower mean square error). However, in our experiment, the best result occurs when we set number of tree estimators to 100, and it does not turn better as the number of trees increases. TABLE VI below displayed the table of Mean Square Error (MSE) when training Gradient Boosting with different number of trees. Furthermore, TABLE VII showed the confusing matrix when evaluated Gradient Boosting Machines on our first dataset (number of tree = 100).

TABLE VI. RESULTS USING GRADIENT BOOSTING MACHINES

	MSE	Precision
Number of Trees=50	0.34574	63.87%
Number of Trees=100	0.34193	64.23%
Number of Trees=150	0.34152	64.17%
Number of Trees=200	0.34408	63.81%
Number of Trees=250	0.34499	63.70%

 TABLE VII.
 RESULT OF GBM ON FIRST DATASET (NUMBER OF TREE IS 100)

	Class 1	Class 2	Class 3	Class 4
Class 1	12019	20648	64	68
Class 2	6491	45685	71	47
Class 3	1125	2511	1077	13
Class 4	581	1546	9	796
Accuracy	59.45%	64.9%	88.0%	86.1%

C. Prediction Results with Second Dataset

After working on our first dataset, we decided to develop our second dataset which also derived from original PPB data with demographic data based on census tract that we obtained from FactFinder. After preprocessing original PPB data, we have 6 features dataset. Based on demographic dataset from FactFinder. we develop new 9 features to form our second dataset. The 9 features are included: below poverty level status, age, sex, race, education status, employment status, working time status, poverty level status, past 12 month working status. Based on the data distribution from the demographic data we have, we will generate missing data using random variable with pre-determined distribution. This resulting dataset contains 15 features with the same size with our first dataset (over 92,000 records). We also applied under-sampling technique on our second dataset to create a training dataset size of 12,145 samples which contains 4,000 samples of class 1; 6,000 samples of class 2; 1,221 samples of class 3; and 924 samples of class 4. The following subsection with discuss the results of applying different machine learning algorithms on our second dataset.

1) SVM

After the model were trained using SVM with Gaussian kernel on above training dataset, we tested the model on the whole sample of the second dataset and got the overall 79.39% correct prediction of crime types. The confusing matrix of testing SVM model and the classification accuracy for each class are described in TABLE VIII. We got good results to classify class 2, class 3 and class 4 but very bad accuracy with class 1.

TABLE VIII. RESULT OF SVM ON SECOND DATASE	TABLE VIII.	RESULT OF SVM ON SECOND	DATASET
--	-------------	--------------------------------	---------

	Class 1	Class 2	Class 3	Class 4
Class 1	1398	2	0	0
Class 2	18816	70324	367	281
Class 3	2	34	854	1
Class 4	0	30	0	642
Accuracy	7.0%	99.9%	70.0%	69.5%

2) Random forest

Next, we trained a model using Random Forest deep tree with minimum leaf size equals to 5 on second training dataset to complete the crime prediction task. We trained different Random Forest models which ranged from 50 trees and 300 trees, and then recorded the classification performance. In our experiment, the best accuracy results occur when we set number of tree to 250 (65.79% accuracy). In TABLE IX, the result confusing matrix and classification accuracy on each class when training Random Forest with 250 trees.

TABLE IX. RESULT OF RANDOM FOREST ON SECOND DATASET

	Class 1	Class 2	Class 3	Class 4
Class 1	11878	473	0	0
Class 2	8014	47647	122	133
Class 3	278	17978	1047	341
Class 4	46	4292	52	450
Accuracy	58.75%	67.69%	85.75%	48.7%

3) Gradient boosting machines

Next, we trained a model using Gradient Boosting using AdaBoost training technique on second training dataset to complete the crime prediction task. We trained different Gradient Boosting Tree models which ranged from 50 trees and 300 trees, and then recorded the classification performance. In our experiment, the best accuracy results occur when we set number of tree to 300 (61.67% accuracy). In TABLE X, the result confusing matrix and classification accuracy on each class when training GBM with 300 trees.

ΓABLE X.	RESULT OF GRADIENT BOOSTING TREE ON SECOND
	DATASET

	Class 1	Class 2	Class 3	Class 4
Class 1	12869	459	0	0
Class 2	7016	42298	18	14
Class 3	236	16491	1167	45
Class 4	95	11142	36	865
Accuracy	63.65%	60.09%	95.57%	93.61%

4) Neural networks with scaled conjugate backpropagation and resilient backpropagation

Lastly, we trained neural networks on our second dataset using two popular backpropagation training techniques which are Scaled Conjugate Backpropagation (SCG) and Resilient Backpropagation (RP). We constructed our neural network using 2 hidden layers. Different combination of different number of nodes of two layers were tested and classification performance were recorded.

With SCG, we obtained the best result of 74.02% classification accuracy with 60 nodes of hidden layer 1 and 40 nodes of hidden layer 2. With RP, we could obtain the best result of 74.24% classification accuracy with 40 nodes of hidden layer 1 and 25 nodes of hidden layer 2. In TABLE XI. and TABLE XII, the confusing matrix of classification using the two best neural network models that obtained from SCG and RP training methods.

TABLE XI. RESULT OF NEURAL NETWORK ON SECOND DATASET WITH SCG TRAINING

	Class 1	Class 2	Class 3	Class 4
Class 1	12659	952	4	1
Class 2	7170	52402	484	424
Class 3	295	12304	590	297
Class 4	92	4732	143	202
Accuracy	62.62%	74.44%	48.32%	21.86%

TABLE XII. RESULT OF NEURAL NETWORK ON SECOND DATASET WITH RP TRAINING

	Class 1	Class 2	Class 3	Class 4
Class 1	13373	1134	0	0
Class 2	6493	54785	548	494
Class 3	278	11513	581	308
Class 4	72	2958	92	122
Accuracy	62.2%	77.8%	47.6%	13.2%

VI CONCLUSION

For the crime prediction problem offered as a competition by the U.S. National Institute of Justice, we

have started with preprocessing the dataset from Portland Police Bureau. Then, we have attempted to select some helpful features to represent the attributes of the samples in a proper manner. We have divided the total forecast area into hotspots. We have also linked demographic data from other public sources such as Fact Finder, U.S. census bureau and it has resulted in improving the performance of our models significantly. Finally, we have trained some models with the features we employed and calculating the probabilities of different categories of crimes. Because of the large dataset and problem of imbalanced data, we employ under-sampling technique on our dataset to reduce the training set size to less than 20,000 samples and the accuracy of the model is affected by the accuracy of under-sampling method.

With our first dataset in which demographic was used as it is, as shown in the results, SVM does not seem to be a suitable model for this task because of the bad classification accuracy when compared to other used methods. Ensemble methods such as Random Forest or Gradient Boosting turned out to be the two best models when compared the performance of them with SVM. These two methods can handle with big training sets and the training time is faster than the training time of SVM model.

With our second dataset that we use demographic data as reference to generate missing features, SVM and Neural Network show the best accuracy models when compared to other two ensemble methods of Random Forest and Gradient Boosting Machines. Overall, the classification accuracy of different machines on our second dataset are better than our first dataset. However, there are imbalanced classification accuracy between four desired classes which resulting models have large misclassification of one of four classes while the classification of other three classes are very good.

In the era of big data, analytics are increasingly being used for modeling, prediction, knowledge extraction, and decision making, etc. How to make the best use of datasets that are missing important features poses an often very challenging problem in data mining tasks. This research demonstrates a successful approach to build learning machine models with insufficient features. The authors are continuing the work to explore methods to handle imbalanced data, and to develop a more general model to predict crime type, time, and place using the best performing algorithms.

REFERENCES

- C. N. Trueman. (Mar 2015). Why do people commit crime. [Online]. Available: http://www.historylearningsite.co.uk/sociology/crime-anddeviance/ why-do-people-commit-crime.
- [2] D. M. Gottfredson, "Assessment and prediction methods in crime and delinquency." *The Plenum Series in Crime and Justice*, Book Part VI, pp. 337-372, 1995.
- [3] H. Z. Yu, F. X. Liu, and K. Q. Zou, "Improved fuzzy BP neural network and its application in crime prediction," *Journal of Liaoning Technical University* (Natural Science), 2012.
- [4] G. Mohler, "Marked point process hotspot maps for homicide and gun crime prediction in Chicago," *International Journal of Forecasting*, vol. 30, no. 3, pp. 491-497, July-September 2014.

- [5] Office of Justice Programs. Why Crimes Occur in Hot Spots. [Online]. Available: http://www.nij.gov/topics/lawenforcement/strategies/hot-spot-policing/pages/why-hot-spotsoccur.aspx.
- [6] 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 (IJDKP)*, vol. 5, no. 4, July 2015.
- [7] DENVER Open Data Catalog, http://data. denvergov.org/dataset /city- and-county-of-denver-crime. [Accessed: 20- May- 2015].
- [8] National Institute of Justice, https://www.nij.gov/ funding/ Pages/ fy16-crime-forecasting-challenge. aspx#data
- [9] American FactFinder, https://factfinder.census.gov/ faces/ nav/ jsf/ pages/index.xhtml
- [10] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. the Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152. ACM, 1992.
- [11] B. Léon, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun et al., "Comparison of classifier methods: a case study in handwritten digit recognition," *IAPR International Conference*, vol. 2, 1994, pp. 77-82.
- [12] T. Simon and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. the Ninth ACM International Conference on Multimedia.* ACM, 2001.
- [13] B. Leo, "Random forests," *Machine Learning*, vol. 45, no. 1, 2001, 5-32.
- [14] F. Yoav and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.
- [15] Møler and Martin Fodslette, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525-533, 1993.
- [16] O. José and C. A. R. Garc á, "Detecting pathologies from infant cry applying scaled conjugate gradient neural networks," *European Symposium on Artificial Neural Networks*, *Bruges (Belgium)*. 2003.
- [17] R. Martin and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE International Conference on Neural Networks*, 1993, pp. 586-591.
- [18] MathWoks, Inc., Matlab.NeuralNetworksToolbox, V.9.1.0 [Online]. Available: https://www.mathworks.com/
- [19] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data Mining and Knowledge Discovery Handbook*, Springer US, vol. 4, pp. 853-867, 2005.
- [20] Rahman, M. Mostafizur, and D. N. Davis. "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, p. 224, 2013.



Trung T. Nguyen received the B.E degree in Computer Science and Engineering from HCM University of Technology, Vietnam, in 2006. He also received M.S degree in Computer Science from HCM University of Technology, Vietnam, in 2009 and M.S degree in Business Information Systems from Heilbronn University, Germany in 2013. In 2016, he joined the School of Computing, The University of Southern Mississippi, as a PhD

Student in Computational Science. His current research interests include computer vision and image processing, data mining and pattern recognition, machine learning, deep learning as well as their applications. Mining and pattern recognition, machine learning, deep learning as well as their applications.



Amartya Hatua received his B.Tech degree in Information Technology from Calcutta Institute of Engineering and Management, India in 2010. He completed his M.Tech in Information Technology from National Institute of Technology Karnataka, India in 2013. In 2016, he joined the School of Computing, The University of Southern Mississippi, as a PhD Student in Computational Science. His current research interests include computer vision and image processing, data mining and pattern recognition, machine learning, deep learning as well as their applications.

Andrew H. Sung received B.S. degree in Electrical Engineering from National Taiwan University, M.S. degree in Mathematical Sciences from University of Texas at Dallas, and Ph.D. degree in Computer Science from State University of New York at Stony Brook (1984). He has served in the Air Force of Taiwan as Second Lieutenant for 2 years, and Assistant, Associate, and Professor of Computer Science and Engineering in three public universities in the U.S. over 30 years. Since 2014, he has been the Director and Professor of the School of Computing, The University of Southern Mississippi, in Hattiesburg, Mississippi. Dr. Sung has over 200 publications and is currently working in the areas of data mining, multimedia forensics, and cybersecurity.